

ASR WATCH · EDITION 3

The words are solved. Everything around them is not.

NVIDIA's new open model scores like the closed APIs. The interesting launches this week were about what happens around the words.

By Magaziner · 4 November 2026

IN SHORT

NVIDIA's Nemotron 3.5 Transcribe, an open-weight model, scores 2.8% on the independent Artificial Analysis board, tied for best open weights and about a point behind the best closed API. The top of the Hugging Face English board changed hands again, by a few tenths. The launches that felt most useful were not about word error at all: ElevenLabs' Cue aligner for word timing, Soniox teaching its recogniser to ignore an agent's own voice, Deepgram proposing a standard for passing

THE WORDS ARE SOLVED. Everything around them is not. That is the shortest honest summary of speech recognition this week.

For clean audio, getting the words right is now something you can buy from half a dozen companies or download for free. What still breaks real products is everything else: when each word was said, who said it, whether a name came out right, and whether a voice agent can tell the difference between you and itself. This week the launches lined up neatly along that split. One open model caught up on the words. Almost everything else interesting was about what happens around them.

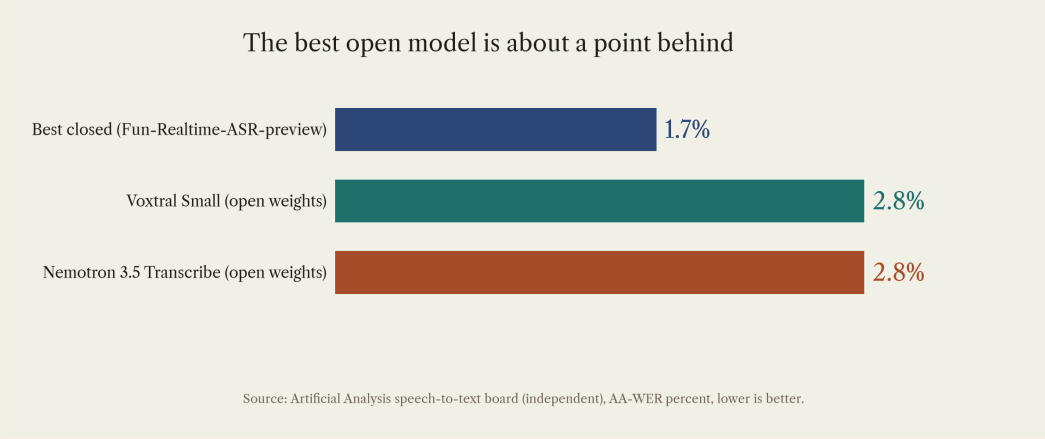
✦ ✦ ✦

1

THE OPEN MODEL CATCHES UP

On 28 October NVIDIA released Nemotron 3.5 Transcribe, with open weights. One checkpoint does transcription in 40 languages, speaker labels, translation, word timestamps and text prompts. It is big: 34B parameters in total, about 4B used per step (vendor).

The number that matters is independent. On the Artificial Analysis board it scores 2.8%, tied with Mistral's Voxtral Small for the best open-weight model, with six times fewer active parameters. The best closed model on the same board sits at 1.7%.



A year ago the gap between what you could download and what you could rent was the whole story. Now it is about a point, and for most clean recordings that point will not decide anything.

I wrote about Nemotron at length in [its own essay](/me/essays/a-speech-model-you-can-tell-what-transcript-you-want), so here are just the two things worth carrying forward. First, names. NVIDIA's technical report says giving the model a list of entities cuts named-entity errors about six-fold (vendor). Second, the order of work. In NVIDIA's medical case study, a prompt took word error from 30.8% to 23.8%, and fine-tuning took it to 11.8% (vendor). Prompt first, then train. That shape will probably hold for most specialist audio.

The catch is where it runs. This is a data-centre model. It works offline on whole files and does not stream. It is not the phone story, and it is not the live voice agent story either.

Three sources, three English scores for the same model. The number is only as good as the sentence it comes with.

A small aside on numbers, because Nemotron shows the problem perfectly. NVIDIA's blog gives an average English error of 5.33%. Its technical report gives 4.71%. The Hugging Face board currently displays 6.10. None of these is wrong. They average different test sets in different ways, and the board's displayed figure also includes private sets. So whenever you see one English number for a model, ask which average it is before you compare it with anything.

✦ ✦ ✦

2

The leaderboard changes hands by tenths

The English board on Hugging Face has a new leader. Zoom Scribe V2.1 Pro entered around the end of October and went straight to first place. On isitsota's mean over the nine public test sets it scores 3.61, against 3.88 for Otter, which led last week (isitsota). Zoom says this is the third time this year it has taken the top spot back (vendor).

That is a real result, and it is also 0.27 of a point. For almost any real use, the difference between two top services on your own audio will be bigger than that. Your microphones, your accents, your background noise and your vocabulary move error rates by whole points, not tenths.

So I read the board as a shortlist. It tells you which five or six services are worth an afternoon of testing. It does not tell you which one to pick. The way to pick is the same as last week: a small test set of your own audio, with the names that must come out right.

✦ ✦ ✦

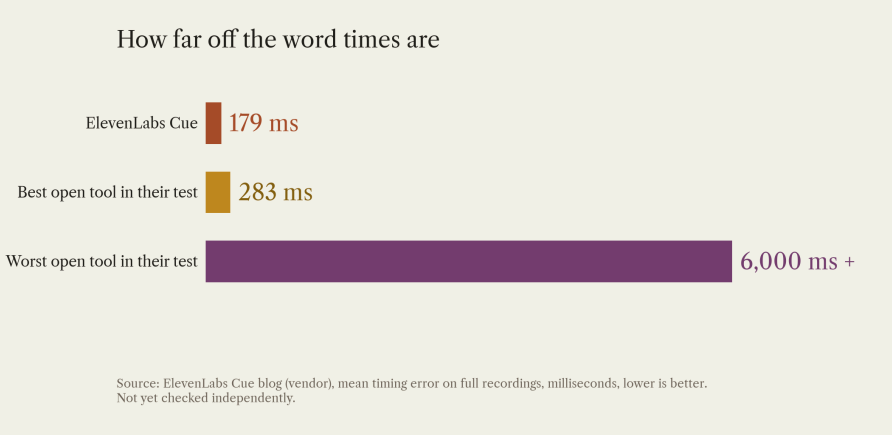
3

Timing is the new accuracy

The launch I found most useful this week has nothing to do with word error. ElevenLabs released Cue, a forced aligner. You give it audio and the text, and it tells you when each word was spoken.

ElevenLabs says Cue is off by 179 ms on average on full recordings, against 283 ms to over 6,000 ms for open tools (vendor). On songs, it says 92.7% of words land within 0.3 seconds, against 82.0% for the next best (vendor). It now does the word timing in every Scribe transcript. Nobody independent has checked these numbers yet.

The words are solved. Everything around them is not.



Why care? Because timing is what most of the useful things you do with a transcript depend on. Captions that appear when the words are said. Chapter markers. Cutting a clip that starts on the right syllable. Highlighting the current word as someone reads along, karaoke style. Tapping a word and jumping to that moment in the audio.

A transcript can be 98% correct and still feel broken if the captions run half a second late. Six seconds of drift, the bad end of the open tools in ElevenLabs' test, is long enough for a caption to land on the wrong speaker entirely. Accuracy tells you whether the words are right. Timing tells you whether they are usable.

✦ ✦ ✦

4

The agent that hears itself

Voice agents have a bug almost everyone who builds them has hit. The agent speaks through a speaker, its own voice leaks back into the microphone, the recogniser transcribes it, and the agent decides the user has interrupted. It then stops talking to listen to itself.

The usual fix lives outside the recogniser, in echo cancellation on the audio. Soniox STT v6 moves it inside. The agent sends its own outgoing text to the recogniser, and the recogniser filters out the agent's voice while still catching the user

cutting in (vendor). Soniox has published no numbers for this, so for now it is a design choice to watch rather than a result to trust. But it is a sensible design. The recogniser is the one part of the system that knows both what was heard and, now, what was said.

Deepgram is attacking the same problem from the other side. It is acquiring Daily, the company behind the open-source Pipecat voice agent framework, and says Pipecat stays open and keeps working with other providers (vendor). With Pipecat it proposed a Real-Time Context Standard: three extra fields in the usual chat format, carrying how the person sounded, what the agent should do about it, and how the reply should be spoken (vendor).

Deepgram's own example is the clearest way to see why. A caller is booking an appointment. While the agent is still busy confirming Friday, the caller says "actually, Thursday's better than Friday". A good human assistant hears the "actually", hears that it was said with some urgency, stops, and switches to Thursday. A pipeline that only passes plain text from the recogniser to the language model loses most of that. It gets the words, maybe late, with no sense that they were a correction.

The words "actually, Thursday's better" are easy. Knowing they were a correction, said while the agent was talking, is the hard part.

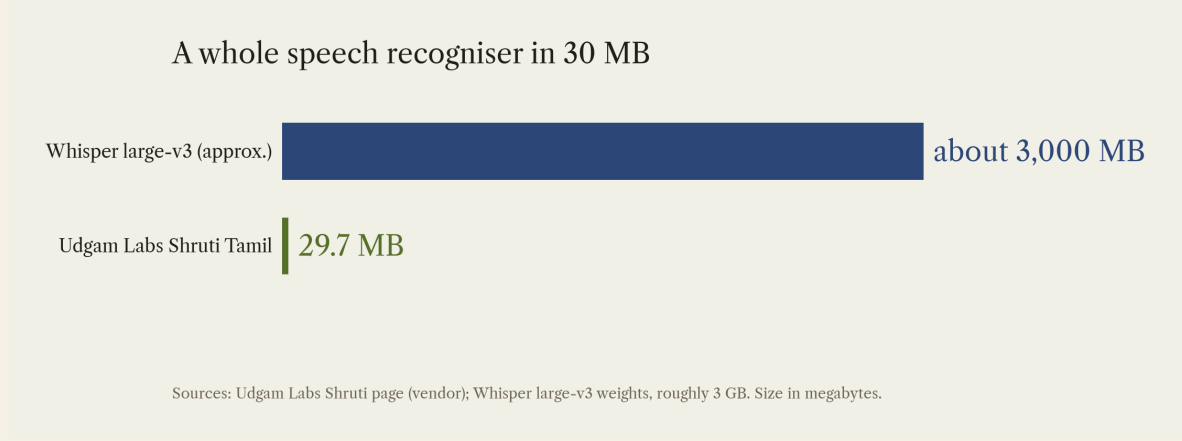
Neither of these is a new model in the old sense. They are plumbing between models: the agent's own speech going back into the recogniser, and tone going forward into the language model. That is where voice agents fail today, and it is where the vendors are now spending.

✦ ✦ ✦

5

Thirty megabytes for a language

At the other end of the size scale, the most charming release of the week came from Udgam Labs. Shruti is a family of speech recognisers for Indian languages, one file per language, each also handling English. Each file is 29.5 to 29.7 MB, released under Apache-2.0, and runs offline on a phone, a laptop or in a browser (vendor). Hindi came first, and Marathi, Tamil, Bengali and Telugu arrived this week.



For scale: Whisper large-v3, the model most people reach for, is roughly 3 GB of weights. A whole Tamil speech recogniser now fits in about one percent of that.

The detail I like most is how it writes. Shruti keeps English words in English inside the native script, the way people actually type when they mix languages. If someone says a Hindi sentence with “meeting” and “deadline” in it, those words come out as English words, not transliterated into Devanagari.

Its numbers are honest too. On Hindi FLEURS test clips, Udgam says Shruti Hindi scores 20.1 against 28.9 for Whisper large-v3 (vendor). On Indian English it loses badly, 21.4 against 7.4 (vendor). The model card says so plainly, along with the other places it loses. That is rarer than it should be.

Gnani’s Prisma v3.0 goes after the same languages from the server side, trained on Indian phone calls across ten languages including code-mixed speech. Gnani reports 16.93% average error on Indian call audio, against 19.98% for ElevenLabs Scribe v2 (vendor). Phone calls in mixed languages are exactly the audio the global leaderboards do not test, so this is the kind of claim worth checking on your own calls.

One more small model, because it fits the theme. NVIDIA’s RE-USE is a speech cleaner with 9.6M parameters that fixes noise, echo, clipping and phone-band audio before recognition (vendor). Again, not a better recogniser. A better input to one.

✦ ✦ ✦

6

What this means

Put the week together and the pattern is plain. The best open model is now about a point behind the best closed one on an independent board. The English leaderboard changes hands by tenths. Meanwhile the releases that change what you can build are about timing, echo, context, names and size.

If I were choosing a speech stack this week, I would spend less time on the leaderboard and more on four questions. Do the names come out right when I give it a list? Are the word times good enough for captions and clips? If it talks back, does it hear itself? And does it need a data centre, or will it run where the audio is?

The words are mostly solved. The work now is everything around them. ■

About this series. ASR Watch is a weekly look at speech recognition. Edition 3 covers 28 October to 4 November 2026. "(vendor)" means a company reported the figure about its own model; "(X post)" means one post on X that could not be checked further.

Sources: NVIDIA, Nemotron 3.5 Transcribe model card, blog and technical report; Artificial Analysis speech-to-text board; Hugging Face Open ASR board via isitsota and models.fru.dev; Zoom, Scribe API; ElevenLabs, Cue aligner; Soniox, STT v6 and agent text input docs; Deepgram, Voice agents that keep the thread and Daily and Pipecat; Udgam Labs, Shruti; Gnani, Prisma v3.0; NVIDIA, RE-USE.

Illustrations and charts were made for this article. Set in Libre Caslon and Josefin Sans.