NVIDIA Nemotron On-Device Speech Recognition

Nemotron is Hedy's new on-device speech recognition engine: fully local, more accurate than earlier local engines, and it labels who said what.

Published by Julian Pscheid · June 23, 2026

Read this article online: https://www.hedy.ai/post/nemotron-on-device-speech-engine/

Six colleagues of different ethnicities in an active discussion around a conference table, one speaking and gesturing while the others listen

Quick answer Nemotron is a new on-device speech recognition engine in Hedy, built on NVIDIA's Nemotron speech model. It runs entirely on your device, transcribes more accurately than earlier local engines, and labels who said what so a multi-person conversation is easy to read afterward. You'll find it under Settings, in the Speech & AI section, in English and Multilingual versions. It joins Whisper and Parakeet and replaces Parakeet going forward.

Two problems have followed local transcription around for a long time. Accuracy lagged behind the cloud, and a transcript of a multi-person conversation came back as one undivided block of text that was hard to read later. Nemotron helps with both, and it runs entirely on your device.

Nemotron is NVIDIA's on-device speech recognition model (https://huggingface.co/nvidia/nemotron-3.5-asr-streaming-0.6b) . NVIDIA released it in June 2026, and Hedy is among the first apps to bring it into a shipping product rather than a demo. It now sits alongside Whisper and Parakeet as one of the speech engines you can choose inside Hedy.

A clear step up in local accuracy

Local transcription has always been the right choice for people who'd rather keep their conversations on their own hardware. The catch was accuracy: on-device models trailed the cloud, sometimes by enough to notice. Nemotron narrows that gap. For everyday meetings and calls, the transcript it produces is a clear step up from what local models managed before, with nothing leaving your device to get there.

If you want the wider picture of how on-device transcription fits into Hedy, see our Local AI for Meetings overview (/post/local-ai-meetings-hedy-3-2/) and the engineering deep-dive on Hedy 3.2 (/post/local-ai-engineering-deep-dive-hedy-3-2/) .

It tells your speakers apart

The bigger change is what a multi-person transcript looks like afterward. Older local transcription handed back one wall of text with no sense of who was talking. Nemotron separates the speakers and labels them: Speaker 1, Speaker 2, Speaker 3, and so on. A back-and-forth meeting reads like a back-and-forth meeting.

Where those labels appear depends on your platform:

• On iPhone and Mac , the labels appear live as people talk and the conversation moves between them.

• On Windows and Android , they're added at the end of the session during processing, so you'll see them once your transcript is ready rather than in real time.

Two versions: English and Multilingual

Nemotron comes in two flavors, and you pick the one that matches how you work. The English version is tuned for English-only conversations. The Multilingual version handles a wide range of languages from a single model, which is the one to choose if your meetings move between languages or aren't in English.

You'll find both under Settings , in the Speech & AI section, next to Whisper. Switching engines takes a couple of taps. For a side-by-side of every option, see how Hedy's speech recognition engines compare (/help/speech-recognition-providers-in-hedy/) .

Moving on from Parakeet

Nemotron replaces Parakeet. It does the same job better, with more accurate transcription. If you're on Parakeet today, switch to Nemotron. Parakeet is being retired, and Nemotron is where that work continues. Whisper stays right where it is as the most broadly compatible engine, and a good default if you're on an older device or simply prefer it.

What you need to run it

On iPhone and iPad, Nemotron needs an iPhone 12 or newer , or an iPad from that generation onward. Older Apple devices won't see the option. It also runs on Mac, Windows, and Android.

Open Settings !’ Speech & AI , pick Nemotron in English or Multilingual, and your next session is transcribed on your device with the speakers labeled. Easier to read afterward, and none of it left your device to get there.

Hedy AI · Live AI Coaching for Important Conversations

Try Hedy free: https://www.hedy.ai/downloads/

https://www.hedy.ai/post/nemotron-on-device-speech-engine/