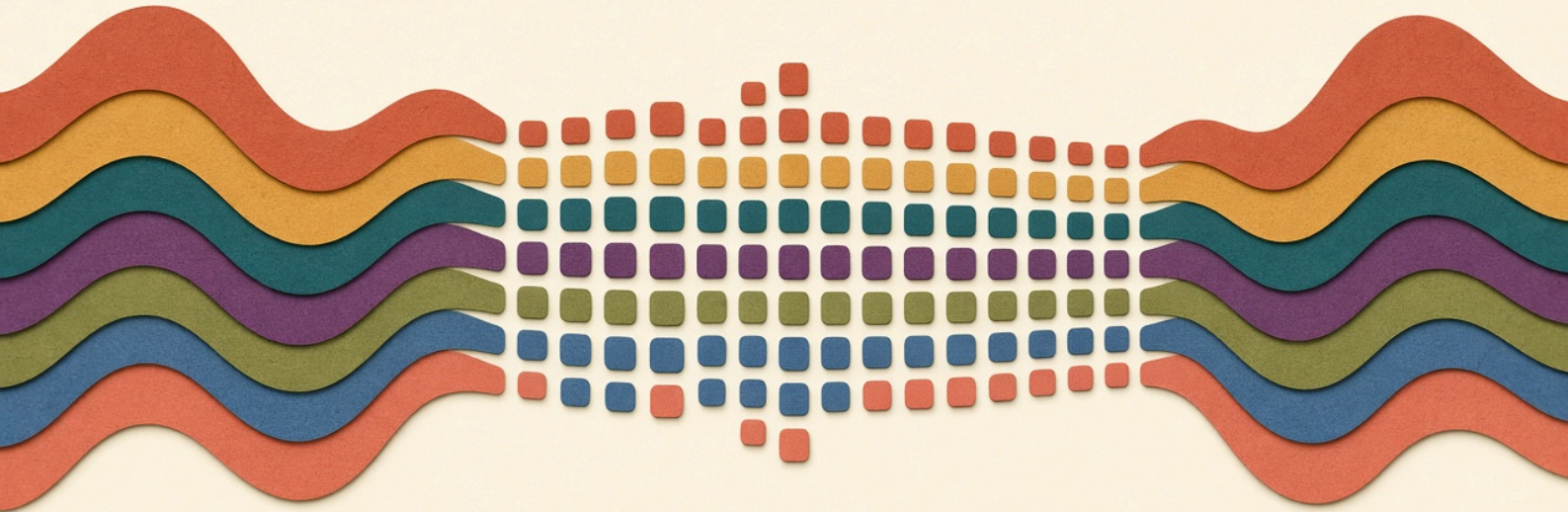




ASR WATCH · EDITION 1

Accuracy is nearly settled in speech recognition

The best open and closed transcription models are now close on accuracy. What separates them is everything around the words: timestamps, speed, size, price and licence.

BY MAGAZINER · OCTOBER 2026

IN SHORT

On the main English leaderboard, the best closed transcription services and the best open models are now within about one to one and a half points of word error rate, and small open models under a billion parameters sit close behind the 2B ones while running much faster. So the useful question for anyone choosing a speech model is no longer which one is most accurate. It's which one gives you the timestamps, speaker

HERE IS THE STATE of speech recognition in one sentence: on English, the gap between the best paid services and the best free models is now about one point of word error rate. On the Hugging Face Open ASR Leaderboard, the top closed system gets about 4 words in 100 wrong, and good open models get between 5 and 6.

My argument in this first edition is that accuracy has stopped being the main thing to choose on. Once the error rates are this close, the decision moves to everything around the words. Do you get word timestamps and speaker labels? How fast does the first word arrive? Does it fit on a phone? What does an hour cost, and can you ship it under the licence? The rest of this edition reads the past few months of news through that lens.

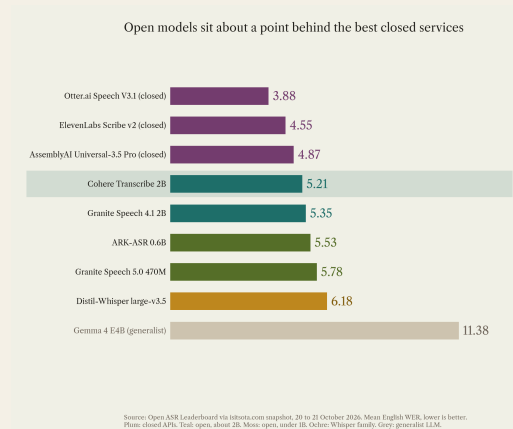


1

THE LEADERBOARD HAS FLATTENED

The Open ASR Leaderboard averages word error rate (WER, the share of words a model gets wrong) over a set of public English test sets. Its methodology paper now covers 86 systems on 12 datasets.

The top of the table is mostly commercial. In a recent snapshot, Otter.ai Speech V3.1 scores 3.88, ElevenLabs Scribe v2 4.55 and AssemblyAI Universal-3.5 Pro 4.87. The best open models follow closely: Cohere Transcribe at 5.21, IBM Granite Speech 4.1 2B at 5.35, and Granite Speech 5.0 470M at 5.78. Distil-Whisper large-v3.5, a compressed Whisper, is at 6.18, and the leaderboard's top open models no longer build on Whisper.



Two things stand out. First, the open models trail the best closed APIs by about one to one and a half points. Second, size barely matters near the top. Granite’s 470M model is less than half a point behind its 2B sibling. Other sub-1B entries, such as Edgeo’s ARK-ASR 0.6B at 5.53, are in the same band.

There is a caveat. Many of the closed-API rows are recent vendor submissions, and these are averages over read and broadcast English. Your audio is not the test set. But the shape is clear enough: on clean English, the accuracy race has bunched up.

The paper also explains why speed differs so much more than accuracy. Models with transformer decoders on top of Conformer encoders get the lowest error, while CTC and TDT decoders (simpler designs that emit text in one pass) get the highest throughput. Cohere Transcribe runs at about 525 times real time in the paper’s table, Canary-Qwen 2.5B at 418 and Qwen3-ASR 1.7B at 148. Those are differences of several times, against accuracy differences of a fraction of a point.



2

MORE ACCURATE, WITH FEWER FEATURES

The clearest example of the new trade-off came from OpenAI. In late July it released [gpt-transcribe](#) and [gpt-live-transcribe](#), one for files and one for live audio. Both accept free-form context, keywords and language hints. OpenAI reports that the

live model's error rate on a real-world audio benchmark fell from 11.65 percent to 9.60 percent against its predecessor, and that adding context raised semantic accuracy on its own benchmark from 38.5 to 44.6 percent. These are OpenAI's own figures.

A month later it set a date. Whisper-1, the gpt-4o-transcribe family and the translations endpoint shut down on 26 February 2027.

The catch is in OpenAI's own migration guide. The new models produce no SRT or VTT subtitles, no word timestamps and no speaker diarization. For those, the guide says to keep using whisper-1 or gpt-4o-transcribe-diarize until they are switched off.

*The newer model is more accurate and does less.
That's the trade-off in miniature.*

If you build subtitles, search over audio, or anything that needs to know who spoke when, the better model is not a drop-in replacement. You have about four months to find one that keeps those features, or to add them back with a separate alignment and diarization step.

Others are pairing accuracy with features. Google's Gemini 3.5 Transcribe, in preview since late August, lists diarization, custom vocabulary and word timestamps across more than 85 languages. AssemblyAI's Universal-3.5 Pro handles code-switching and diarization in 18 languages.



3

SMALL ENOUGH TO DISAPPEAR

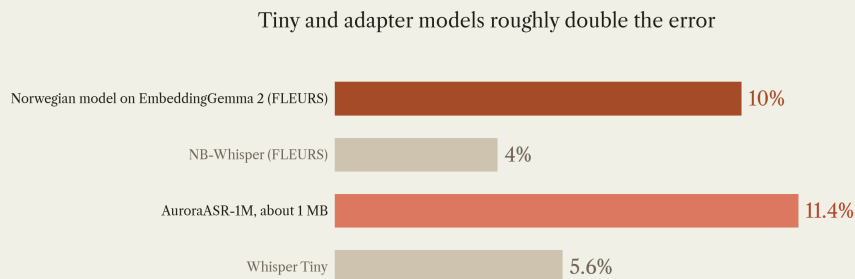
If accuracy is nearly level, size and speed become the interesting axis, and here the movement is fast.

Fermion's Phonon-2 is a 164 MB English model derived from NVIDIA's Parakeet TDT 0.6B v3 and quantized to about 2.1 bits per weight. Fermion says it turns an hour of audio into text in about 20 seconds on a MacBook Air and claims better accuracy than Whisper large. Further down, Cactus Whistle is 16.9 MB and runs on a CPU, and Lokutor's Oído, a 13M-parameter model, runs on a \$5 microcontroller and reports 3.7 percent WER on LibriSpeech, according to the vendors. These are self-reported numbers, but the direction is not in doubt.

Parakeet itself has become the default local engine for a lot of software, including Canonical's Myna on Ubuntu, according to a weekly roundup. When an operating system ships transcription, it has stopped being a product and become a feature.

The most interesting experiment this month goes further. Several builders have built speech recognisers on top of Google's small EmbeddingGemma 2 model, training only a small added piece, and cheaply. All of the numbers here come from single posts on X that no one else has checked. Haku reports about 288 times real time on a base M3 MacBook Air, after under two hours of training on eight H100s. A Norwegian model by AlfH is said to cost about \$15 to train and run 10 to 40 times faster than NB-Whisper. A Filipino and Taglish model, SEA-ASR-PH, claims 3,817 times real time on one H100.

The cost is accuracy. AlfH's Norwegian model gets 10 percent WER on FLEURS, against 4 percent for NB-Whisper. AuroraASR, a 949K-parameter English model of about 1 MB, reports 11.4 percent against 5.6 percent for Whisper Tiny.



Sources: single X posts by @AlfH (15 Oct 2026) and @TeksEdge (19 Oct 2026), not independently checked. Word error rate, percent, lower is better. Each new model is shown above the baseline it compares itself to.

That is the one place in this edition where accuracy is very much not settled. These tiny and adapter models roughly double the error of the models they compare themselves to. For a search index or rough notes that may be fine. For subtitles it isn't. It's an open question whether more data closes the gap or whether the approach has a ceiling.

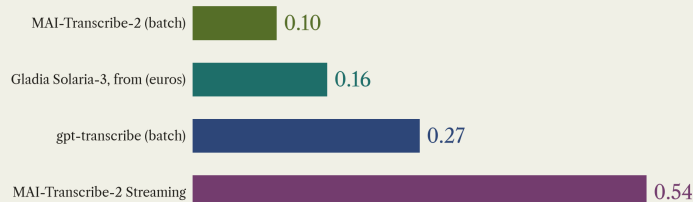


4

PRICE, STREAMING AND THE REST OF THE LIST

Prices for batch transcription have fallen to the point where they rarely decide anything at low volume. Microsoft says [MAI-Transcribe-2](#) costs \$0.10 an hour. [Gladia's Solaria-3](#) starts at €0.16 an hour, and [gpt-transcribe](#) costs \$0.0045 a minute, or \$0.27 an hour. Streaming costs more: Microsoft's [streaming model](#) is \$0.54 an hour, with first partial results in about 100 ms.

An hour of batch transcription costs cents



Sources: Microsoft, OpenAI and hotwired.news (Gladia). List prices per hour of audio, as published. Gladia is quoted in euros; the others in US dollars. Streaming is the last bar.

Vendors now ship batch and streaming models in pairs: OpenAI, Google, Microsoft, AssemblyAI and Mistral all have one of each. And streaming models are taking on turn-taking, deciding when a speaker has finished. Deepgram's [Flux](#) builds turn detection into recognition, and ElevenLabs added turn-prediction to realtime Scribe in October. For a voice agent, knowing when to reply now matters as much as hearing the words right.

Licences are the last item on the list, and the easiest to miss. Moonshine v2's English models are MIT-licensed but its other languages are [non-commercial](#). Tether's [QVAC TranscribePsy](#) describes itself as both research-only and Apache 2.0. Cohere Transcribe is Apache 2.0 but needs you to name the language and has no timestamps or diarization.

One caution on generalist models. Google's Gemma 4 can transcribe, but its E4B model scores 11.38 on the leaderboard, and an [independent benchmark](#) found the 12B model fails badly on meeting audio and refuses 7.4 percent of the time. The author's

advice is plain: don't swap it in for a dedicated model.



5

HOW I WOULD CHOOSE NOW

So, for anyone picking a speech model this autumn, I would start with a checklist rather than a leaderboard. Write down whether you need word timestamps, speaker labels, subtitles, streaming latency, a language beyond the big ones, on-device use, and a commercial licence. Then filter. Most of the time several models will pass, and they will be within a point of each other on WER. Pick the cheapest or smallest of those, and test it on your own audio, because the leaderboard is clean English and yours probably isn't.

Use context and keyword hints before fine-tuning. OpenAI, Google and Microsoft all now accept them. Fine-tune when the domain is narrow and the payoff is measurable: NatWest reports 20 percent fewer errors after tuning NVIDIA's Nemotron model on UK accents, and ElevenLabs reports about 35 percent lower WER on clinical audio for Scribe v2 Medical. Both are their own numbers.

If you are on whisper-1 or gpt-4o-transcribe, the date that matters is 26 February 2027. Check which features you rely on before you switch.

Accuracy still improves every month, and on hard audio it still matters a lot. But for most English transcription, the error rate is no longer what separates the good choices from the bad ones. The details around the words are. ■

About this series. ASR Watch is a weekly look at speech recognition. Edition 1 covers roughly April to October 2026. Vendor figures are self-reported unless noted, and leaderboard numbers are as of 20 to 21 October 2026.

Sources: Hugging Face Open ASR Leaderboard and its methodology paper; isitsota snapshot (20 to 21 October 2026); OpenAI gpt-transcribe announcement, deprecations, migration guide and pricing; Google, Gemini 3.5 Transcribe; AssemblyAI, Universal-3.5 Pro; Fermion, Phonon-2; Microsoft, MAI-

[Transcribe-2](#) and [its streaming model](#); Deepgram, [Flux](#); ElevenLabs, [Scribe v2 Medical](#); [Moonshine v2](#); James Ding, [Gemma 4 ASR benchmark](#); hotwired.news Voice AI weekly roundups for weeks [40](#) and [42](#); X posts by [@TslShahir](#), [@AlfH](#), [@seanphan](#) and [@TeksEdge](#), unverified beyond the posts themselves.

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