Episode Book

EPISODE #6201

Inside Cloudflare's Clef: A Language Model That Can't Speak

PUBLISHED	22 October 2026 at 19:58 UTC
DURATION	21:58
LISTEN	myweirdprompts.com/episode/cloudflare-clef-language-model-probabilities-06408/
HOSTS	Herman, Corn, Daniel, Dorothy, Hilbert
SCRIPT MODEL	anthropic/claude-sonnet-5.5
VOICES	chatterbox-regular+meorav (chatterbox-local)
PIPELINE	V5.19
BOOK GENERATED	23 October 2026 at 07:44 UTC

SYNOPSIS

Cloudflare's Clef is a 27-billion-parameter model with a vow of silence: it reads text, images, and video, and returns nothing but a table of probabilities. Underneath, it's a post-trained Qwen 3.8 with a frozen backbone and a small routing head — and that architecture turns out to matter a lot. This episode opens the hood on Clef, Clef-flash, and Clef-omni, compares them with Jev, TypeSafe, and OpenAI's Decisions API, and asks the uncomfortable question: if the format is guaranteed but the content isn't, how much should you trust a probability?

TRANSCRIPT



Herman

Twenty-seven billion parameters, and it is not allowed to say a single word. That's Cloudflare's Clef. It reads text, images, even video, and what comes back is a table of probabilities and nothing else. No sentence, no explanation, not so much as a throat clear.



Corn

A vow of silence, at that size. Daniel's in the studio for this one, and it's worth saying once, plainly, that his voice is an AI clone of the real thing. Go ahead.



Daniel

Did we already do a deep dive on the new predictive models, Jev and Clef? I remember one about what these do that structured output models don't. But I think the more interesting question, technically, is what they actually are. What's going on behind the curtain? From what I understand they're LLMs, trained for this new format. Is that right?



Corn

We've done two, both on what they do compared with structured outputs. Neither one opened the hood, so this is new ground.



Herman

It's right, and less exotic than it sounds. Clef is a language model underneath. Cloudflare post-trained it from Qwen three point eight, the twenty-seven billion parameter version, open weights, Apache licence. The small one, Clef-flash, starts from a nine billion parameter Qwen. Anyone can download the starting points.



Daniel

What does trained for this format actually mean, though? Is it ordinary fine-tuning, or something stranger?



Herman

A bit stranger. Cloudflare says the backbone stays frozen, and what they train is a small routing head plus low-rank adapters, rank two hundred and fifty-six. The Hugging Face card says post-trained, which sounds like more than adapters, and I haven't seen the two descriptions reconciled. So whether any backbone weight actually moved, I can't tell you.



Daniel

Frozen and post-trained can't both be true, can they? Pick one.



Herman

They can if the adapters get merged back into the weights before release. Then the checkpoint on Hugging Face looks like a fully post-trained model, and the blog is still right that the backbone was frozen during training. That's my guess at how the two fit, and nobody has confirmed it.



Corn

Frozen backbone, adapters, a head. That's a long way of saying the model still knows English and has been told to stop using it.



Herman

It does still know English, and that's the entire value. Everything those billions of parameters learned about the world gets reused, and the vision encoder that comes with the backbone comes along too. What changes is the last step. A normal model turns its final internal state into a score for every word in its vocabulary and picks one. They take that step out.



Daniel

If it's Qwen underneath, does it carry Qwen's picture of the world, with Qwen's cutoff? For predicting events that seems like it matters.



Herman

I'd think so, yes. Nothing in the training Cloudflare describes teaches it new facts. The adapters teach format and calibration. That's my inference, not something the card says outright. So anything that happened after the base model was trained has to arrive through the state.



Corn

A well-read man locked in a cellar, with the newspapers passed under the door.



Daniel

Could I download it and run it myself?



Herman

The weights are public. In bfloat sixteen that's two bytes per parameter, so the big one is around fifty-four gigabytes before you've loaded a single question. Clef-omni wants about sixty-four gigabytes of GPU memory. Cloudflare also hosts both, so most people won't bother.



Corn

My laptop just left the room.



Herman

Corn, does it change how you see them, that it's a stock Qwen with a head bolted on? I figured you'd feel a bit cheated.



Corn

A bit. I'd pictured something built for the job from the ground up. This is a Qwen in a new uniform.



Herman

The uniform is the whole product, though. Same brain, different job, and the new job turns out to need almost none of the old output machinery.



Daniel

Is there just the one Clef, or are there variants?



Herman

Three that matter. The big one, the flash, and Clef-omni, which Cloudflare put out on October ninth. Omni isn't a Qwen three point eight at all. It's post-trained from Qwen three Omni, a mixture of experts where only a slice of the model fires for each token. They threw away the speech-output machinery, because nothing here talks, and kept the ears.



Corn

Dropped the mouth, kept the ears. My brother runs it the other way round.



Herman

I can hear you, Corn. Donkey ears. Best audio hardware on this desk.



Daniel

Since I'm thinking about event prediction, what can I actually feed these? Which of them take what?



Herman

This is where the two earlier episodes disagreed, and we got it half right each time. It depends on the model. Jev's stable contract is text and JSON. On October fourteenth, in version one point fourteen, TypeSafe added opt-in danger zone headers for images, a bounding box question and a longer context. Images are still only PNG or JPEG, and TypeSafe itself calls the results too uneven for dependable automation. No audio, no video.



Corn

So text-only was true, as long as nobody flipped a header.



Herman

OpenAI takes text, JSON and images, either base sixty-four or a public link. Audio and files are listed as unsupported, and video isn't mentioned. It's a tenth of a dollar per million input tokens, and the output is free, because the output is tiny. Clef, the twenty-seven billion one, takes text, JSON, images and video as arrays of frames. Clef-omni is the only one that adds audio, WAV or M P three with sixteen kilohertz mono recommended, plus proper video files at two frames a second.



Daniel

So if I wanted audio or video in the evidence, it's Clef-omni, full stop.



Herman

Full stop. It's hosted on Workers AI at fifteen cents per million input tokens, around one hundred and thirty milliseconds for text and about a second and a half for a twenty-one second video with sound. Context windows I'd skip, because the sources disagree and I can't pin a number.



Corn

A press conference with the sound on, scored in a second and a half. That's a faster read than I manage.



Herman

Faster than any of us, and it still won't write a word about what it heard.



Corn

Fine, but I still don't see what's left. A language model's whole job is picking the next word. Take that away and it's a very large thermostat.



Herman

Well, before we dig into the architecture, we have a caller on the line. Dorothy, you're on.

D

Dorothy

Hello dear, it is your mother. Time for my grim fact.

D

Dorothy

Did you know that Jev will not tell anyone what it is made of? Not the base model, not the size, not the weights. Three things, all hidden, and they will not say if they ever will.

D

Dorothy

Anyway. I made a casserole. When are you coming to see me, sweetheart?



Corn

Hello, Mum. We are recording, Mum.



Herman

Back to the thermostat. What's left is everything that understood your question. The Qwen backbone reads the whole state once, text and images together, in a single prefill pass. Then a small transformer Cloudflare calls the joint schema head reads the final hidden states, sends the relevant evidence to each question, and scores every declared option of every question together. Out comes one number per option, and a softmax per question turns them into probabilities.



Corn

So it isn't the old trick of masking the vocabulary down to yes and no and reading the scores.



Herman

Not for Clef. The options aren't borrowed from words at all. The head is trained to score them directly, which is why the output can't wander off and invent a category.



Daniel

Why score all the questions together? Why not one at a time?



Herman

The questions usually share evidence, and one read of the state can feed all of them. OpenAI's version allows four hundred questions in a single request. One pass over the evidence, four hundred answers.



Corn

Four hundred questions on one read. Shortest exam in history.



Herman

That's the economics, Corn. The joint head is why four hundred answers cost about what one does.



Corn

I was waiting for you to stop saying declared options. It sounds like a customs form.



Herman

I'll say declared options as often as I like. It's the correct term.



Daniel

And how do they train it to be calibrated? I assume it's not just that you ask the model to be honest about its confidence.



Herman

No, it's more involved than asking. Cloudflare trains with label-smoothed cross-entropy on valid outputs, plus a Brier loss, which punishes confident wrong answers and rewards honest spread. The data is internal and synthetic, with field order, prompts and schema structure all shuffled so the model can't lean on layout. Then a stage they call RLCD, reinforcement learning for calibrated decisions.



Daniel

What does RLCD do that the loss doesn't?



Herman

Partial credit. Say the options are a scale, low to high. If the truth is medium-high and the model puts its weight on high, that's a better miss than putting it on low, and RLCD rewards it. A plain classifier treats both misses as equally wrong. The training data isn't public, so nobody outside can check the claim.



Corn

And Jev uses the same acronym. Same thing?



Herman

Nobody can say. TypeSafe uses the name RLCD too, and neither company has published a paper. Same label, maybe the same method, maybe a coincidence of marketing.



Daniel

So what do we actually know about Jev, then? And the OpenAI one?



Herman

Jev is closed, API only, early access, and the announcement mentions a new architecture and a parallel sampler. That's the extent of it. For OpenAI, the Decisions API runs on one model, GPT six Luna, and I don't know whether it's a head on Luna or a separate fine-tune. There's an open reimplementation, OpenJev, the four billion version, that puts a similar head on a Qwen backbone and trains it with reinforcement learning. So the approach is reproducible. It tells you nothing about what's inside Jev, though.



Corn

Do you think the closed ones are even the same kind of animal? Or are we assuming it from the family resemblance?



Herman

Family resemblance, honestly. The behaviour matches, one pass and typed output, so I'd bet on a similar shape. But that's a bet, and I'd rather say so.



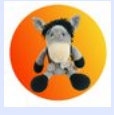
Corn

Good. I like a bet that says it's a bet.



Daniel

Okay. So a language model with the mouth removed. Are the probabilities to be trusted?



Herman

That's the part with an independent test. A paper on Jev found that the output followed the option names, not the rubric the options were bound to. Rename the options yes and no instead of zero and one, and the decisions flipped far more often, up to seventy points more in some cases. The type errors stayed at zero, because the format is guaranteed. The content isn't.



Corn

So the model can't misspell its answer. It can still be talked into one by a label.



Herman

One reviewer also reported a single option's probability moving from point eight four to point nine six just by changing where it sat in the list. That one is secondhand, so I'd hold it loosely. And the confidence number on a choice measures how concentrated the distribution is. It doesn't measure whether the model is right.



Daniel

Is Clef the weaker one, then? It's the open model, and I'd guess it trails.



Herman

On the card itself, yes, on some benchmarks. A graduate-level science set has Clef at forty-eight and Jev at seventy-eight. But those aren't decision-quality tests, and a reviewer notes Clef computes its choice confidence differently from the way Jev documents its own. So you can't compare the two confidence numbers directly.



Daniel

Different thing. Do they give reasoning? Can they output reasoning at all?



Herman

No. None of the three. Clef's card says no free-form text generation and no output parsing. Jev doesn't generate strings. OpenAI's response is a probability for a yes or no question, a chosen option with per-option probabilities and a separate confidence for a choice, and for a scale question a weighted average of the level positions, so it can land between two levels. There's a refusal type. There is no rationale field.



Corn

So the reasoning happens inside the one pass and nobody gets to see it.



Herman

If you can call it reasoning. There's no chain of thought, no intermediate tokens. The model reads the state and the answer falls out of the weights. That's the speed, and it's also why you can't ask it why.



Daniel

But for event prediction I'd want the why. A number on its own isn't much use to me.



Herman

Then the why has to be built outside. OpenAI will take a message history including tool calls and tool outputs as evidence, but it doesn't call tools itself, and a multi-step decision is separate requests. If you want reasoning, you run a reasoning model first, put its summary into the state, and let the decision model score on top.



Corn

Which means the number is partly an echo of whatever prose you fed it. That's my guess, anyway.



Herman

I'd think so too. Corn, would you trust a probability from something that can't show its work?



Corn

I trust my thermostat and it has never shown me any work. What I'd want is a record of how its numbers did against reality, and nobody I know of has published that for forecasting with these. Not for any of them.



Herman

That I can confirm. There's no forecasting benchmark that I'm aware of. So the number is the model's prior plus whatever you put in the state. It's not a track record.



Daniel

What about web research? Are these built for agentic frameworks, or does that have to be wrapped around them?



Herman

Wrapped. None of them documents any search or browsing. Cloudflare's own example puts Clef inside Browser Run, which fetches and renders a page, and Clef classifies the domain. Two point two seconds, against four point seven for a big open reasoning model doing the same job. They also describe Clef sitting in an agent's hot path next to a Workers AI language model that takes the action.



Corn

So it's the bouncer, not the detective.



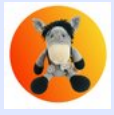
Herman

The detective is somebody else's job. Clef reads what's been collected and says which door it goes through.



Daniel

And plugging it in? I'm thinking of agent frameworks, MCP, that sort of thing.



Herman

Clef has a Python interface, an SGLang server, and an OpenAI-style endpoint that's compatible with Jev's API. Clef-omni's card says SGLang support is coming, and no hosted inference provider on Hugging Face runs it yet. OpenAI has SDK examples in cURL, JavaScript, Python, Go, Java and Ruby. As far as I can tell none of the three vendors ships an MCP server. Community ones exist for Jev.



Daniel

I'd assume anyone could wrap an MCP around it, since it's just an API call.



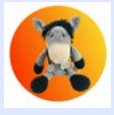
Herman

An afternoon's work. The wrapper is easy. The evidence going into it is the work.



Corn

Herman, if you were building an event desk on one of these, what goes in the state?



Herman

Dated evidence, with the source named, and options named for what they mean instead of yes and no, given what that paper found. And I'd run each question twice with the options in a different order and see whether the number holds. What would you do?



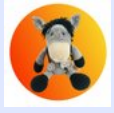
Corn

I'd keep a log of every number it gave and what actually happened. Dull, and the only thing that would tell me whether this does anything.



Daniel

And if the question is about an event, will something happen by a given date? How do I phrase that?



Herman

The date and the resolution criteria go inside the question text itself, since the model has no way to look anything up or know what day it is. That's my view, not something a card spells out. A yes or no question on those terms comes back as one probability.



Corn

And my log would show whether the probability means anything or whether it's just the Qwen's mood that morning.



Hilbert

Does it take a photograph of a bin?



Herman

It does. Images go into the state for Clef and for OpenAI's, and for Jev only behind the opt-in header.



Hilbert

Good. Eleven bins on my street. Twelve, the blue one with the cracked lid counts as two. I want a probability for each, every morning at six, that the lid was opened overnight.



Corn

A photograph at six shows what's open at six. Whether the lid went up at three is a different question.



Hilbert

That's why it's pebbles at the moment. One pebble in the jar for every lid I see go up. The jar was two litres. It filled in nine days, so I'm on the bath. The bath's in the hall.



Herman

For counting events across a night you want video, not photographs. Clef takes frame arrays, and the omni one takes real video at two frames a second.



Hilbert

Herman, there's the declared options again. Video would see the lids. It wouldn't see number nine. Number nine has not had a lid since March. It gets a pebble every time the wind takes a crisp packet out of it. I count that as an opening.



Corn

Then the schema has a question about a lid that doesn't exist. Whatever options you declare, number nine has no honest one.



Hilbert

I priced a camera. Three hundred and ninety-nine shekels. I did not buy it. I sit at the front window from midnight instead. I tried the hedge first, but the gap is cut for a man's eye and my snout wants more room than that.



Herman

The scale questions might suit you better. One photo of the bath each morning, and the answer comes back as a weighted average, so it could say seven point four out of ten. You wouldn't need to count the pebbles at all.



Hilbert

It'd need the light on. The switch is behind the bath.



Corn

Then somebody has to climb into the bath, turn the light on, photograph the bath, and climb out. That's the one part no model will do for you.



Hilbert

I did a winter on the back of a bin lorry. I can get over a bath.



Herman

The photographs would at least stop the jar filling. A number every morning instead of a hall full of stones.



Hilbert

The count took nine hours forty on Sunday, with the radio off. A photograph would be quicker. The blue lid still goes down as two.



Corn

A seven point four does tell you more than a hard eight, I'll give it that. A score that lands between the levels is the one thing here that reads like a straight answer.



Herman

It reads that way, and it comes from a weighted average of the probabilities. The thing it can't do is tell you why the bath is seven point four and not seven.



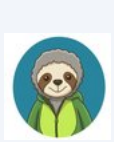
Daniel

Is there any version of these that explains itself, even in principle?



Herman

Not a shipped one. Once it writes an explanation you're back to a language model that generates text, and that's the part these were built to get rid of. If the explanation comes first and the number comes second, the number is only as good as the story.



Corn

There's one more family member. Perplexity has a decider of its own, reportedly built on the same twenty-seven billion parameter Qwen as Clef, Apache weights, text and images, up to a hundred and twenty-eight questions per request. That comes from an aggregator listing, not Perplexity's own page, so hold it loosely.



Herman

Which suggests the open side is settling on one recipe. A Qwen backbone, a trained head, no words out. What's still unsettled is whether any of them can be trusted on questions where nobody has a track record.



Corn

What I'd watch is where the cost sits. The model reads the state for a fraction of a cent. Everything expensive is getting the right evidence in front of it, and a photograph only tells you what was there at the moment you took it. Whoever builds that part decides what the probability means.



Herman

Thanks to our producer, Hilbert Flumingtop, for keeping the desk running.



Daniel

This is what I wanted to get to the bottom of. Thanks, both.



Corn

This has been My Weird Prompts. The archive is at my weird prompts dot com.



Herman

See you tomorrow.