PURETENSOR
INTELLIGENCE DISTILLED

We Benchmarked the Open Relatives and Did Not Swap

Open-Weight Dictation Speech Recognition on 400 Clips With a Paired Bootstrap

Paper: PT-R-2026-057 v1.0

Author: PureTensor Ltd

Date: 21 October 2026

Status: Published

Abstract

On 20 October 2026 a vendor launched a hosted transcription family and published a chart of word error rate in which, as we recorded it from the public chart, the six leading entries were all API-only products with no downloadable weights. Our rule, set the next day, is that dictation speech recognition runs on open weights only, because a dictation stream is the most private text we produce and a hosted service would carry every spoken word off our own hardware. The hosted entries were therefore out regardless of score; the question was whether the open models in the same class beat what we already run by enough to justify a swap.

We answered it on one morning with 200 English clips (FLEURS) and 200 Icelandic clips (Samromur), every engine on the same clips, and a paired bootstrap over clips for the intervals. On English, Voxtral Small 24B scored 3.92% word error against 4.80% for our server lane (Whisper large-v3-turbo, beam 5), and Voxtral Mini 3B scored 4.38% against 5.24% for the on-device Parakeet TDT 0.6B v3. Fun-ASR-Nano was last at 5.72%. Voxtral Small has the lowest point estimate, and only its gain over Parakeet (1.32 points) is firm. The two other gains whose 95% intervals exclude zero (Small against the server, Mini against Parakeet) are marginal: both include zero at 99%, and the Mini gain ends at -0.05 points. A fourth model, Voxtral Mini 4B Realtime, returned empty and malformed output; we report that arm as invalid. On Icelandic, the server lane scored 18.57% against 19.70% for the same model run greedily on a laptop; none of the new candidates lists Icelandic, going by their model cards.

We did not swap. Voxtral Small needs about 46 GB of GPU memory and about one second per clip against roughly a tenth of a second, and it does not list Icelandic. Voxtral Mini 3B was not run on the laptop that carries over 98% of our dictation. A one-point gain on clean read speech does not justify either cost, and we have no transcribed sample of the speech we actually dictate.

1. A Chart We Could Not Use

We run voice dictation for our own use on two lanes. The first is on-device: a 24 GB Apple M5 Pro laptop keeps two recognisers resident through MLX (Apple's array framework) behind hotkeys. English goes to NVIDIA Parakeet TDT 0.6B v3, a 0.6B-parameter transducer model. Icelandic goes to Whisper large-v3-turbo with greedy decoding. The second is a server lane: a

GPU host runs faster-whisper with large-v3-turbo in float16 at beam size 5 (the decoder keeps five candidate transcriptions in play and picks the best), reached over HTTP by other devices.

The usage log on the laptop says which lane matters. Between 7 and 21 October it recorded 911 microphone takes: 901 English (549.5 minutes of audio) and 10 Icelandic (5.3 minutes). English is 98.9% of takes and 99.0% of minutes, with a mean English take of 36.6 seconds. A further 24 English takes (1.1 minutes) came from audio files. In practice Parakeet is the production dictation system. These are aggregates; we did not read the text.

The vendor's chart ranked a composite benchmark. As we recorded it, its six leading entries were hosted, with no weights to download at the time. We make no claim about their quality beyond their chart position, and we never ran them. The rule we set the next day (21 October) was that dictation recognition must be open-weight, never an API. That left the open relatives of the chart's leaders, which we can run ourselves.

This is a decision test, not a leaderboard: 400 public clips, one run per engine, one morning's work.

2. What We Ran

The English clips are 200 utterances drawn at random (fixed seed) from a pool of 646 FLEURS English (US) utterances; the staging script did not record which split it read. The Icelandic clips are 200 utterances drawn with the same seed from the first of two shards of the Samromur 21.05 test set (5,000 rows; we did not download the second), so they are 200 of the first 5,000 rows, not of the whole set. Audio was converted to 16 kHz mono.

	English	Icelandic
Clips	200	200
Audio, total	1,901.7 s (31.7 min)	1,190.8 s (19.8 min)
Clip length, mean (min to max)	9.51 s (3.96 to 20.18)	5.95 s (2.47 to 10.45)
Reference words after normalisation	4,315	1,777

Table 1. The two clip sets.

Seven configurations ran on the same clips, with a language hint where accepted. Four were GPU jobs on a single workstation Blackwell card with 96 GB: Voxtral Small 24B (2507), Voxtral Mini 3B (2507), Fun-ASR-Nano (2512) and Voxtral Mini 4B Realtime (2602). All used transformers 5.17 in bf16 at default settings, greedy decoding and a batch of one, with up to 256 new tokens (the Realtime model: one token per 80 ms of audio plus 64). The fifth was the production server lane, called over HTTP, with the round trip included in its time. The last two were the on-device engines, Whisper large-v3-turbo (greedy) and Parakeet TDT 0.6B v3 (bf16 conversion), run on the laptop in a separate process and never through the live dictation daemon.

Each engine ran one untimed warm-up and then the 200 clips one at a time, with the CUDA clock synchronised around each GPU clip. To give a job a whole GPU, the launcher drained the resident services on the compute node and restarted them afterwards. Three launches were needed: one for Fun-ASR-Nano (which also held the first Realtime attempt, killed after 8.6 seconds), one for the Realtime rerun, and one for Voxtral Mini 3B and Small 24B, which were downloaded later. Everything is a single run. No engine was repeated.

Scoring uses word error rate (WER): substitutions plus deletions plus insertions, over reference words, pooled across clips rather than averaged per clip. English text on both sides passes through the Open ASR Leaderboard English normaliser. Icelandic uses the basic multilingual normaliser, which keeps diacritics. Character error rate (CER) is pooled the same way. For

confidence intervals we resampled the 200 clips with replacement 2,000 times, recomputed the pooled WER of both systems on the same resample, and took the 2.5th and 97.5th percentiles of the difference. The scoring script's incumbent is Parakeet for English and on-device Whisper for Icelandic; we also computed intervals with the production server as incumbent.

Every WER, CER, error-count and latency figure below comes from re-running the retained scoring script on the retained hypothesis files, and matches our logged table. The server-incumbent intervals and the 99% endpoints came from an inline calculation that was not retained; an independent reimplementation matched them to within about 0.02 points (0.07 for one Icelandic bound). Read those endpoints as accurate to roughly 0.05 to 0.1 points. No conclusion changes.

3. English Results

Engine	WER %	CER %	Errors (S / D / I)	p50 ms	p90 ms	Delta vs Parakeet, pts [95% CI]	Delta vs production server, pts [95% CI]
Voxtral Small 24B	3.92	2.00	169 (93 / 58 / 18)	984	1388	-1.32 [-2.18, -0.48]	-0.88 [-1.60, -0.16]
Voxtral Mini 3B	4.38	1.89	189 (129 / 36 / 24)	265	366	-0.86 [-1.69, -0.05]	-0.42 [-1.10, +0.22]
Production server, Whisper turbo, beam 5	4.80	1.97	207 (142 / 36 / 29)	106	140	-0.44 [-1.04, +0.17]	none
Whisper turbo, on device, greedy	4.96	2.07	214 (146 / 39 / 29)	230	258	-0.28 [-0.91, +0.32]	+0.16 [-0.14, +0.44]
Parakeet TDT 0.6B v3, on device	5.24	2.28	226 (160 / 45 / 21)	97	127	none	+0.44 [-0.15, +1.06]
Fun-ASR-Nano	5.72	2.39	247 (180 / 33 / 34)	279	403	+0.49 [-0.26, +1.23]	+0.93 [+0.20, +1.66]

Table 2. English, 200 FLEURS clips (4,315 reference words), sorted by WER. S, D and I are substitutions, deletions and insertions. A negative delta means fewer errors than the incumbent. The Realtime arm is in Section 7.

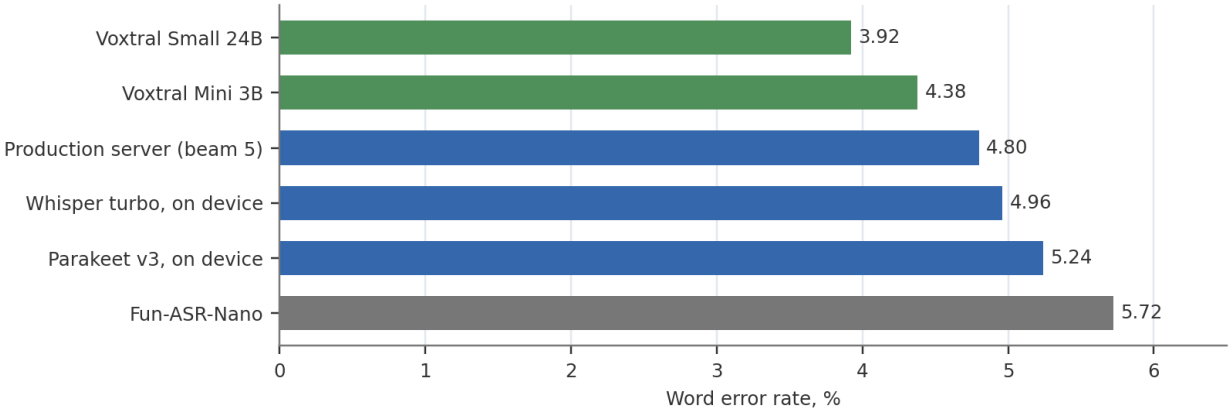


Figure 1. English word error rate per engine on 200 clips. Green bars are the Voxtral models, blue bars our current deployments, grey the Fun-ASR-Nano model. Values are in Table 2.

The six valid engines span 1.80 points of WER (3.92 to 5.72) on read speech of about 22 words per clip. The Voxtral models make fewer substitutions (93 and 129, against 142 to 180). Voxtral Small makes the most deletions among the leaders (58, against 36 for Mini and the server);

these counts are small.

A per-clip view agrees with the pooled one. Counting clips where the first system made fewer errors, more errors, or the same number, Voxtral Small against Parakeet is 64, 25 and 111, and Voxtral Mini against Parakeet is 57, 37 and 106. Most clips are ties, as one expects from clean read speech.

Four comparisons in Table 2 exclude zero at 95%: Small against Parakeet, Small against the production server, Mini against Parakeet, and Fun-ASR-Nano against the production server, where it is worse (+0.93, interval +0.20 to +1.66). The rest are inside noise. Fun-ASR-Nano ran once through a community transformers port at default settings, so its last place may reflect the port as much as the model.

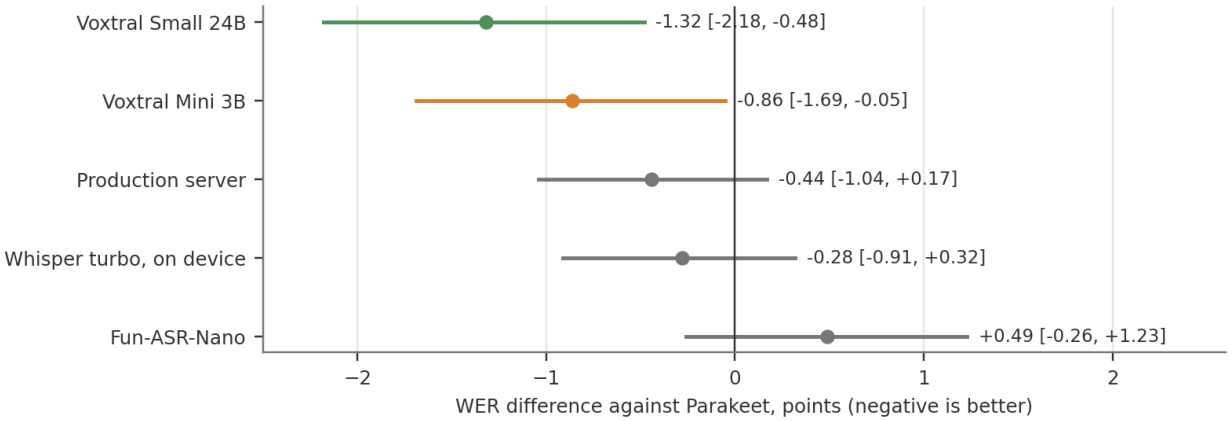


Figure 2. Difference in WER against Parakeet with 95% paired-bootstrap intervals; negative is better. Green marks an interval clearly below zero; orange marks Mini, whose interval ends at -0.05 and is marginal (Section 4). Values are in Table 2.

4. How Much "Significant" Carries

Our first reading of this run called two results significant: Small against the production server, and Mini against Parakeet. That was true at the 95% level as run. It is too strong a description of what we have, so we stress-tested both, and Small against Parakeet, with a second bootstrap of 10,000 resamples and a different seed. This check came after the original run. We did not stress-test Fun-ASR-Nano against the server.

Comparison	95% CI, pts	99% CI, pts	Share of resamples with difference >= 0
Small vs Parakeet	[-2.16, -0.46]	[-2.42, -0.18]	0.14%
Mini vs Parakeet	[-1.69, -0.02]	[-1.99, +0.24]	2.2%
Small vs production server	[-1.58, -0.16]	[-1.82, +0.07]	0.96%
Mini vs production server	[-1.07, +0.23]	[-1.27, +0.45]	10.8%
Small vs Mini	[-1.16, +0.27]	[-1.36, +0.50]	10.8%

Table 3. Supplementary check, 10,000 resamples; negative means the first system made fewer errors.

Small against Parakeet holds at 99%. The other two results do not: Mini against Parakeet ends at -0.05 points in the original run and -0.02 in the larger one, and includes zero at 99%, and so does Small against the production server. We also made five or six English comparisons without multiplicity correction, so a few intervals clearing zero at 95% is what chance alone would

sometimes produce. Our summary: Voxtral Small has the lowest point estimate of every engine and is clearly better only than Parakeet; its edge over the production server is marginal; Mini is better than Parakeet by a margin close to the noise floor of 200 clips; and Small and Mini cannot be told apart.

5. Icelandic: A Gap Between Deployments

Engine	WER %	CER %	Errors (S / D / I)	p50 ms	p90 ms	Delta, pts [95% CI]
Production server, Whisper turbo, beam 5	18.57	4.52	330 (277 / 27 / 26)	94	109	-1.13 [-1.96, -0.40] vs on-device
Whisper turbo, on device, greedy	19.70	4.89	350 (292 / 30 / 28)	227	255	none (incumbent)

Table 4. Icelandic, 200 Samromur clips (1,777 reference words). With the server as incumbent the interval is +1.13 [+0.40, +1.94].

Only these two deployments ran on Icelandic, deliberately: none of the new candidates lists Icelandic among its supported languages. That rests on the model cards' language lists; we did not run those models on Icelandic audio, and do not claim that no open model could cover the language.

Icelandic WER of 18 to 20% is about four times the English figure. Most errors are substitutions (277 of 330 and 292 of 350) and CER is far lower (4.5 to 4.9%), consistent with inflected word forms being scored wrong as whole words.

The 1.1-point gap favours the server, but it is a gap between two deployments, not a measurement of beam search: they differ in decoding (beam 5 against greedy), runtime (CTranslate2 in float16 against MLX) and weights (the laptop's are a conversion). Icelandic takes are about 1% of production dictation (10 of 911).

6. What Speed and Memory Cost

Engine	Environment	p50 ms	mean ms	Real-time factor (processing / audio)	Peak GPU memory
Parakeet v3	laptop, MLX	97	101	0.011	not measured
Production server	GPU host, HTTP	106	118	0.012	not measured
Whisper turbo	laptop, MLX	230	232	0.024	not measured
Voxtral Mini 3B	GPU, transformers	265	271	0.028	9,011 MiB
Fun-ASR-Nano	GPU, transformers	279	284	0.030	1,650 MiB
Voxtral Small 24B	GPU, transformers	984	1,010	0.106	46,440 MiB

Table 5. English per-clip latency and peak GPU memory. Single request, one pass; the three environments (GPU with default transformers settings, server over HTTP, laptop in MLX) are not comparable like for like.

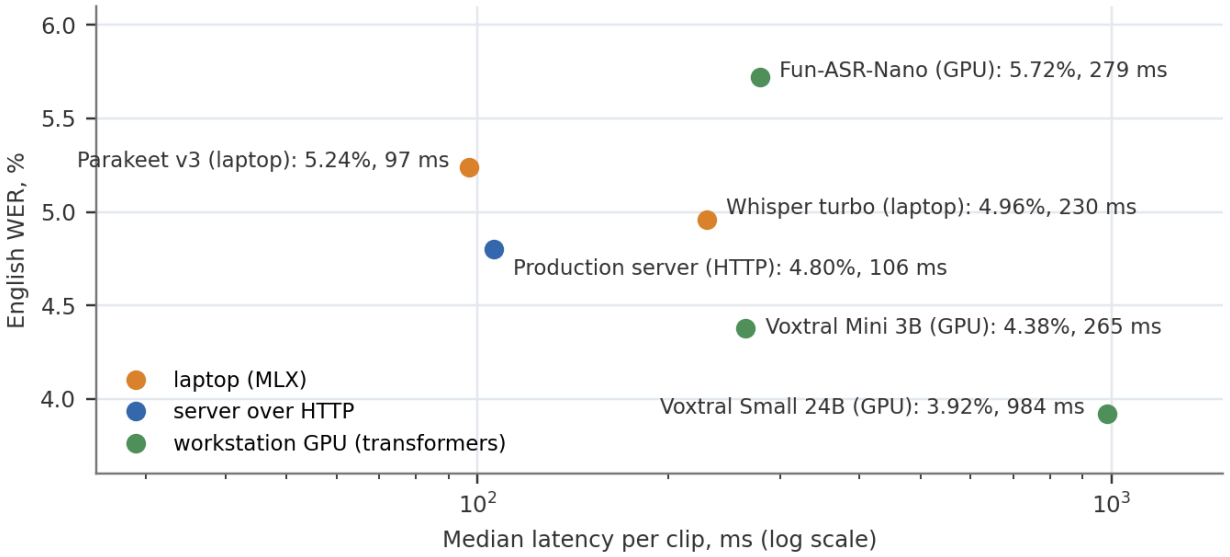


Figure 3. English WER against median latency on a logarithmic axis; the legend marks each engine's environment. Values are in Tables 2 and 5.

Voxtral Small has the lowest WER and is also the slowest valid engine: a median of 984 ms per clip, about nine times the production server's 106 ms, and 46,440 MiB of peak GPU memory. Voxtral Mini 3B takes 265 ms on the workstation GPU, 2.7 times Parakeet's 97 ms on the laptop. That ratio compares different hardware and software stacks, but it does show that a model of this size with a language-model decoder was not faster than a 0.6B transducer even on a much larger processor.

The benchmark understates the cost for dictation: these clips average 9.5 seconds, a production take 36.6 seconds, and the wait after releasing the hotkey is felt on every take. Voxtral Mini 3B was not run on the laptop, so we have no measurement there. On memory, such an engine would sit beside two resident dictation models on a 24 GB machine that also holds a local language model.

7. The Invalid Arm

Voxtral Mini 4B Realtime (2602), the streaming sibling in the same family, produced output that is not a transcription result. In the transformers port, 16 of 200 clips (8%) returned an empty string, and at least one non-empty output had its spaces stripped, a whole sentence as one run of letters and commas. The pooled WER scored as 15.37% (401 deletions, 210 substitutions, 52 insertions against 4,315 words); the other 184 clips alone scored 8.22%.

We cannot tell whether the fault lies in the model, the port, or our generation settings, so we report the arm as invalid and publish no WER for it as a measurement of the model. Its median latency of 2,461 ms includes this misbehaviour and is not the model's speed.

8. Why We Did Not Swap

We changed nothing in production. On-device English, more than 98% of takes, keeps Parakeet. Voxtral Mini 3B makes about 16% fewer word errors on read speech (4.38% against 5.24%; 189 errors against 226), but the interval of that gain ends within 0.05 points of zero, and, going by the latency above, we expect its language-model decoder to be noticeably slower on the laptop. A gain of 0.86 points is one fewer error in about 116 words, while the delay after releasing the key is felt on every take. Because Mini was not run on the laptop, that delay is an expectation, not a measurement.

On the server lane, Voxtral Small is the only open model whose gain over the production server clears zero at 95% (not at 99%): 0.88 points, about 18% fewer errors (169 against 207). It needs about 46 GB of GPU memory and about one second per clip against a tenth of one, and lists no Icelandic. A marginal gain is not worth a dedicated card.

For Icelandic, none of the new candidates lists the language, so Whisper stays.

9. What These Numbers Will Not Carry

The sample is 200 clips per language, from one clean, read, news-and-encyclopedia corpus for English and one crowd-read corpus for Icelandic. One point is about 43 words of 4,315. Each engine ran once, and latencies are single-pass wall times in three different environments.

FLEURS and Samromur are public datasets and may be inside some models' training data; we cannot test that, and exposure may differ by model. Whether the first 5,000 Samromur rows are in random order is not recorded. These Samromur utterances are not the set used in earlier, smaller internal Icelandic tests, so the 18.57% and 19.70% figures cannot be compared with those, and the same holds for earlier English figures on 40 clips with an earlier Parakeet version.

The script that produced the on-device hypotheses was not retained; we know the engines and settings from notes, so those two arms cannot be regenerated from the record. The usage aggregates come from the laptop's daemon log, which is not part of the retained benchmark material. The bootstrap resamples clips, not speakers or sentences, and clips are not independent if one speaker recorded many (not checked). There was no multiplicity correction across five or six comparisons on English and one on Icelandic.

We never ran the hosted entries, we recorded the chart's contents rather than rechecking them, and our test sets differ from its composite. We cannot say whether the public chart's ordering reflects the same effect as our 0.9 to 1.3 points on read speech.

The deepest gap is that we have no transcribed sample of the speech we actually dictate: mixed technical vocabulary, proper names, accents and microphone noise. An earlier internal finding is that scoring against another recogniser's output inherits its hallucinations (duplicated words, invented text on silence), so such a set must be human-transcribed. We have none yet.

What survives: on 200 read FLEURS clips, Voxtral Small 24B has the lowest point WER of the tested engines and clearly beats Parakeet, but not Voxtral Mini, and only marginally the production server. The production server beats the laptop's greedy Whisper on Icelandic. Fun-ASR-Nano, in our port, is last among the valid engines. The open winners' latency and memory cost is large. What does not survive: any statement about accuracy on our own dictation, any ranking of Voxtral Mini against the production server, any statement about Voxtral Realtime, and any statement that no open model could cover Icelandic, since we only checked language lists. Whether Voxtral Mini 3B is acceptable on the laptop, and how the engines compare on speech with silences and hesitations, remain open.

10. What We Changed, and the Next Test

We recorded the rule and kept the engines. A candidate that is a point better on clean read speech is not a candidate until measured on our own recorded speech against human-checked references. The next test we would run is to build that set, with enough speakers, recording conditions and silences to resemble real use, transcribe it by hand, and run the shortlist on it: Voxtral Small on the server lane, and Voxtral Mini 3B on the laptop, where its latency would be measured for the first time.

11. Methods and Artefacts

Software: transformers 5.17.0, torch 2.11.0 (CUDA 12.8 build), mistral-common 1.12.0, jiwer 4.0.0, soundfile 0.14.0 and numpy 2.2.6. Scoring uses the Open ASR Leaderboard English normaliser and the basic multilingual normaliser shipped in transformers. Hardware is described in Section 2.

Statistics: a paired clip-level bootstrap with 2,000 resamples (seed 0), 95% percentile intervals at ranks 50 and 1,949. The supplementary check used 10,000 resamples with a different seed and also reports 99% intervals. Usage statistics come from the on-device daemon's log; text content was not read.

We retain the clip lists and references for both languages (the audio is public-dataset material), the staging, inference, HTTP and scoring scripts, every engine's hypothesis file with per-clip latency, and the job and per-engine logs. The script that generated the on-device hypotheses and the model weight caches are not retained.

PureTensor operates a sovereign AI infrastructure fleet (on-premises GPU compute, storage, and serving) run day-to-day by autonomous agents under human direction. This paper is PT-R-2026-057. All measurements were taken on the morning of 21 October 2026 and the usage aggregates cover 7 to 21 October; raw artefacts are retained.