Episode Book

EPISODE #6147

Hebrew AI: Ears, Mouth, and Brain

PUBLISHED	19 October 2026 at 20:00 UTC
DURATION	26:28
LISTEN	myweirdprompts.com/episode/hebrew-ai-speech-language-models-75710/
HOSTS	Corn, Herman, Larry, Hilbert
SCRIPT MODEL	anthropic/claude-sonnet-5.5
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.10
BOOK GENERATED	20 October 2026 at 07:46 UTC

SYNOPSIS

Can you run Hebrew AI entirely offline? It turns out the answer splits three ways — ears, mouth, and brain — and each one has a different winner. Open Hebrew speech recognition, built on Knesset recordings, is beating Google, Amazon, and OpenAI on a public leaderboard. Hebrew text-to-speech is dominated by closed models, with one clever phoneme project changing the math. And Hebrew-native language models exist, but the details matter. This episode walks through what's actually available right now, open and closed, and how good it really is.

TRANSCRIPT



Corn

Does anyone actually build AI for Hebrew, or are we all watching English models do an impression of it?



Herman

Ears, mouth or brain? The answer is different for each, and that's the fun of it.



Corn

Hannah asked it from the ears side. She's started using Whisperian, the offline transcription app Daniel has recommended before, the one he likes because it keeps transcription separate from post-processing. She was looking at his English and Hebrew mixed-speech app and asked whether there's a Hebrew speech-to-text model at all. Daniel gave her the reasons the big multilingual models struggle with Hebrew, then admitted he's nearly sure he's seen open-source Hebrew speech recognition, more sure about Hebrew text-to-speech, and he knows there are Hebrew-native language models. But he wouldn't swear to any of it. So what exists as of now, open and closed, and how good is it?



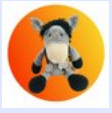
Herman

His hunches are right on all three. Open Hebrew speech recognition exists, Hebrew speech synthesis exists, Hebrew language models exist. What I like is that each one comes with a twist he didn't mention, and the twists are different in each modality.



Corn

Ears first, then.



Herman

Ears first. The name to know is ivrit.ai. It's an Israeli project, and its flagship is a fine-tune of Whisper large version three turbo, Apache two point oh licence, trained in April of last year. The training data is mostly the Knesset plenum recordings, about four thousand seven hundred hours of them, plus a few hundred hours of crowd-transcribed audio.



Corn

Four thousand seven hundred hours of the Knesset. Half a year of people interrupting each other.



Herman

Which is a surprisingly good training set, actually, because it's long, it's real speech, and it has every register from formal to furious. Now, how good it is. The organisation runs a public Hebrew transcription leaderboard, updated this month, that averages word error rate across five test sets. Top of the board is an ivrit.ai build at about eight point eight percent. Second is the turbo build at about nine percent. Third is Soniox, a commercial service, at about nine point nine, and that's the best commercial entry.



Corn

A small open fine-tune beating the paid services. I'd have bet the other way.



Herman

Most people would. Now look at what's underneath. Amazon Transcribe is around eleven and a half. Plain Whisper large version three, the thing everybody actually runs, is about fourteen and a half. Deepgram's Nova three is about fourteen point six.



Corn

So fine-tuning on Hebrew roughly halves the error of the model it started from?



Herman

Close to it, going from the plain large version three number to the ivrit.ai one. It's the whole story in one comparison, really. Scale didn't win. Hebrew data won.



Corn

Hebrew data over scale. I'd put that on a mug.



Herman

The data being public is half the reason this works, by the way. The Knesset recordings, the crowd-transcribed sets, the evaluation sets, the WhatsApp-style one, they're all published as datasets. The original paper from twenty twenty-three counted over three thousand three hundred hours of Hebrew speech, and the collection has grown since. So anyone can fine-tune from the same starting point, and some people have.



Corn

Public data, public models, a leaderboard they run themselves. That's a very tidy little ecosystem for a language with a small market.



Herman

Small market, very motivated people. I think that's the whole recipe.



Corn

Where do the big names land?



Herman

Further down. ElevenLabs Scribe version one is about sixteen and a half. OpenAI's GPT four oh transcribe is about eighteen, and on one of the test sets, the Kan broadcast one, it hits thirty-nine percent. The mini version is about twenty-three. Google Speech is twenty-eight point six.



Corn

Google Speech at nearly thirty percent. For a language with that many native speakers on their doorstep.



Herman

Bottom of the table is NVIDIA's Nemotron three point five streaming model, used zero-shot, so with no Hebrew training, at about thirty-five and a half. Deepdub, the Israeli voice company, took that same model family and fine-tuned it on its own Hebrew data,



Herman (continued)

and it says the result is good. That's their number, not an independent one.



Corn

Thirty-five percent. That's less transcription than a rough guess.



Herman

There are caveats on the table, and they matter. The newer commercial models aren't in that table: Scribe version two, the newer Gemini audio, Chirp three. And two of the test sets are ivrit.ai's own, so a home-advantage effect is possible. I'd say the ranking among the open models is solid and the exact gap to the commercial ones is softer than it looks.



Corn

ElevenLabs advertises Hebrew speech recognition, doesn't it? I remember a number that looked very good.



Herman

Their own Hebrew page quotes about three percent on the FLEURS set. The same page also shows fifteen. From what I can tell, the page disagrees with itself, so I wouldn't build anything on it.



Corn

Three or fifteen. I'd go with the one that matches the leaderboard.



Herman

Now the part Hannah cares about, which is running it offline. ivrit.ai publishes faster-whisper builds and an ONNX build of the turbo model, both on its Hugging Face page. I'd have guessed a ggml build too, for the llama-dot-cpp style runtimes, but the only ones I saw on the page were for the Yiddish models. So I can't confirm a Hebrew ggml build.



Corn

Hold on, Yiddish?



Herman

That's the twist. There's a family called yi-whisper in the same organisation. It sounds like it could be a Hebrew thing. It's Yiddish, a different language with its own training data. Anyone grabbing the wrong checkpoint because the name looked right is going to have a strange afternoon.



Corn

Yi, as in the thing you say when your transcript comes out in the wrong language.



Herman

There's also a small one worth knowing about, a streaming model called zipformer, sixty-six million parameters. It runs in a browser through WebGPU, it's CC-BY licensed, and it scores about five percent on one of the test sets and nearly nine on the WhatsApp-style one. Averaged across the leaderboard it's eleven point one, which still beats Amazon.



Corn

Sixty-six million. That's smaller than most phone keyboards' prediction models, I'd guess.



Herman

I wouldn't swear to that comparison, Corn. But it was originally built by one developer, Yakov Kolani, and ivrit.ai adopted it. Which is a nice reminder of how much of this scene is a handful of people.



Corn

Daniel's one of them, in a small way. His Whisper-Hebrish took the turbo model from about seventeen percent to about six on his own mixed-language set. That set is small, so it's his number for his voice, not a benchmark.



Herman

His Parakeet fine-tune for code-switching was the same instinct on a different base model. And that's relevant, because NVIDIA does not have an official Hebrew Parakeet. Parakeet version three, as far as I know, is a European-language model. Hebrew isn't in it.



Corn

Which matters, because Parakeet is what Whisperian's local engine runs.



Herman

It does, along with the cloud options through your own API keys. Keep that in your pocket for the end.



Corn

Do you think the home-advantage worry is serious, Herman? Or is that me being a sloth about it?



Herman

Less serious than it sounds, I think. The Knesset data and the eval sets are different audio. But I'd want to see someone outside ivrit.ai run the same test. There's a model called GigaAM-He whose card has its own side-by-side where the ivrit turbo model still comes out ahead. I only saw that through a search snippet, though, so treat it as a hint.

L

Larry

Tired of buying your Hebrew speech model one download at a time like some kind of animal? Introducing Weights Monthly, the only subscription that delivers a freshly trained model to your doorstep on a hand-labelled USB stick, every single month, whether your computer has a USB port or not. Choose your tier. Tiny is a gentle nudge toward better transcription. Turbo is for the professional who dictates at speed and loses things. And our top tier, Large Plus Plus Plus, includes a personal courier who stands beside you and reads your own mumbling back to you in a calmer voice. Each box ships with a free vowel sticker, colour not guaranteed. Cancellation is simple. Simply submit your request in triplicate, in handwriting, in a language you do not speak, to a post office box we move quarterly. Cancellations are processed on the thirtieth of February. Refunds are issued in the form of additional boxes. Weights Monthly. You'll never stop receiving it.



Corn

...so what I'd actually want to know is whether the ivrit.ai number survives a bad microphone.



Herman

The WhatsApp set is exactly that. Noisy, compressed, people talking over each other, and every system gets worse on it, which is why it's the one I'd watch.



Corn

Mouth, then. I'm braced for this being a shorter and sadder story.



Herman

Shorter, and not sad. Just a different shape. In speech recognition, the open model is on top. In Hebrew speech synthesis, the top of the arena is all closed models.



Corn

There's an arena?



Herman

ivrit.ai runs a Hebrew TTS arena, blind listening votes turned into Elo ratings. I read it today. In first place is Google's Gemini three point one Flash TTS at about fifteen hundred and eighty-seven. Second is Deepdub, an Israeli company, at about fifteen seventy-four. Third is ElevenLabs version three at fifteen fifty-two.



Corn

You read that table out like a racing commentator.



Herman

Someone has to. Deepdub being Israeli makes second place feel like a home game, but it's a blind vote. Fourth is Soniox, fifth is OpenAI's realtime model at about fifteen hundred. And then the first entry that isn't labelled proprietary is ninth, Blue version two, at about fourteen seventy-four. That's roughly a hundred points behind the leader. I'm inferring it's open from the labelling, not from the page saying so, and the vote counts aren't shown, so I'd read it as a general shape.



Corn

A hundred points of Elo. In human terms?



Herman

Listeners would pick the leader most of the time, but not overwhelmingly. And ninth place there is still above OpenAI's older TTS models, which sit around fourteen forty-five.



Corn

Ninth out of a field with Google and ElevenLabs in it. I'd call that a podium of sorts.



Herman

The open side rides on one idea. A project called Phonikud. It's a Hebrew grapheme-to-phoneme system, meaning it turns written Hebrew into a precise pronunciation, full IPA with stress marks, built as adaptors on top of a diacritizer. It's being presented at Interspeech this year. The code is CC-BY, though the training datasets are non-commercial, which matters if you want to build a product on it.



Corn

So it solves the problem Daniel described in the prompt, upstream of the voice.



Herman

Yes, and that's the neat bit. The ambiguity, that the same written letters can be several different words, is resolved before the voice model ever sees the text. The claim from the authors is that a small local voice model fed Phonikud phonemes gets close to large proprietary systems, and that you can fine-tune an English model on about an hour of Hebrew if you give it those phonemes. That's their claim, mind you. A follow-up called ReNikud came in June, and I've only seen its abstract.



Corn

An hour of Hebrew. I've napped longer than that.



Herman

And you'd have made less progress. The usable models built on this line are LightBlue, which is Hebrew-first with English and MIT licensed, and Blue version three, which is Hebrew-first multilingual, also MIT. LightBlue's developer claims twelve hundred and sixty times real time on an RTX three thousand ninety, and thirty-five times real time on a plain CPU. Self-reported. Blue version three is a flow-matching model at forty-four point one kilohertz, and the card has no quality numbers at all.



Corn

Thirty-five times real time on a CPU. So I could synthesise this episode while I'm reaching for my tea.



Herman

With a long reach. The release dates aren't clear to me, LightBlue looks to be around eight months old. There's also a sibling called blue-yi, and yes, that's Yiddish again.



Corn

Same trap, new directory.



Herman

One more thing about the arena. ivrit.ai published the preference data from all those blind votes as its own dataset, updated just a few days ago. So the open Hebrew TTS people get a free scoreboard and the raw votes to learn from. That's a loop the closed vendors don't get to use in public.



Corn

Voting in public, with the ballots published. Closed vendors can't copy that without opening their ballot box.



Herman

What would you use, Corn? If you were building something that speaks Hebrew tomorrow.



Corn

Honestly, for a personal project I'd start with the open one. Local, free, and I can fix it myself when it mispronounces something. If it were going out to customers who'll judge the voice in four seconds, I'd pay for one of the top four.



Herman

Same split I'd draw. And Azure has Hebrew neural voices, Hila and Avri, which I've only seen listed on a third-party page, not Microsoft's own. So I'd check that before leaning on it.



Corn

Brain, then. Everyone's favourite.



Herman

This one has two entrants, and they're a little bit rivals. First, DictaLM three point zero, from DICTA, an Israeli nonprofit. It launched in December, with the paper out in February. Three sizes: twenty-four billion built on Mistral Small, twelve billion on NVIDIA's Nemotron Nano, a hybrid state-space design, and one point seven billion on Qwen three. Each



Herman (continued)

comes in base and instruct or thinking flavours, with tool calling and a sixty-five thousand token context.



Corn

Which is the one I can try without owning a datacentre?



Herman

The twenty-four billion has a hosted chat page run by DICTA. There are quantised versions too, listed as updated in May, which should bring it within reach of a decent graphics card. The training recipe is about a hundred billion Hebrew tokens and thirty billion English, mixed roughly three to one, so it doesn't forget English. Press coverage says Apache two point oh licence, but I haven't seen the model card say it.



Corn

How good is it? I'm assuming there's a table with a lot of bold numbers.



Herman

There is, and it's honest about what it measured. On Hebrew chat tasks, the twenty-four billion thinking model gets about fifty-seven on summarisation, where Gemma three twenty-seven billion gets about forty-five and Llama three point three seventy billion gets about thirty-eight. Translation, trivia, same direction. Then there's a vowel-pointing task. DictaLM scores about eighty-seven. Llama's seventy billion scores about four.



Corn

Four. Out of a hundred. That's the model knowing it's Hebrew but declining to try.



Herman

That is how task-specific it gets. A model can read your Hebrew perfectly well and be hopeless at one operation on it. And it isn't a clean sweep. On the Winograd-style reasoning set, Llama seventy billion wins, about eighty-three against about seventy-eight.



Corn

Good. I trust a table more when the home team loses one.



Herman

The small model is the part that surprised me. The one point seven billion DictaLM scores about ten on summarisation. Qwen three at the same size, the model it was built from, scores under one. On translation it's about two against zero.



Corn

Ten out of a hundred is still a bad grade. But it's the difference between a bad grade and not turning up.



Herman

And on the vowel task, fifty-three against three. A model that size fits on a laptop, Corn. Would you run a Hebrew model locally if it were that small, or is it still too weak to bother?



Corn

For cleaning up a transcript, maybe, if the cleanup is simple. For anything I'd want to trust with a nuance, I'd go up to the twelve or the twenty-four. Small models are for tidying, not thinking.



Herman

That's roughly where I'd land.



Corn

Fair. I'd still like to know who the table left out.



Herman

Here's what the table doesn't have, though. Only open models are compared. Gemini two point five Pro appears as a reference for win-rates, and GPT four oh is the judge, but there's no row where a frontier closed model sits next to DictaLM on the same benchmark. I couldn't find that head-to-head anywhere, and I'm not going to invent a ranking.



Corn

So if someone says the closed frontier models beat it, or that it beats them, that's a feeling, not a measurement.



Herman

Right now, yes. The second entrant is Hebatron, from a project called HebArabNlp. Thirty-one point six billion parameters total but only about three billion active at a time, a mixture-of-experts design, built on NVIDIA's Nemotron three Nano. It was made by PwC Israel and MAFAT, the national NLP program, with AWS. Apache two point oh. About ninety-nine billion training tokens, three and a half billion of them colloquial Hebrew, and around two million instruction samples.



Corn

A government-backed rival to the nonprofit. That's a sentence I didn't expect to say out loud today.



Herman

The authors claim a Hebrew reasoning average of seventy-three point eight against DictaLM's sixty-eight point nine, and about nine times the throughput, which follows from only three billion active parameters. Those are their own figures, from a paper in May. I haven't seen an outside group repeat them.



Corn

The two are on the same leaderboard, I take it, since MAFAT and DICTA both back it.



Herman

They do co-run a Hebrew LLM leaderboard. Tasks like natural language inference, question answering, sentiment, Winograd, translation and Israeli trivia. I couldn't load the rows, so I can't tell you whether any closed models are on it.



Corn

Daniel said he knew of Hebrew-native LLMs. I assume he meant these.



Herman

Probably. One more name for completeness. AI twenty-one's Jamba one point five officially supports Hebrew, and their CEO told Calcalist that Hebrew is not a significant market. It's now a legacy model.



Corn

Not a significant market, said an Israeli company. I'd have chosen different words in an interview.



Herman

The practical backdrop for all of it is the tokenisation tax. One blog post from July measures Hebrew at about one and a third times the tokens of English on OpenAI's newer tokenizer, close to two times on Claude's, and around four times on the old GPT four one. Blog-level numbers, not rigorous. Whether DictaLM's tokenizer fixes that, the paper doesn't say.



Hilbert

...and Pinchas says it's the nine-metre one, same as last month, and I say Pinchas, I paid for the nine-metre one in March. He says that was the other nine-metre one. He keeps two.



Herman

Two nine-metre cables, same shop?



Hilbert

Same counter. Hillel Street, two doors from the pharmacy. Six and a half months, by the way. The Knesset hours. Not half a year. A hundred and ninety-six days if you ran it without stopping.



Corn

You counted that.



Hilbert

Before the first break. The rule is the microphone sits in the kitchen and the laptop sits in the stairwell, on the third step, and between them is nine metres of cable, no more, no less. Eight is too close. The Hebrew comes out soft. Pinchas has sold me cable for nineteen years and he has never once sold me a ten.



Herman

Nine metres from the kitchen to the third step. That's a long run for a microphone signal, but the loss is small.



Hilbert

Herman, you said leaderboard eleven times. I cut two. The step belongs to Pinchas's sister-in-law. Forty shekels a day, and the laptop has its own tenancy, with a clause about Purim.



Corn

Does she check the cable?



Hilbert

Every morning at seven. She measures it with a tailor's tape. Corn, you're peaking on the b's, I'll ride it down. Anyway, Pinchas rents me the cable by the hour. Nine shekels. He keeps it in the shop between uses, in a biscuit tin with my name on it, next to the offcuts of every cable he's ever cut for me, which he also charges for. Eleven years of offcuts. They're in the tin.



Herman

He bills you for cable you didn't use.



Hilbert

And for the tin. I record everything with the phone in my shirt pocket, though. That's Pinchas's other rule. He says nobody in the history of speaking has talked into the good microphone.



Corn

A phone in a shirt pocket, at the end of nine metres of cable.



Hilbert

Pinchas says it's on a long lead. I've never seen the lead.



Herman

A shirt pocket is the hard case, Hilbert. Cloth rustle, a mouth at the wrong angle. It's close to the WhatsApp set on the leaderboard, where everyone's error rate gets worse.



Hilbert

Forty minutes to get one clean sentence out of the stairwell. Pinchas says that's the Hebrew's fault. Fair enough, though. A phone doesn't need a cable. I'll tell him. He won't like it.



Corn

That leaves Hannah, who has an app, a microphone and a question.



Herman

Whisperian's local engine is Parakeet, and as far as I know Parakeet is the wrong tool for Hebrew. Its page doesn't mention Hebrew, and it doesn't mention importing your own model either. That's a question for the developer, or a ten-minute test, not something we can answer for her.



Corn

I'd just ask. Developers of small apps tend to answer, and a custom-model option would be a feature worth hearing about.



Herman

What do you think she should do if the answer's no, Corn?



Corn

Use something else for the Hebrew, offline. And since she keeps the transcription step apart from the clean-up step, which is the whole reason Daniel likes the app, she can feed the transcript back in afterwards.



Herman

There are two routes. One is an app that can run the ivrit.ai faster-whisper or ONNX builds locally, which is the accuracy route, around nine percent average error on that leaderboard. The other is the sixty-six million parameter zipformer through the sherpa-onnx runtime, which is the small and light route, around eleven. Both run on her own machine.



Corn

And the cloud options Whisperian offers through API keys?



Herman

OpenAI's transcribe and Deepgram both sit between fourteen and eighteen on the same board, so they'd work, but noticeably worse than the open Hebrew models. For a Hebrew-and-English mix, Daniel's Whisper-Hebrish was built for exactly that, though only measured on his own set. I'd try it before assuming anything.



Corn

So the order is, ask about custom models, then try ivrit.ai, then try Daniel's.



Herman

In that order, and with her own recordings, not a clean test file.



Corn

Hebrew speech recognition got better by feeding a small model the right data, not by waiting for a bigger one. Speech synthesis hasn't flipped that yet, because the best voices are still behind a paywall, and language models are a two-horse race that nobody has measured against the frontier.



Herman

Each modality has a different answer, and that's the useful picture. The open models are close, the closed ones are not extinct, and the evidence for the LLM side is mostly the authors' own tables.



Corn

What stays open for me is whether the next closed model that lands on that Hebrew arena beats the top by a mile or a hundred points.



Herman

That, and whether the next Hebrew model you hear sounds right with the vowels in place. The dots, the vowel marks, are the subject of the next episode, and there's more to them than a letter-to-sound table. The test that counts is your own voice, recorded on your own phone, in your own kitchen, because that's the hard case.



Corn

Thanks to our producer, Hilbert Flumingtop, for keeping the levels sane.



Herman

This has been My Weird Prompts. You can find every episode at my weird prompts dot com.



Corn

See you tomorrow.