

ASR WATCH · EDITION 1 · 21 OCT 2026

Small speech models got good

What moved in speech recognition from April to October 2026, and what it means for teams that build with small models.

A practitioner's report · Sources on the last slide

Space, Page Down or arrow keys to move. Or use the links at the bottom.

THE SHORT VERSION

Three things to take away

~1 pt

WER gap between the best open models and the best paid APIs on English, about 1 to 1.5 points.³

26 Feb 2027

OpenAI shuts down whisper-1 and the gpt-4o transcribe models in its API.⁷

Small open models are good enough. Models under 1B sit close to the 2B class and run much faster.^{3,4}

Whisper's API has an end date, and the replacements drop timestamps, subtitles and speaker labels.⁸

Speech on embedding models. A small adapter on EmbeddingGemma 2 gives very fast, cheap ASR for one language, at a real accuracy cost. A new 163 MB Parakeet fine-tune matches Parakeet on English.^{34,35,37}

WER is word error rate, lower is better. RTFx is times faster than real time. "vendor" = self-reported. "X post" = one post, not checked.

Paid APIs lead, but small open models are close

MODEL	TYPE	SIZE	WER
Otter.ai Speech V3.1	API	n/a	3.88
Zoom Scribe V2 Pro	API	n/a	4.08
Azure Speech 06-2026	API	n/a	4.26
ElevenLabs Scribe v2	API	n/a	4.55
MOSS-Transcribe-Diarize	Open	n/a	5.17
Granite Speech 4.1	Open	2B	5.35
Edge0 ARK-ASR	Open	0.6B	5.53
Zipformer cr-ctc-transducer XL	Open	290M	5.56
Granite Speech 5.0 TurboCTC	Open	470M	5.78
Distil-Whisper large-v3.5	Open	n/a	6.18

Rows from the isitsota snapshot and the HF results dataset.^{1,2,3} API rows have no speed figure. Which rows are vendor-submitted is not clear.

Whisper's API has an end date, and the replacements drop features

What changed

- gpt-transcribe (files) and gpt-live-transcribe (realtime), released 28-29 Jul. Both take context, keywords and language hints.⁵
- gpt-transcribe costs \$0.0045 a minute, about \$0.27 an hour.⁶
- whisper-1 and the gpt-4o transcribe family shut down 26 Feb 2027.⁷
- Live model: 19.70% error on Common Voice (22 languages) vs 20.33% for gpt-realtime-whisper-1.

VENDOR⁵

Why it matters

- No native SRT or VTT, no word timestamps, no diarization. OpenAI says to keep the old models for those until shutdown.⁸
- If you use timestamps for chapters, aligned transcripts or clip search, this is a stack change, not a model swap.
- Open Whisper weights are not affected.

Sub-1B open models are production grade

The evidence

- Leaderboard paper (86 systems, 12 sets): CTC and TDT decoders are fastest, transformer decoders most accurate.⁴
- Cohere Transcribe 5.42 WER at 525 RTFx; Canary-Qwen 2.5B 5.63 at 418; Qwen3-ASR 1.7B 5.76 at 148.⁴
- Parakeet TDT 0.6B v3: 25 European languages, 6.34% average WER. **VENDOR**^{9,10} Default local engine in apps like Canonical's Myna.¹³
- Parakeet Unified 0.6B: offline and streaming in one model, 160 ms minimum latency.¹¹ Nemotron 3.5 Streaming 0.6B: 40 locales, 80 ms to 1.12 s latency.¹²

Why it matters

For bulk transcription that feeds search or metadata extraction, a ~0.5B model on your own GPUs is within about a point of the 2B class and much faster. It is the same trade as with small text models.

How embedding-model ASR works

EmbeddingGemma 2 puts text, images, audio and video in one space, so a spoken word lands near the written word.³² Builders read words straight out of that space.

Word lookup (Haku)

The audio encoder stays frozen. A small reader maps each slice of audio to the nearest of ~229,000 English words; a tiny model adds punctuation. Only the reader is trained: under 2 h on 8 H100s, on openly licensed speech.^{32,33}

CTC head (Norwegian, Filipino)

norsk-asr uses the 12-layer audio conformer with only the top layer trained, a 6-layer conformer reader with CTC over 83 character units, and an optional 4-gram LM.³⁴ sea-asr-ph is the encoder plus a CTC head, streaming in 0.48 s chunks, with word timestamps and SRT/VTT.³⁵

And in reverse

The same backbone was used to plan speech for TTS: 79.5x real time on a base M3 Air, quality "not close to the best". [X POST](#)³⁹ A later post claims 300x on an M5 Pro. [X POST](#)¹⁵

DEVELOPMENT 3 · THE NUMBERS

Very fast and cheap, clearly less accurate

MODEL	LANGUAGE	SIZE	SPEED	WER VS REFERENCE	LICENSE
Haku EG2 Transcribe	English	715 MB app models	288x, base M3 Air ³²	~5% vs 2-3% for leaders; test set not named ³²	Base Apache 2.0; free app
norsk-asr	Norwegian Bokmål	~850 MB with LM	1,000-1,600x H100; ~10x laptop CPU ³⁴	FLEURS 10.0 vs 4.2 NB-Whisper large; NPSC 9.7 vs 16.5 ³⁴	Apache-2.0 ³⁴
sea-asr-ph	Filipino, Taglish	364M params	3,817x H100; 860x A10G ³⁵	FLEURS fil 17.2 vs 10.6 Whisper large-v3; ties Whisper medium ³⁵	Apache-2.0 ³⁵

Training data: 1,680 h of open Norwegian speech³⁴ and 1,914 h of CC-BY Filipino labels, released as a dataset.³⁵ All figures are the builders' own. norsk-asr's NPSC lead is mostly transcription style; counting only misheard words, NB-Whisper is 1.5 to 2.5x better.³⁴

Unverified, but the pattern is real

Not checked

- "About \$15 to train" for Norwegian and "10 to 40x faster than NB-Whisper" come from one X post. [X POST](#)¹⁶
- Haku's speed table mixes its own runs with others' benchmarks on different Macs; the page itself calls it "a rough guide".³²
- None of these models is on the Open ASR Leaderboard yet.¹

Why it matters

- The cost is tiny: a frozen backbone, a few hours of training, a few thousand hours of open data.^{32,34,35}
- Error rates are 1.5 to 2.5x the best per language, so this fits search, draft captions and topic or entity extraction, not verbatim transcripts.
- Google ships on-device products on the same backbone.¹⁸

Underdog Bark 3.0: Parakeet accuracy in a 163 MB file

What it is

- English ASR, a low-bit fine-tune of NVIDIA Parakeet TDT 0.6B v3 for on-device MLX. Not built on EmbeddingGemma. CC-BY 4.0.³⁷
- 4.94% average WER on the leaderboard's seven English sets, against 4.941% for Parakeet v3 (2.5 GB): a tie.
VENDOR³⁷
- Meetings (AMI): 7.82% vs 9.42% Parakeet and 13.63% Whisper large-v3. **VENDOR**³⁶
- Family: Clean 180 MB drops "um"s (5.83%), Plus 180 MB 4.76%, Pro 1.25 GB 4.29%. **VENDOR**³⁶

What to watch

- Speed claims differ: "1 min in under 0.25 s on a MacBook" in the launch post **X POST**³⁸, about 1.5 s a minute (41x) on an M2 with stock mlx-audio on the model card.³⁷
- Phonon-2 did the same trick at 164 MB and 5.21%; Bark edges it.³⁶
- Self-scored with the leaderboard's method, not listed on it yet.³⁶
- Page dated 15 Oct; announced on X 21 Oct.^{36,38}

Local transcription is a commodity

Fermion Phonon-2 · 164 MB

Parakeet v3 at ~2.1 bits. 1 h of audio in ~20 s on a MacBook Air; claims better than Whisper large. [VENDOR](#)¹⁹ Core ML port: 1 h in 5.6 s on an M5 Pro. [X POST](#)²⁰

Cactus Whistle · 16.9 MB

CPU, 7 languages, "mostly beats Whisper base". [VENDOR](#)¹⁴

Lokutor Oído · 13M

Runs on a \$5 ESP32-S3; 3.7% WER on LibriSpeech. [VENDOR](#)¹⁴

AuroraASR-1M · ~1 MB

436x real time on one CPU thread, but 11.4% WER vs 5.6% for Whisper Tiny. [X POST](#)²¹

QVAC TranscribePsy · 254M

10 min in 1.65 s on iPhone 17 Pro Max. [VENDOR](#) License wording contradicts itself.¹³

Why it matters: transcription can happen where the audio is, so you may get text plus metadata instead of audio.

Batch is cents an hour, and every vendor ships a pair

Price per hour of audio

- Microsoft MAI-Transcribe-2: **\$0.10**, 60 languages, 5.2% WER on FLEURS. VENDOR²²
- Gladia Solaria-3: from **€0.16**.¹³
- OpenAI gpt-transcribe: **~\$0.27**.⁶
- MAI-Transcribe-2-Streaming: **\$0.54**, first partials in ~100 ms. VENDOR²³

Batch and live pairs

- OpenAI: gpt-transcribe / gpt-live-transcribe⁵
- Google: Gemini 3.5 Transcribe / -live²⁴
- Microsoft: MAI-Transcribe-2 / -Streaming^{22,23}
- AssemblyAI: Universal-3.5 Pro Async / Realtime²⁵
- Mistral: Voxtral Mini Transcribe V2 / Realtime²⁶

Why it matters: at ten cents an hour, self-hosting only wins at real volume or when audio can't leave your systems.

Give the model your entity list before you fine-tune

Context at inference

- gpt-transcribe: prompts, keywords, language hints.⁵
- Gemini 3.5 Transcribe: custom vocabulary, word timestamps, diarization.²⁴
- MAI-Transcribe-2: keyword biasing.²²

Where people do fine-tune

- ElevenLabs Scribe v2 Medical: ~35% lower WER on clinical audio. **VENDOR**²⁷
- NatWest tuned Nemotron on UK accents: 20% fewer errors on its own test set. **VENDOR**¹³

Why it matters: in content understanding you usually know the names (people, products, places). Bias first, measure entity recall, then decide.

Multimodal LLMs still trail dedicated ASR

11.38

Gemma 4 E4B WER on the leaderboard. The 12B scores 20.87.³

7.4%

of the time the Gemma 4 12B refused, in an independent test.²⁹

Gemma 4 E2B, E4B and 12B do ASR, in clips up to 30 s.²⁸

James Ding: E4B 7.6% mean WER on 7 sets without AMI; 12B hit 104% on AMI. "Don't swap it in for a dedicated ASR model."²⁹

Do this instead: a dedicated speech model in front, its text into the LLM.

ALSO WORTH KNOWING

Turn-taking, languages and licences

Turn-taking moves into STT

Deepgram Flux has built-in turn detection; Anam reports ~50% fewer barge-ins. **VENDOR**³⁰ ElevenLabs realtime Scribe added turn prediction on 12 Oct.¹³

Small languages

Covered by regional fine-tunes: Arabic dialects (Jais-ASR, Falcon-ASR), Japanese (Kotoba), plus Norwegian and Filipino.^{13,16,17}

Licence traps

Moonshine is non-commercial outside English. TranscribePsy's text contradicts itself. Cohere Transcribe is Apache-2.0 but needs the language set and has no timestamps.^{13,31}

Six experiments for this month

- 1 Baseline on your audio.** 2 to 5 hours with good references; measure Parakeet v3, Bark 3.0 or Granite 5.0 470M for WER and cost per hour.

- 2 Score by the downstream task.** Run your tagger on reference text and on each transcript. If tag F1 barely moves between 5 and 10 WER, use the cheaper model.

- 3 Bias before you fine-tune.** Pass your entity list to gpt-transcribe or Gemini 3.5 Transcribe; measure entity recall.

- 4 Prototype an adapter** on a frozen small backbone for one niche language or domain. Treat the X-post numbers as claims until you reproduce them.

- 5 Plan the OpenAI move.** List everything using timestamps or diarization; it needs a new model before 26 Feb 2027.

- 6 Price it.** \$0.10 to \$0.27 an hour from APIs against your own GPU cost per hour.

What I don't know yet

Open questions

- Which leaderboard rows are independent runs, and which are vendor submissions?
- Does embedding-backbone ASR close the WER gap with more data?
- How do the new OpenAI and Google models do on long audio against Parakeet TDT?
- Is MAI-Transcribe-2 API only? It looks that way.
- Will streaming benchmarks become as trusted as the HF leaderboard?

Dropped as unverified

- Parakeet v3 at 6.32 WER / 3,330 RTFx in the paper: from a search summary, row not seen.
- AssemblyAI "Universal-3.6 Pro Realtime": no primary source.
- Nemotron 3.5 Streaming release date: sources disagree.
- @reach_vb as a source to watch: not checked.

Sources

1. [Hugging Face Open ASR Leaderboard \(space\)](#)
2. [Open ASR Leaderboard results dataset](#)
3. [isitsota: Open ASR snapshot \(Oct 20-21, 2026\)](#)
4. [Open ASR Leaderboard paper, arXiv 2510.06961 v4](#)
5. [OpenAI forum: gpt-transcribe and gpt-live-transcribe \(Jul 2026\)](#)
6. [OpenAI: gpt-transcribe model page](#)
7. [OpenAI: deprecations](#)
8. [OpenAI Cookbook: migrating from Whisper to gpt-transcribe](#)
9. [NVIDIA: Granary dataset, Canary-1B-v2 and Parakeet-TDT-0.6B-v3 \(Aug 2025\)](#)
10. [NVIDIA: production-ready open models for European languages](#)
11. [NVIDIA NeMo Speech \(Parakeet Unified\)](#)
12. [Nemotron 3.5 ASR Streaming 0.6B model card](#)
13. [hotwired.news Voice AI, week 42 2026](#)
14. [hotwired.news Voice AI, week 40 2026](#)
15. [X post: Haku, EmbeddingGemma 2 as ASR \(@TslShahir\)](#)
16. [X post: Norwegian STT on EmbeddingGemma 2 \(@AlfH\)](#)
17. [X post: SEA-ASR-PH \(@seanphan\)](#)
18. [X post: Google AI Edge Foresight \(@googlegemma\)](#)
19. [Fermion Research: Phonon-2](#)
20. [X post: FluidAudio Core ML port of Phonon-2](#)
21. [X post: AuroraASR-1M \(@TeksEdge\)](#)
22. [Microsoft: MAI-Transcribe-2 \(Sep 2026\)](#)
23. [Microsoft: first streaming transcription model \(Oct 2026\)](#)
24. [Google: Gemini 3.5 Transcribe \(Aug 2026\)](#)
25. [AssemblyAI: Universal-3.5 Pro Async](#)
26. [Mistral: Voxtral Transcribe 2](#)
27. [ElevenLabs: Scribe v2 Medical GA \(Sep 2026\)](#)
28. [Google AI: Gemma audio capabilities](#)
29. [James Ding: Gemma 4 ASR benchmark \(Jun 2026\)](#)
30. [Deepgram: Introducing Flux](#)
31. [Hugging Face blog: Cohere Transcribe release](#)
32. [Haku: Embedding Gemma Transcribe \(project page\)](#)
33. [X post: Haku thread follow-up, how it works \(@TslShahir\)](#)
34. [norsk-asr repo and README \(Norwegian STT on EmbeddingGemma 2\)](#)
35. [PixelMLL sea-asr-ph model card](#)
36. [Underdog: Meet Underdog Bark 3.0](#)
37. [Underdog Bark 3.0 model card \(Hugging Face\)](#)
38. [X post: Underdog Bark 3.0 launch \(@UnderdogAI, 21 Oct 2026\)](#)
39. [X post: EmbeddingGemma 2 as TTS \(@TslShahir\)](#)