

gpt-transcribe vs Parakeet v3 vs Voxtral vs Scribe: the honest speech-to-text comparison for CTOs

If whisper-1 disappears on 26 February 2027, what do you actually move to? The leaderboard answer is uncomfortable for anyone who likes managed APIs: the most accurate of the four replacements runs on your own GPU.

OpenAI's own replacement, gpt-transcribe, costs $0.0045 per minute for recorded files and is the fastest migration: a model string change. But the most accurate option in this comparison is NVIDIA Parakeet TDT 0.6B v3 at 6.32% WER, open weights under CC-BY-4.0, which you run yourself. The deadline is 26 February 2027, announced 26 August 2026.

The shutdown covers whisper-1, gpt-4o-transcribe, gpt-4o-mini-transcribe, gpt-4o-transcribe-diarize and the /v1/audio/translations endpoint, per the OpenAI deprecations page. The named successors are gpt-transcribe for recorded files, gpt-live-transcribe for live audio, and gpt-realtime-translate via /v1/realtime/translations for translation. Six

months of notice, and by the time most teams read the changelog, closer to three.

The useful part is that you do not have to decide on vendor marketing. The Open ASR Leaderboard paper (arXiv 2510.06961, updated 19 October 2026) evaluated 60+ open and proprietary systems across 11 datasets and reports both word error rate and RTFx, a throughput measure of audio seconds processed per wall-clock second. Accuracy and speed in the same table, from the same harness. That is rare enough to build a migration plan on.

Four systems sit between 6.32% and 7.05% WER

NVIDIA Parakeet TDT 0.6B v3 sits at rank 7 on the leaderboard with 6.32% WER and RTFx 3332.74, the fastest throughput among the leading systems. It ships under CC-BY-4.0 open weights and is reported to cover 25 European languages. A 0.6B parameter model, so it fits comfortably on hardware you probably already rent.

NVIDIA Canary 1B Flash is rank 8 at 6.35% WER, RTFx 1045.75. Statistically it is a coin flip against Parakeet on accuracy and roughly a third of the throughput. It exists in this comparison because it is the obvious fallback if Parakeet's language coverage or output format does not fit your pipeline.

ElevenLabs Scribe v1 is rank 12 at 6.88% WER. No RTFx is reported, because it is a closed API and the leaderboard cannot measure GPU-second throughput on someone else's infrastructure. That absence is itself information: you are buying latency SLAs, and throughput you cannot model.

Mistral Voxtral Mini 3B is rank 16 at 7.05% WER with RTFx 109.86. The accuracy gap to Parakeet is 0.73 points. The throughput gap is roughly 30x per GPU-second. Voxtral is an audio-capable language model rather than a pure transcriber, which explains the cost and some of why teams still want it.

And gpt-transcribe itself: OpenAI has not published a leaderboard position for it in the data I have, so I cannot give you a WER figure for it alongside the others. What I can give you is price. $0.0045 per minute for recorded files, $0.017 per minute for live audio, with a reported 10% uplift for regional or data-residency processing, which puts EU-processed

gpt-transcribe at roughly $0.00495 per minute.

ATTRIBUTE	PARAKEET TDT 0.6B V3	CANARY 1B FLASH	SCRIBE V1	VOXTRAL MINI 3B	GPT-TRANSCRIBE
Open ASR WER	**6.32%**	6.35%	6.88%	7.05%	n/a
Leaderboard rank	**7**	8	12	16	n/a
RTFx (throughput)	**3332.74**	1045.75	n/a (closed API)	109.86	n/a (closed API)
Weights	**Open, CC-BY-4.0**	n/a	Closed	n/a	Closed
Price per minute	Your GPU cost	Your GPU cost	n/a	Your GPU cost	$0.0045 ($0.00495 EU)
Language coverage	25 European languages	n/a	n/a	n/a	n/a

The blanks are real. I am not going to fill a comparison table with numbers nobody published. Where a cell says n/a, the leaderboard paper or the vendor docs simply do not carry that figure, and if you need it you should measure it on your own audio before signing anything.

Doing nothing is survivable only if you self-host

Whisper large-v3 remains MIT-licensed and self-hostable across roughly 99 languages, which makes it the only genuinely zero-migration path for teams that were never calling the OpenAI API in the first place. If your Whisper runs on your own box, 26 February 2027 is a date on someone else's calendar.

If you call whisper-1 over HTTP, you have a hard deadline. The cheapest move is also the laziest: swap the model string to gpt-transcribe and keep going. That buys you a working pipeline at $0.0045 per minute. It also leaves you on a closed model whose accuracy you cannot independently verify against the leaderboard, with a second deprecation clock that starts the day you ship.

My take: the laziest move is defensible for teams with low audio volume, where engineering time costs more than the API ever will. Above that, you are paying a per-minute rent on a capability you could own, and the rent compounds for as long as the product lives.

Parakeet beats Voxtral on accuracy and runs 30x faster

Parakeet v3 and Voxtral Mini are both open-weight models you run on your own hardware. Parakeet is more accurate (6.32% vs 7.05%) and processes roughly 30x more audio per GPU-second (RTFx 3332.74 vs 109.86). There is no trade-off here. One model is better on both axes that matter for batch transcription.

RTFx 3332 means one GPU-second of compute handles over 3300 seconds of audio under the leaderboard's harness. Your production numbers will be lower, because real pipelines have I/O, chunking and queueing that a benchmark does not. But the ratio between Parakeet and Voxtral should survive, and the ratio is what decides your GPU bill.

So why would anyone pick Voxtral? Because Voxtral Mini is an audio-capable language model, and if your pipeline is transcribe-then-reason, running one model instead of two removes an entire handoff. I have no sourced cost comparison for that consolidated architecture, so treat it as a hypothesis to test rather than a recommendation. The 30x throughput gap sets a high bar for what the consolidation has to save you.

Managed transcription has turned into a price fight

Across the managed options, the per-minute numbers have converged to within a factor of a few, and the accuracy spread among the leaders is well under a percentage point. Deepgram Nova-3 batch runs $0.0043 per minute monolingual (about $0.258 per hour) and $0.0052 multilingual (about $0.312 per hour), with a reported 5.26% WER on a nine-domain real-world test. That 5.26% is not from the Open ASR Leaderboard harness, so do not line it up against the 6.32% as if the two were measured the same way. Different test set, different rules.

AssemblyAI Universal-2 is listed at $0.15 per hour, with diarization at $0.02 per hour prerecorded and $0.12 per hour streaming. Google Chirp 3 runs $0.016 per minute for the first 500,000 minutes, dropping to $0.003 per minute with dynamic batch. Azure Speech

sits higher: $0.18 per hour batch, $0.36 fast, $1.00 real-time.

The outlier is real-time. gpt-live-transcribe costs $0.017 per minute against $0.0045 for recorded files, a 3.8x premium for the same audio arriving differently. Azure's $1.00 per hour real-time against $0.18 batch is a 5.5x premium. If your product treats streaming as the default because it feels more modern, that decision now has a line item.

Regional and data-residency processing adds a reported 10% uplift on eligible OpenAI models, taking gpt-transcribe to roughly $0.00495 per minute. Small per minute, but it is a recurring tax on a decision that self-hosting makes free. If you are already thinking about where your inference physically runs, I wrote about the hosted vs self-operated maths for a Swiss sovereign NLP stack, and the same arithmetic applies here.

Volume, language coverage and licence decide which one you pick

Pick **gpt-transcribe** if your audio volume is low, your team is small, and the migration is a model string change you can ship this sprint. At $0.0045 per minute you are buying time, and time is the scarce resource between now and February 2027. The cost of this choice is that you cannot verify its accuracy against the leaderboard, and you are back here the next time OpenAI retires a model family.

Pick Parakeet TDT 0.6B v3 if you process meaningful volume, your languages are European, and you have anyone on staff who can keep a GPU service alive. It is the most accurate of the four on the leaderboard and by far the fastest, under a permissive CC-BY-4.0 licence. This is the choice that ends the deprecation treadmill.

Pick Canary 1B Flash if Parakeet's 25-language coverage misses something you need and Canary covers it. The accuracy difference is 0.03 points, which is noise. You trade roughly 3x throughput for whatever Canary gives you that Parakeet does not.

Pick Scribe v1 if you want a managed API that lands in the leaderboard's top 12 and you are willing to pay for operational simplicity. The missing RTFx figure means you cannot model its cost per GPU-second, but that is the point of a managed service.

Pick Deepgram Nova-3 if price per minute is the binding constraint. At $0.0043 monolingual it is the cheapest managed option in this data, marginally under gpt-transcribe before the EU uplift.

Pick Whisper large-v3 if you already self-host it and it works. MIT licence, roughly 99 languages, zero migration. Its leaderboard position is not in my data, so I cannot tell you how far behind Parakeet it sits on accuracy. If accuracy is not currently a complaint, that gap is not urgent.

WHERE I LAND

For the typical reader of this piece, a team with real audio volume and a GPU budget: Parakeet TDT 0.6B v3, with gpt-transcribe as the stopgap you ship in week one so the February deadline stops being a risk. Do the swap first, do the migration second. Running both in parallel for a month on your own audio is the only accuracy test that counts, because none of these leaderboard numbers were measured on your domain, your accents, or your microphones.

The leaderboard settles throughput, not your domain

Eleven datasets is a serious evaluation and 60+ systems is a serious sample, but the whole exercise still measures general-domain audio. Medical dictation, call-centre recordings over a bad PSTN leg, Swiss German in a meeting room: none of those are represented by a leaderboard average, and the ranking between Parakeet at 6.32% and Scribe at 6.88% could invert entirely on your corpus. A 0.56 point spread is well inside the range that domain mismatch can erase.

What the leaderboard does settle is the throughput question, and that is the one most teams get wrong by two orders of magnitude when they guess. The 30x between Parakeet and Voxtral is a hardware-budget fact, and it holds regardless of how your accents land. Benchmark accuracy on your own audio; trust the published throughput ratios.

The other thing I would watch is diarization. OpenAI is retiring gpt-4o-transcribe-diarize along with the rest, and the replacement list names gpt-transcribe, gpt-live-transcribe and gpt-realtime-translate without a diarization equivalent in the data I have. If speaker labels

are in your product, confirm what you are migrating to before you assume the rest of the pipeline transfers. AssemblyAI prices diarization separately at $0.02 per hour prerecorded and $0.12 streaming, which at least tells you what the market thinks the feature is worth.

Three months is enough time to do this properly if you start with the stopgap and run the real evaluation behind it. It is not enough time to discover in February that your diarization has no home. If you are weighing the self-host decision against the per-minute bill and want a second opinion on the arithmetic, drop me a note and we can go through your numbers.



WRITTEN BY

Artur Markus

Software engineer in Basel building AI infrastructure and agentic systems, with a background in pharmaceutical automation, GxP compliance, and validation. I help teams turn AI from a chatbot into a reliable operational layer.

Book a meeting →