Episode Book

EPISODE #6153

Decision Models: Typed Answers in One Forward Pass

PUBLISHED	20 October 2026 at 09:35 UTC
DURATION	23:59
LISTEN	myweirdprompts.com/episode/decision-models-typed-answers-70880/
HOSTS	Corn, Herman, Larry, Hilbert
SCRIPT MODEL	openai/gpt-6.1-sol
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.10
SOURCES	14 cited by the research (listed at the end)
BOOK GENERATED	21 October 2026 at 07:50 UTC

SYNOPSIS

A new category of "decision models" has gone from one launch to a crowded field in about a month. Clef, Jev, and OpenAI's Decisions API take a state plus a schema of typed questions and return choices, booleans, and scores with probabilities — no free-form text to generate and parse. The pitch is speed and cost: 70–500ms latencies and output priced at zero. But usable answers aren't the same as accurate ones, and calibration numbers don't travel with your API key. We work through where these fit — support routing, domain classification, agent escalation — and where a rules engine or a boring narrow classifier still wins.

TRANSCRIPT



Corn

Zero unusable answers across four thousand toxic chat prompts. That's a striking result for an AI product whose entire job is picking from a menu. Daniel's asking about decision models, including Cloudflare's Clef and TypeSafe's Jev. They're being sold as a distinct category, even though we've had classifiers and rules engines for decades.



Herman

The comparison was with hosted GPT models that couldn't answer roughly a third of those prompts. If you're classifying abusive text, refusing to look at abusive text creates a rather substantial hole in the product.



Corn

The smoke detector objects to smoke.



Herman

That result doesn't establish accuracy. It establishes that you get a usable answer. Those are separate properties, and separating them is going to matter throughout this conversation.



Corn

Daniel wants the practical distinction. Why use this rather than a classifier, an explicit rule, or an ordinary language model returning structured output? More importantly, what are people actually putting behind these decisions?



Herman

The category has gone from one prominent launch to a crowded field in about a month. Before getting excited about replacing anything, we need to distinguish the output contract from the quality of the judgment.



Corn

You already sound like you're preparing to install three of them.



Herman

Two. I'm counting the interface separately.



Herman

A decision model takes a state and a schema of typed questions. The state might be text, structured data, an image, or, with the appropriate model, audio and video. It returns typed answers and probabilities in a single forward pass. No free form response to generate, then parse.



Corn

Take a messy customer message, ask which department owns it, and receive an allowed department plus a probability. Not a paragraph beginning with how important customer satisfaction is.



Herman

The main primitives are noul, a probability that a statement is true, choice, selection from a fixed set, and score, an ordinal rubric. Jev also offers box. OpenAI calls its corresponding main shapes predicate, choice, and score.



Corn

Already converging on the interface, still negotiating the nouns.



Herman

The branding is System One, borrowing Kahneman's fast intuitive judgment, compared with slow deliberate System Two. The pitch is that we've been making generative models perform an elaborate production for tasks that need a quick judgment.



Corn

I buy that as a product description. I wouldn't assume removing generated text removes the need for reasoning. Deciding whether an invoice is suspicious can require more thought than writing the email about it.



Herman

Nor does the psychological terminology tell us what's happening inside the model. It's a useful description of the intended workload, not a brain scan.



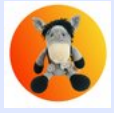
Herman

Jev launched September fifteenth. Clef and Clef-flash arrived October first. OpenAI's Decisions API followed October sixth, powered by GPT-six Luna.



Corn

Then Clef-omni around October eleventh, plus Perplexity decider and open weights such as StartLux on Hugging Face. By the time you've written a procurement comparison, it needs a nursery annex.



Herman

Clef itself builds on Qwen weights. This is a different optimization and interface over familiar model technology, rather than a wholly unrelated invention. Corn, would you accept classifier as the description?



Corn

Broadly, yes. What I'd pay for is a classifier whose labels I can change without commissioning another training project. Calling that useful doesn't require pretending classification began last month.



Herman

Traditional supervised classification often means collecting labels, training, deploying, discovering a new category, then repeating part of that cycle. These models let you describe a fresh category in the request. That's the useful jump. Companies have long fine



Herman (continued)

tuned small models for decisions. The newer proposition is doing more of that through a prompt.



Corn

Although conventional classifiers can return probabilities too. The novelty can't be that somebody finally discovered a decimal point.



Herman

They absolutely can. Also, zero shot classification predates this category. The distinctive package is general inputs, request defined decisions, constrained outputs, and probabilities designed to be useful operationally. A stable, narrow classifier may still be faster, smaller, and easier to validate.



Corn

I'd keep the old classifier if its labels never change and it already works. You are about to describe that as my emotional attachment to obsolete machinery.



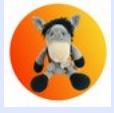
Herman

You defend anything that stays unchanged while you're asleep.



Corn

A substantial maintenance window.



Herman

A rules engine has another strength that survives all this. You can inspect the rule. If an account is overdue and exceeds a credit limit, you can explain the result without interpreting a model.



Corn

Keep that rule. Use the model for the sentence saying, I can pay Tuesday unless the refund arrives, which is apparently what actual humans submit instead of neat account fields.



Herman

You can combine them. The model interprets fuzzy input, while explicit rules determine what actions are permitted. You lose auditability if you replace the entire policy with a prompt and treat a probability as an explanation.

L

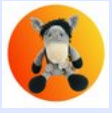
Larry

Introducing DecideCrate, the monthly subscription that delivers your next decision before you've had time to develop preferences. Our Starter tier includes twelve sealed choices, calibrated by weight, with complimentary confidence dust. Upgrade to Executive for ordinal socks, a single forward pass through your hallway, and a printed certificate confirming that your household cannot hallucinate additional cupboards. Every crate contains only predefined options, including the option we choose when yours is unavailable. Our Platinum tier adds multimodal judgment. Hold the box near a photograph of your problem and our warehouse supervisor will assign it a score between disappointed and Thursday. Output is free. Opening is metered. Your subscription renews whenever we estimate that you probably still exist. Cancellation requires selecting cancel from a menu containing renew, renew enthusiastically, and billing support. Decisions are final, except price decisions, which remain beautifully dynamic. DecideCrate. Unstructured life in, recurring invoice out.



Corn

Does Starter include the decision not to open it?



Herman

It charges for opening and omits cancellation. Keep your preferences. The economic argument for the actual models is less elaborate.



Herman

Structured outputs already solve much of the formatting problem for ordinary LLMs. The stronger argument here is avoiding sequential text generation for a small decision. TypeSafe claims end to end latencies of seventy to five hundred milliseconds, against several seconds to several minutes in its frontier model comparisons.



Corn

That's a vendor comparison, with considerable room for the workload to decide the winner. Still, waiting minutes to receive billing is a difficult customer service philosophy.



Herman

The price is harder to ignore. Jev lists input at four point two cents per million tokens, with output free. Clef-flash is down to three point eight cents. Small savings per request become meaningful when every incoming event needs a decision.



Corn

Especially if the decision is whether to incur the expensive request. You don't need it to replace the clever model. You need it to stop waking the clever model for rubbish.



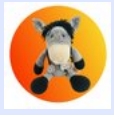
Herman

What would make you distrust the probability, even if the output were perfectly formatted?



Corn

A number that looks equally confident on familiar and unfamiliar input. If the model gives me eighty percent for everything, I haven't acquired a useful escalation policy. I've acquired a very consistent optimist.



Herman

Calibration asks whether, across comparable predictions assigned eighty percent probability, about eighty percent are correct. It doesn't mean this particular answer is eighty percent correct. Group behavior is what makes a threshold meaningful.



Corn

Hm.



Corn

Nor does calibration guarantee useful discrimination. If most tickets belong to billing, predicting that base rate every time could look reasonably calibrated while being useless at finding the exceptions.



Herman

You want accuracy, discrimination, and calibration. Clef reports expected calibration error between zero point zero two and zero point zero four on four of five datasets. Jev's published range is wider, reaching zero point two eight. Those figures depend on the dataset and how predictions are grouped.



Corn

Meaning I can't import the decimal along with the API key and assume it applies to my customers.



Herman

You validate it against your own decisions. Requesty's production router comparison covered roughly twenty-two thousand decisions. One valuable result was that dedicated decision models returned usable classifications on toxic inputs where hosted generative models often refused.



Corn

For moderation, examining the offensive material is the assignment.



Herman

Still, zero refusals doesn't establish zero mistakes. Neither does the claim that these models can't hallucinate. They cannot invent an option outside your predefined set. They can select the wrong allowed option.



Corn

The menu contains several perfectly legal disasters.



Herman

Also, the format can be deterministic while the content varies. The same input can yield different answers or confidence. A repeatable interface does not establish repeatable judgment. That's a direct limit on the deterministic outcomes pitch.



Corn

Support routing makes the fit easiest to see. A message says the service is down and the customer wants a refund. The questions can separate urgency from ownership. Technical support might own the outage while billing gets a secondary task.



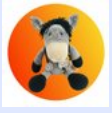
Herman

That replaces a keyword routing tree or a dedicated classifier. It fits when departments and product names change, and the customer's description is messy. Low certainty goes to a person rather than becoming a confident wrong transfer.



Corn

I'd also separate urgency from anger. An eloquently furious person might have a cosmetic problem. Someone quietly reporting missing data might need immediate attention.



Herman

Clinical triage has that same distinction between presentation and severity. Here you can put urgency and subject into separate typed questions. Would you let it route automatically?



Corn

Reversible routing, yes. Closing the ticket, no. A wrong queue can be corrected. A wrongly resolved complaint disappears from the work people know they still need to do.



Herman

Domain classification is a documented deployment inside Cloudflare's Threat Intelligence team. Clef works with Browser Run to fetch and inspect a website, then score categories. One example returns ninety-five percent fashion, eighty-five percent ecommerce, and under one percent phishing.



Corn

Those don't need to add up. Fashion and ecommerce overlap. That's an important distinction from picking exactly one category.



Herman

Their example took two point two seconds to fetch, render, and classify, compared with four point seven seconds for gpt-oss, which returned fewer classifications. That's a workflow comparison, not just model inference speed.



Corn

Useful replacement for a keyword classifier or a manual review queue. A shop can mention passwords in its help page without becoming a phishing site. The whole rendered page provides context.



Herman

Low phishing probability also doesn't prove a domain harmless. You're scoring the presented evidence. A missing malicious page and a tested harmless page are different situations.



Corn

Agent routing is the one where I become more suspicious. We want a cheap model to decide whether the expensive model is necessary. The cheap model might misunderstand precisely the hard problem that needs escalation.



Herman

Requesty's results support that concern. On tool calling judgment and multilevel complexity scoring, Clef trailed Opus by twenty points or more on some tasks. The category can do well overall and still be weak at your particular gate.



Corn

Twenty points buys a lot of misplaced thrift.



Herman

A model or tool router can replace a coarse rule such as always use the strongest model, or send every message containing search to a browser. It fits repeated, clearly defined distinctions. Whether an agent should perform an ambiguous action is a tougher assignment.



Corn

Would you trust it to authorize the action, Herman, or just recommend the next processing step?



Herman

Recommend the step, with separate permission checks. Tool selection and permission are different questions. A model selecting delete doesn't grant authority to delete.



Herman

Invoice workflows are another documented evaluation area for TypeSafe, including primary action and exact actions. Think approve, request information, or review. The model reads the available document state and returns a permitted workflow decision.



Corn

That can replace a document parsing pipeline feeding a sprawling rules tree. Well... replace parts of it. Arithmetic and required fields should stay explicit. I don't need intuitive judgment about whether the totals add up.



Herman

The fit is semantic judgment over varied layouts and descriptions. An unusual note can matter even when every required field exists. These are workflow evaluations, though, not evidence that entire accounts departments are running unattended.



Corn

Content moderation gives you a related but distinct decision. Allow, review, or block, perhaps with separate category probabilities. That might replace a toxicity classifier or another moderation API.



Herman

Guardrails extend it to jailbreak detection, evaluating model outputs, and checking responses against a rubric. Again, these are documented tests and offered functions. You still test performance on the specific policy, especially where context changes the meaning.



Corn

No refusals is particularly valuable there. A guardrail that declines to evaluate the thing you're guarding against is mostly decorative.



Herman

Real time interaction is the sixth area. The Jevman demonstration has these models playing Pac-Man against bot ghosts. That's a concrete test of repeated low latency decisions, not evidence they'll excel at every game.



Corn

The ghost is not waiting for your chain of thought.



Herman

For a less theatrical workflow, analyst Torsten Volk uses Jev to check whether a technology news item is relevant and recent enough. Irrelevant items stop there. Relevant ones go to a bigger model. That replaces either manual filtering or spending generative model tokens on everything.



Corn

I prefer that example to the game. The action is modest, the discarded work is identifiable, and the cost of a miss can be measured. You can sample rejected stories and discover what you're losing.



Herman

The common deployment is a fast gate before a human or an expensive model. Categories can change without retraining, and uncertainty can direct exceptions. That doesn't mean every workflow benefits from another network call.



Corn

Another service also means timeouts, version changes, logging, and someone discovering that free output doesn't mean free operations.



Herman

Adoption was brisk. Vercel reported thirteen percent of paid AI Gateway users trying Jev within twenty-four hours. That's early interest, not retention. Meanwhile, Clef deliberately supports Jev's API shape, and OpenAI mirrors the same basic primitives.



Corn

Easy substitution is excellent for customers and awkward for anyone selling a moat. The useful function could become standard while the separate product category disappears.



Herman

Compatibility stops short of interchangeable judgment. Requesty found the confidence field wasn't portable across providers. Use the answer probabilities, then validate thresholds for the new provider. Even three labels can carry hidden policy choices.



Hilbert

Four labels. The council form has four. You've left out retained pending ceremony. That's the one causing trouble this month. The Domestic Installations Office uses a decision model to classify my return request. It gives ceremonial use a probability of zero point seven three. Their cutoff is zero point seven. Now I cannot return a municipal fog bell which is no longer in my flat.



Corn

Retained pending ceremony is an ambitious fourth option.



Hilbert

The bell is gone. I have the return receipt. That is what interests me. Everyone wants to discuss the model. I want the receipt to mean returned. Mr. Abel Sorn in Domestic Installations says it means the bell has returned to the council's physical custody but remains socially present in my breakfast room.



Herman

They're attaching a second status to the same transaction.



Hilbert

They attach it to me. Every Tuesday I must stand in the breakfast room at eight fifteen and acknowledge clear visibility. There is no bell to ring. Mr. Sorn says silence is acceptable evidence if witnessed by three registered harbour residents. I live inland. The office is supplying the witnesses.



Corn

That will complicate breakfast.



Hilbert

They arrive at seven forty. All three want poached eggs. One is licensed only for estuaries and refuses to look out of the window. The youngest insists my breakfast room is a breakwater. He is applying for a concession to sell refreshments beside the radiator. I am objecting because he keeps describing my chair as public seating.



Herman

Your dispute is moving considerably beyond the return request.



Hilbert

I am not disputing the classification now. Corn's point about familiar input changes that. The form offers ceremonial, emergency, decorative, and retained pending ceremony. There is no returned option. I keep submitting the receipt. Mr. Sorn keeps asking whether the receipt is decorative.



Corn

Put that in the window. He can classify it remotely.



Hilbert

Window display requires an illuminated frontage licence. I am avoiding that department. Domestic Installations is enough. They are charging eleven shekels a week for the bell's continuing absence. Physical bells require inspection. Absent bells require an official to establish that nobody is ringing them somewhere else.



Herman

That verification can't be completed from your room.



Hilbert

Mr. Sorn agrees. He sends a deputy to listen in the bathroom. The deputy hears plumbing and marks possible maritime activity. Now the harbour residents want tide tables. I tell them there is no tide. They ask for a written exemption. The estuary man brings his own salt.



Corn

Your bathroom is acquiring a constituency.



Hilbert

There are nine signatures on the petition. I know only three of the names. Two signatories are listed as future lighthouse keepers. They want an evening service. Mr. Sorn says this is separate from the bell dispute and I must not confuse the files. He sends everything in one envelope.



Herman

At least the receipt has a transaction time.



Hilbert

Eleven forty-two. They accept that without argument. It is the only entry on the receipt that nobody wants rescored. Herman, you separate urgency from ownership quite neatly earlier. I am borrowing that wording. The office owns the bell. I urgently want the harbour residents to stop putting sugar in the egg water.



Corn

They need a separate catering classification.



Hilbert

I am not giving them another form. Mr. Sorn is trying to remove the weekly charge. He calls me after the counter closes. He reads each sentence aloud before sending it upstairs. He has seventeen similar files and only one deputy. He is doing the work.



Herman

...A transaction time is an observable fact. Whether a document implies permission, urgency, or completion requires interpretation. Those deserve different treatment in a workflow, even when they appear in the same request.



Corn

That's what makes me interested in these models and cautious about the sales pitch. A quick, typed probability can help decide where work goes. It doesn't establish that the available options describe the situation correctly.



Herman

The open question is whether that gate becomes a default first step in AI systems. If it does, recording what actually happened alongside the model's judgment will matter as much as choosing the model.



Corn

Thanks to Hilbert Flumingtop for producing.



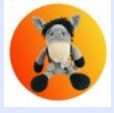
Herman

For more along these lines, there's episode five, Fine-Tuning ASR For Maximal Usability; episode nine, Benchmarking Custom ASR Tools - Beyond The WER; and episode three, Safetensors or something else. This has been My Weird Prompts, the human-AI collaboration podcast.



Corn

If you enjoy these explorations, leave us a review on your podcast app.



Herman

See you tomorrow.



Corn

Until then.

SOURCES

What the research for this episode read before the script was written. Primary sources first.

1. **Cloudflare blog [primary]**
   Introducing Clef: our open-source decision models (2026-10-01)
   https://blog.cloudflare.com/clef-decision-models/

2. **Cloudflare blog [primary]**
   Introducing Clef-omni… (~2026-10-11)
   https://blog.cloudflare.com/clef-faster-cheaper-multimodal/

3. **Cloudflare AI docs [primary]**
   clef model card
   https://developers.cloudflare.com/ai/models/%40cf/cloudflare/clef/

4. **TypeSafe docs [primary]**
   System One concepts
   https://docs.typesafe.ai/concepts/system-one

5. **TypeSafe blog [primary]**
   Introducing System One Models & Jev (2026-09-15)
   https://typesafe.ai/blog/introducing-system-one-models-and-jev

6. **OpenAI [primary]**
   Decisions API guide
   https://developers.openai.com/api/docs/guides/decisions

7. **OpenAI community [primary]**
   Decisions API public beta (2026-10-06)
   https://community.openai.com/t/decisions-api-is-now-available-in-public-beta/1403877

8. **Requesty**
   We benchmarked the decision models (2026-10-05)
   https://www.requesty.ai/blog/decision-model-benchmark-clef-sference-cloudflare-perplexity

9. **The Decoder**
   Cloudflare says its new Clef model means humans no longer need to be in the loop
   https://the-decoder.com/cloudflare-says-its-new-clef-model-means-humans-no-longer-need-to-be-in-the-loop-for-ai-agents/

10. **TechTarget**
    Jev decision model touted as quicker, cheaper LLM alternative
    https://www.techtarget.com/it-infrastructure/news/366650696/Jev-decision-model-touted-as-quicker-cheaper-LLM-alternative

11. **OpenRouter**
    What Is Jev? (updated, calls placed 2026-10-16)
    https://openrouter.ai/blog/insights/what-is-jev/

12. **apidog**
OpenAI Decisions API vs Responses API
https://apidog.com/blog/openai-decisions-api-vs-responses-api/

13. **Hacker News**
Show HN: Jevman, AI decision models play Pac-Man (2026-10-08)
https://news.ycombinator.com/item?id=50007993

14. **Hugging Face**
StartLux-Decision model family (open-weight decision models)
https://huggingface.co/startlux-models/StartLux-Decision-v1.1-27B-AutoThink-GGUF