Open Speech Recognition Models Beyond the Leaderboard: Nemotron 3.5 Transcribe and Qwen3-ASR on Public Audio

Transcribe.so

Research report, version 1, 2026. Measurements taken 2026-10-29 to 2026-10-30.
https://transcribe.so/papers/open-asr-models-2026-10.pdf

Abstract

We compared two open speech recognition models, NVIDIA Nemotron 3.5 Transcribe 30B-A4B and Qwen3-ASR-1.7B, on 1,830 items (about 7.3 hours) of public audio: FLEURS read speech in 18 languages, LibriSpeech test-clean, excerpts of an AMI meeting and Earnings-22 earnings calls. Both models ran through the same hosted inference API, and Nemotron also ran locally on Apple silicon from NVIDIA's official weights, at 4-bit and in NVIDIA's NVFP4 checkpoint. On read speech the models were close in most languages: with numbers spelled out on both sides, the pooled FLEURS word error rate over 9 word-scored languages was 7.3% (95% CI 6.7 to 7.9) for Qwen3-ASR-1.7B on the hosted API and 8.1% (7.4 to 8.9) for Nemotron run locally, with large gaps either way in single languages. Three findings are invisible in a single word-error-rate column. First, out of the box Nemotron wrote numbers as words on all three runtimes we tried, which added 6.9 to 7.0 points of raw word error rate on earnings calls; the hosted API offered no way to change the instruction that controls this. Second, in our local 4-bit build, the model card's written-numbers instruction coincided with runs of at least 15 silently deleted words in 3 of 5 excerpts from one meeting; with the default instruction the same excerpts had none. This association rests on 5 excerpts and was not reproduced on a reference runtime. Third, languages outside the model card's list came back from the hosted API with no warning and error rates of 34.3% (Persian) and 36.7% (Malay). Locally, Nemotron needed 18.2 to 24.4 GiB of memory per stream against 4.4 GiB for an 8-bit Qwen3-ASR-1.7B build. Per-item scores are published.

1 Introduction

Speech recognition models are usually compared by one word error rate per test set. That number answers a narrow question: how often the words differ from a normalized reference. Teams that run an open model themselves meet other questions first. Does the output write numbers the way a reader expects? Does a documented instruction change behave the same on every runtime? What happens on audio in a language the model was not trained for? Will it fit in the memory available?

This report measures those questions for two open models released with weights on Hugging Face, NVIDIA Nemotron 3.5 Transcribe 30B-A4B [1] and Qwen3-ASR-1.7B [2], on public test sets only. We ran both through one hosted inference API, and ran Nemotron locally on an Apple M4 Pro laptop through the open-source mlx-audio runtime. A companion blog post presents the same results for a general audience: `https://transcribe.so/blog/nemotron-transcribe-vs-qwen3-asr-2026`.

The authors sell a transcription service. The Conflict of interest section describes that interest and the design choices made because of it.

2 Systems

Table 1 lists the arms. Arms A and B are the two models on the same hosted inference API, which we do not name; its instruction to the model is fixed: our requests could not change it. Both hosted arms received each clip's language as a hint. Arms C and D run NVIDIA's official weights (Hugging Face revision `9d8cd6b7`), which we converted to MLX at 4-bit on the CPU, with mlx-audio 0.6.3 [3]: arm C uses a written-numbers instruction asking for all spoken words, verbal fillers and numbers in their standard written form (paraphrased from the model card; the exact string passed in the run was not logged); arm D uses mlx-audio's default instruction, "Produce a verbatim transcript of the audio. Preserve named entities, abbreviations, numbers, dates, measurements, acronyms, and technical terms as clearly as possible. Keep the same language as spoken and do not translate." Arms F and G run NVIDIA's NVFP4 checkpoint (revision `66c6adc4`) on mlx-audio's main branch at commit `5f42a925d3a1ce76b8d78876bc7dd3bedd050fe3`, which added word timestamps for the official checkpoints [4], with word timestamps and speculative decoding on: arm F with the default transcription task, arm G with the speaker-attributed task, whose instruction also asks for written numbers. The local arms received no language hint. Arms E and H were run for memory only. All local runs used batch size 1 on a 48 GB Apple M4 Pro laptop.

Table 1. Systems compared. Arms E and H were run for memory only.

Arm	Model	Runtime	Precision	Instruction	Language hint	Tested
A	Qwen3-ASR-1.7B	hosted inference API (third party, not named)	as reported by the service	none (the language of each clip was passed as a hint)	yes	2026-10-29
B	NVIDIA Nemotron 3.5 Transcribe 30B-A4B	the same hosted inference API as arm A	as reported by the service	fixed by the service; requests could not change it	yes	2026-10-29
C	NVIDIA Nemotron 3.5 Transcribe 30B-A4B	local, Apple M4 Pro 48 GB, batch 1, mlx-audio 0.6.3	MLX 4-bit (language model), converted by the authors	written-numbers instruction (all spoken words and verbal fillers, numbers in standard written form), paraphrased from the model card; the exact string passed in the run was not logged	no	2026-10-29 to 2026-10-30
D	NVIDIA Nemotron 3.5 Transcribe 30B-A4B	as arm C	as arm C	mlx-audio default: "Produce a verbatim transcript of the audio. Preserve named entities, abbreviations, numbers, dates, measurements, acronyms, and technical terms as clearly as possible. Keep the same language as spoken and do not translate."	no	2026-10-30
F	NVIDIA Nemotron 3.5 Transcribe 30B-A4B (NVIDIA NVFP4 checkpoint)	local, Apple M4 Pro 48 GB, batch 1, mlx-audio main at commit 5f42a925d3a1ce76b8d78876bc7dd3bed-d050fe3 (pull request 1092), word timestamps and speculative decoding on	NVFP4 (published by NVIDIA)	mlx-audio default transcription task (same text as arm D)	no	2026-10-30
G	NVIDIA Nemotron 3.5 Transcribe 30B-A4B (NVIDIA NVFP4 checkpoint)	as arm F	as arm F	mlx-audio speaker-attributed task: the model card's speaker-tag instruction, which also asks for numbers in standard written form	no	2026-10-30
E	Qwen3-ASR-1.7B	local, Apple M4 Pro 48 GB, batch 1, an MLX 8-bit build (memory baseline only)	MLX 8-bit	none	no	2026-10-29
H	NVIDIA Nemotron 3.5 Transcribe 30B-A4B	local, Apple M4 Pro 48 GB, batch 1, mlx-audio 0.6.3 (memory only)	MLX 8-bit	as arm C	no	2026-10-30

3 Corpora

All audio comes from public test sets (Table 2). FLEURS [5] supplies single read sentences in many languages; for the three-arm comparison we used the clips all of arms A, B and C ran, and a larger set of clips for the two hosted arms (Appendix A). LibriSpeech test-clean [6] supplies read English with forced-aligned word times. For long-form audio we concatenated consecutive utterances into windows: 5 excerpts of about 148 s from a single AMI meeting, EN2002a, headset mix [7], and 24 windows of about 162 s from 3 Earnings-22 earnings calls [8]. The reference for each window is the concatenation of its utterances' reference transcripts. Earnings-22 is licensed CC BY-SA; we republish its item identifiers and scores only, never its text. No customer or private audio was used.

Table 2. Corpora. All audio is public; no customer or private audio was used.

Set	Source	Split	License	Unit	Items	Mean length
FLEURS	google/fleurs	test	CC BY 4.0	single read sentences	1,721	12.2 s
LibriSpeech	LibriSpeech test-clean, word alignments from gilkeyio/librispeech-alignments (Montreal Forced Aligner)	test_-clean	CC BY 4.0	single read utterances	80	8.3 s
AMI	edinburghcstr/ami, config ihm	test	CC BY 4.0	consecutive headset utterances of meeting EN2002a concatenated into windows of about 150 s	5	148.3 s
Earnings-22	distil-whisper/earnings22, config chunked	test	CC BY-SA 4.0 (scores and item ids only are republished; no text)	consecutive segments of three earnings calls concatenated into windows of about 165 s	24	162.5 s

4 Metrics

- **Word and character error rate.** Word error rate (WER), the edit distance between normalized reference and output divided by the number of reference words [9]; character error rate (CER) for Mandarin Chinese, Japanese and Korean. Normalization lowercases, folds Unicode, removes punctuation, bracketed tags and a short list of fillers, the same primitives as our earlier caption study [10]. Rates are pooled as total edits over total reference units.
- **Raw and numbers-spelled.** Raw rates count a year spoken and written as words ("twenty twenty one") against the same year in digits as errors. The numbers-spelled rate first writes every number as words on both sides with num2words [11] (in English, four-digit numbers in the range of recent centuries are read as years), so it measures recognition rather than number formatting. Raw minus spelled is the cost of number formatting, in percentage points.
- **Digit-group ratio.** Digit groups in the output divided by digit groups in the reference: one means numbers come back as digits as often as the reference writes them, zero means none do. It can exceed one when a system splits numbers into more groups than the reference.

- **Skipped passages.** A run of at least 15 consecutive reference words deleted in the word alignment of one long-form window. We count it on AMI and Earnings-22 windows only; on short clips a run of that length would be most of the clip.
- **Statistics.** 95% intervals are paired percentile bootstrap intervals over items [12], 2,000 resamples with a fixed seed (2861); pooled rows resample within each language. Where a comparison has 5 items or fewer we print "n too small for a CI" instead of an interval.

5 Results

5.1 *Accuracy on read speech*

Across the 990 FLEURS clips in 9 word-scored languages that all three arms ran, the pooled numbers-spelled WER was 7.3% (95% CI 6.7 to 7.9) for Qwen3-ASR-1.7B on the hosted API, 9.0% (8.2 to 9.8) for Nemotron on the same API, and 8.1% (7.4 to 8.9) for Nemotron run locally at 4-bit. The paired difference between the local Nemotron build and Qwen3-ASR-1.7B was 0.8 points (0.3 to 1.4). Single languages differed more than the pooled figure suggests (Table 3): the local Nemotron build was 2.7 points lower than Qwen3-ASR-1.7B on Arabic (1.2 to 4.1), while Qwen3-ASR-1.7B was far lower on Indonesian (4.9% against 16.3% hosted and 17.1% local) and on Mandarin Chinese (4.9% against 9.1% and 7.3% CER). English (10) and Japanese (10) clips are too few for more than an indication.

Table 3. FLEURS test, error rate with numbers spelled out on both sides, on the clips all three arms ran (95% CI). * fewer than 50 clips: indicative only. Differences are paired, in percentage points; negative means lower error than Qwen3-ASR-1.7B.

Language	Metric	Clips	A: Qwen3-ASR-1.7B, hosted	B: Nemotron, hosted	C: Nemotron, local 4-bit	B minus A, pts	C minus A, pts
Arabic	WER	120	14.2% (11.8 to 16.7)	14.0% (11.7 to 16.7)	11.5% (9.4 to 14.1)	-0.2 (-2.1 to 1.7)	-2.7 (-4.1 to -1.2)
German	WER	120	5.3% (4.0 to 6.9)	7.0% (5.1 to 9.1)	5.6% (3.9 to 7.4)	+1.6 (0.2 to 3.3)	+0.2 (-0.8 to 1.3)
English*	WER	10	8.5% (1.6 to 21.6)	8.5% (0.9 to 24.1)	10.0% (2.6 to 24.1)	0.0 (-2.9 to 3.1)	+1.5 (0.0 to 3.6)
Spanish	WER	130	3.5% (2.6 to 4.5)	3.3% (2.3 to 4.6)	2.9% (2.0 to 3.9)	-0.2 (-1.0 to 0.7)	-0.6 (-1.3 to 0.0)
French	WER	120	5.3% (4.1 to 6.7)	5.5% (4.3 to 6.8)	5.4% (4.2 to 6.6)	+0.2 (-0.9 to 1.3)	0.0 (-1.1 to 1.2)
Hindi	WER	130	11.7% (9.8 to 13.6)	13.4% (10.0 to 17.3)	11.0% (8.3 to 14.2)	+1.7 (-1.2 to 5.0)	-0.7 (-3.0 to 2.1)
Indonesian	WER	120	4.9% (3.4 to 6.8)	16.3% (13.9 to 18.7)	17.1% (14.5 to 19.9)	+11.4 (9.6 to 13.4)	+12.2 (10.2 to 14.3)
Japanese*	CER	10	5.0% (2.7 to 7.9)	0.9% (0.0 to 2.0)	2.0% (0.8 to 3.4)	-4.1 (-7.3 to -1.3)	-2.9 (-5.5 to -0.6)
Korean	CER	130	4.1% (2.4 to 6.5)	3.9% (2.6 to 5.5)	3.7% (2.4 to 5.4)	-0.2 (-2.4 to 1.1)	-0.4 (-2.0 to 0.7)
Portuguese	WER	120	6.4% (4.4 to 8.8)	5.9% (4.1 to 8.2)	5.5% (3.6 to 7.9)	-0.4 (-1.3 to 0.4)	-0.9 (-1.5 to -0.1)
Russian	WER	120	8.2% (6.3 to 10.4)	9.4% (7.2 to 12.1)	8.6% (6.5 to 11.1)	+1.2 (-0.3 to 2.8)	+0.4 (-0.8 to 1.8)
Mandarin Chinese	CER	130	4.9% (3.4 to 6.7)	9.1% (7.1 to 11.3)	7.3% (5.4 to 9.4)	+4.1 (2.9 to 5.4)	+2.4 (1.2 to 3.7)
Pooled, 9 word-scored languages	**WER**	**990**	**7.3%** (6.7 to 7.9)	**9.0%** (8.2 to 9.8)	**8.1%** (7.4 to 8.9)	**+1.7** (1.1 to 2.4)	**+0.8** (0.3 to 1.4)

On LibriSpeech test-clean (80 clips) the rates were 1.71% for Qwen3-ASR-1.7B and 1.45% for Nemotron on the hosted API, and 1.39% for the NVFP4 build (Table 4); every paired interval includes zero, so this sample does not separate the models on clean English read speech.

Table 4. LibriSpeech test-clean, 80 clips.

Arm	WER, numbers spelled (95% CI)	Minus A, pts (95% CI)
A: Qwen3-ASR-1.7B, hosted	1.71% (1.03 to 2.52)	
B: Nemotron, hosted	1.45% (0.85 to 2.16)	-0.27 (-0.72 to 0.17)
C: Nemotron 4-bit, written numbers	1.66% (0.95 to 2.46)	-0.05 (-0.57 to 0.43)
F: Nemotron NVFP4, default	1.39% (0.79 to 2.10)	-0.32 (-0.74 to 0.05)

5.2 *Number formatting on earnings calls*

On 24 Earnings-22 windows, Nemotron's raw WER was 13.1% on the hosted API against 7.1% for Qwen3-ASR-1.7B (Table 5). With numbers spelled on both sides the two were 6.2% (5.3 to 7.1) and 6.6% (5.6 to 7.8), a paired difference of -0.5 points (-1.2 to 0.2). The gap in raw WER is number formatting. Without an instruction that asks for written numbers, Nemotron spelled numbers out on all three runtimes we tried (arms B, D and F, digit-group ratio 0.00 each), which cost 6.9 to 7.0 points of raw WER; Qwen3-ASR-1.7B returned digits (0.98). The hosted API applies its own instruction, so a hosted user could not request written numbers. Locally, the model card's written-numbers instruction raised the ratio to 0.36 (arm C) and the speaker-attributed task to 0.25 (arm G), short of the reference.

Table 5. Earnings-22, 24 windows (8,480 reference words).

Arm	Raw WER (95% CI)	WER, numbers spelled (95% CI)	Raw minus spelled, pts	Digit-group ratio
A: Qwen3-ASR-1.7B, hosted	7.1% (6.0 to 8.3)	6.6% (5.6 to 7.8)	0.5	0.98
B: Nemotron, hosted	13.1% (11.1 to 15.1)	6.2% (5.3 to 7.1)	7.0	0.00
C: Nemotron 4-bit, written numbers	11.5% (9.3 to 13.8)	6.6% (5.6 to 7.6)	4.9	0.36
D: Nemotron 4-bit, default	13.5% (11.4 to 15.5)	6.5% (5.7 to 7.4)	6.9	0.00
F: Nemotron NVFP4, default	13.8% (11.7 to 16.0)	6.9% (6.0 to 7.9)	6.9	0.00
G: Nemotron NVFP4, speaker-attributed	12.4% (10.2 to 14.7)	7.1% (6.2 to 8.0)	5.3	0.25

5.3 *Instructions and skipped passages*

With the written-numbers instruction, our local 4-bit build (arm C) deleted runs of at least 15 consecutive reference words in 3 of 5 AMI excerpts; the longest run in each affected excerpt was 155, 32, 17 words, and the output carried no sign of the gap. Its WER over the five excerpts was 22.1%, against 12.1% for the same build with mlx-audio's default instruction (arm D), 13.4% for Qwen3-ASR-1.7B and 13.7% for Nemotron on the hosted API (Table 6). The same build and instruction had 0 of 24 Earnings-22 windows with such a run. NVIDIA's NVFP4 build with the default task (arm F) had one run of 22 words in 1 of its 29 long windows (Table 7).

We report this as an association observed on 5 excerpts of a single meeting, not as a property of the model or of NVIDIA's instruction. We did not reproduce it on a reference runtime or at higher precision, and our conversion, the port, the instruction and their interaction are all candidate explanations. Anyone running the model on long audio can guard against silent gaps by comparing output length with audio duration or by aligning the transcript to the audio.

Table 6. AMI meeting EN2002a, 5 excerpts, WER. Runs: deleted runs of at least 15 reference words. n too small for a CI.

Excerpt	A: Qwen3-ASR-1.7B, hosted	B: Nemotron, hosted	C: Nemotron 4-bit, written numbers	D: Nemotron 4-bit, default	F: Nemotron NVFP4, default
1	14.7%	11.1%	10.2%	12.0%	9.6%
2	14.0%	17.3%	55.2% (runs: 155, 30)	14.2%	12.4%
3	13.8%	12.6%	19.0% (runs: 32)	12.4%	12.4%
4	9.7%	17.3%	17.8% (runs: 17)	10.9%	9.3%
5	14.5%	10.3%	7.1%	11.1%	11.7%
All	**13.4%**	**13.7%**	**22.1%**	**12.1%**	**11.2%**

Table 7. Long-form windows with at least one run of 15 or more consecutive deleted reference words (run lengths in words). Windows that failed to complete are not counted.

Arm	AMI excerpts with a run	Earnings-22 windows with a run
A: Qwen3-ASR-1.7B, hosted	0 of 5	0 of 24
B: Nemotron, hosted	0 of 5	0 of 24
C: Nemotron 4-bit, written numbers	3 of 5 (155, 32, 30, 17)	0 of 24
D: Nemotron 4-bit, default	0 of 5	0 of 24
F: Nemotron NVFP4, default	0 of 5	1 of 24 (22)
G: Nemotron NVFP4, speaker-attributed	0 of 3	0 of 24

On the 3 excerpts where every NVFP4 run completed, the speaker-attributed task (arm G) had a WER of 8.3% against 14.3% for Qwen3-ASR-1.7B on the hosted API (Table 8). Of the other excerpts, 2 failed in arm G: one with an alignment error from the word-timestamp step and one with a GPU timeout.

Table 8. AMI, the 3 excerpts every NVFP4 arm completed. n too small for a CI.

Arm	WER
A: Qwen3-ASR-1.7B, hosted	14.3%
B: Nemotron, hosted	11.4%
F: Nemotron NVFP4, default	11.4%
G: Nemotron NVFP4, speaker-attributed	8.3%

5.4 *Hosted and local runs of the same model*

Nemotron under the same model id did not behave identically on the two runtimes. The clearest gap was Mandarin Chinese: on 130 clips, CER was 9.1% on the hosted API and 7.3% for our local 4-bit build, a difference of 1.8 points; German showed a smaller gap in the same direction (7.0% hosted, 5.6% local, 5.3% for Qwen3-ASR-1.7B). We do not attribute these differences to a cause. The two runtimes differ in precision (bf16 as reported by the service, against our 4-bit conversion), in the instruction (fixed and unseen on the hosted side), in the language hint (sent to the hosted API only) and possibly in decoding settings we cannot see; and the hosted weights cannot be verified against a revision.

5.5 *Languages outside the model card*

Persian and Malay are not among the languages listed on the Nemotron 3.5 Transcribe model card. Sent through the hosted API, every clip still returned a transcript, with no error and no warning, at 34.3% WER for Persian and 36.7% for Malay, against 25.4% and 10.3% for Qwen3-ASR-1.7B on the same clips (Table 9). A caller that does not check the language first gets a plausible-looking transcript with a high error rate.

Table 9. FLEURS test, languages not listed on the Nemotron 3.5 Transcribe model card (WER, 95% CI). * fewer than 50 clips: indicative only.

Language	Clips	A: Qwen3-ASR-1.7B, hosted	B: Nemotron, hosted
Persian*	27	25.4% (20.9 to 30.5)	34.3% (29.7 to 39.0)
Malay*	29	10.3% (6.5 to 14.3)	36.7% (30.9 to 42.4)

5.6 *Memory and deployment*

Peak memory per stream on the laptop was 18.2 to 19.7 GiB for the 4-bit Nemotron build and up to 24.4 GiB for the NVFP4 build with word timestamps, against at most 4.4 GiB for an 8-bit Qwen3-ASR-1.7B build (Table 10). An 8-bit Nemotron conversion peaked at 33.3 GiB and above and completed 3 of 4 meeting excerpts before a GPU timeout. Converting the official weights to MLX on the GPU also hit a GPU timeout, so we converted on the CPU. Word timestamps for the official checkpoints reached mlx-audio's main branch during our test window and were not yet in a release [4]. We do not report speed: after the GPU timeout the laptop's throughput was not stable, and a single laptop is not a serving benchmark.

Table 10. Peak process memory reported by MLX, Apple M4 Pro laptop, 48 GB, batch 1.

Arm	Peak memory	Items completed
E: Qwen3-ASR-1.7B, MLX 8-bit	3.2 to 4.4 GiB	70 of 70
C: Nemotron 4-bit, written numbers	18.2 to 19.7 GiB	1369 of 1369
D: Nemotron 4-bit, default	19.7 to 19.7 GiB	29 of 29
F: Nemotron NVFP4, default	23.0 to 24.4 GiB	109 of 109
G: Nemotron NVFP4, speaker-attributed	23.5 to 24.4 GiB	27 of 29
H: Nemotron, MLX 8-bit	33.3 to 34.4 GiB	3 of 4

6 Limitations

1. **The skipped passages were not reproduced on a reference runtime or at higher precision.** They come from our own 4-bit MLX conversion, 5 excerpts of one meeting and one written-numbers instruction (paraphrased from the model card; the exact string passed in the run was not logged), so they may stem from the quantization, the port, the instruction or their combination. The NVFP4 build also dropped one run with the default instruction.
2. **Hosted and local runs are confounded.** They differ in precision, instruction, language hint and possibly decoding, and the hosted weights are unverifiable. Differences between them are measured, not explained.
3. **Small samples.** The meeting results rest on one AMI meeting; English (10) and Japanese (10) clips in the three-arm FLEURS subset are indicative only; Appendix A languages with about thirty clips are flagged.
4. **Read speech dominates.** FLEURS and LibriSpeech are read sentences. Conversational, noisy or accented speech beyond the AMI and Earnings-22 windows was not measured.
5. **Normalization choices.** Spelling numbers on both sides depends on num2words support per language, and our filler list and punctuation rules affect absolute rates for every system equally.
6. **One point in time.** Hosted services and open-source runtimes change quickly; these results describe 2026-10-29 to 2026-10-30.

7 Conflict of interest

This report was designed, run and written by Transcribe.so, which sells a transcription service. No external party funded, reviewed or audited it, and neither NVIDIA nor the Qwen team was involved. The design choices made because of this interest: only public audio was used, so that anyone can re-run the comparison; both models were run on the same hosted service and scored with the same scorer; every figure in this report is generated from the published results files by script, and per-item scores are published; we report results that favour each model; and the hosted service is not named, because results for a single service can change without notice and the finding that matters, that a hosted wrapper can fix the instruction and the output format, is general.

8 Data availability

The following files are published on transcribe.so. They contain scores and item identifiers only, no transcripts or reference text.

- Per-table results with intervals (CSV): `https://transcribe.so/data/open-asr-models-2026-10.csv`

- Full results, arm descriptions, model revisions, corpus metadata and per-item scores (JSON): `https://transcribe.so/data/open-asr-models-2026-10-results.json`
- Model cards: `https://huggingface.co/nvidia/NVIDIA-Nemotron-3.5-Transcribe-30B-A4B`, `https://huggingface.co/nvidia/NVIDIA-Nemotron-3.5-Transcribe-30B-A4B-NVFP4`, `https://huggingface.co/Qwen/Qwen3-ASR-1.7B`

The fetching, inference and scoring scripts are in the authors' repository, which is not public at the time of writing. The full Nemotron revision is `9d8cd6b79bfdd427b6558a5fdbf76fcaf8c75341` and the NVFP4 revision `66c6adc4020179567074d0399cda2a4eeb2bb347`.

References

1. NVIDIA (2026). NVIDIA Nemotron 3.5 Transcribe 30B-A4B, model card. https://huggingface.co/nvidia/NVIDIA-Nemotron-3.5-Transcribe-30B-A4B
2. Qwen Team. Qwen3-ASR-1.7B, model card. https://huggingface.co/Qwen/Qwen3-ASR-1.7B
3. mlx-audio contributors. mlx-audio: text-to-speech, speech-to-text and speech-to-speech on Apple silicon with MLX. https://github.com/Blaizzy/mlx-audio
4. mlx-audio pull request 1092 (merged 2026-10-29): word timestamps for the official Nemotron 3.5 Transcribe checkpoints. https://github.com/Blaizzy/mlx-audio/pull/1092
5. Conneau, A., Ma, M., Khanuja, S., Zhang, Y., Axelrod, V., Dalmia, S., Riesa, J., Rivera, C., and Bapna, A. (2023). FLEURS: Few-shot learning evaluation of universal representations of speech. *IEEE Spoken Language Technology Workshop (SLT 2022)*.
6. Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). LibriSpeech: an ASR corpus based on public domain audio books. *ICASSP 2015*, 5206-5210. Word alignments: McAuliffe, M., et al. (2017), Montreal Forced Aligner, *Interspeech 2017*.
7. Carletta, J., et al. (2005). The AMI meeting corpus: a pre-announcement. *Machine Learning for Multimodal Interaction (MLMI 2005)*.
8. Del Rio, M., et al. (2022). Earnings-22: a practical benchmark for accents in the wild. arXiv:2203.15591.
9. Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8), 707-710.
10. Transcribe.so (2026). YouTube automatic captions against creator captions. Research report. https://transcribe.so/papers/youtube-auto-captions-2026-10.pdf
11. num2words contributors. num2words: convert numbers to words in multiple languages, version 0.5.14. https://github.com/savoirfairelinux/num2words
12. Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, 7(1), 1-26.

How to cite

Transcribe.so (2026). *Open Speech Recognition Models Beyond the Leaderboard: Nemotron 3.5 Transcribe and Qwen3-ASR on Public Audio.* Research report. `https://transcribe.so/papers/open-asr-models-2026-10.pdf`

```
@techreport{transcribeso2026openasr,
  author      = {{Transcribe.so}},
  title       = {Open Speech Recognition Models Beyond the Leaderboard: Nemotron 3.5 Transcribe and Qwen3-ASR on
Public Audio},
  institution = {Transcribe.so},
  year        = {2026},
  url         = {https://transcribe.so/papers/open-asr-models-2026-10.pdf}
}
```

Appendix A. Hosted runs on the larger FLEURS sample

Every FLEURS clip both hosted arms ran, excluding the two languages reported as unsupported. This sample is larger than the three-arm subset for most languages and adds languages with about thirty clips each.

Table 11. FLEURS test, every clip both hosted arms ran, numbers spelled on both sides (95% CI). * fewer than 50 clips: indicative only.

Language	Metric	Clips	A: Qwen3-ASR-1.7B, hosted	B: Nemotron, hosted	B minus A, pts
Arabic	WER	149	14.4% (12.2 to 16.8)	14.4% (12.1 to 16.9)	0.0 (-1.5 to 1.6)
Czech*	WER	29	24.0% (17.9 to 31.6)	8.9% (5.6 to 12.6)	-15.0 (-21.5 to -9.6)
German	WER	150	5.3% (4.1 to 6.6)	6.5% (5.0 to 8.2)	+1.2 (0.1 to 2.6)
English*	WER	29	5.8% (2.8 to 10.1)	5.8% (2.5 to 10.7)	0.0 (-1.5 to 1.4)
Spanish	WER	148	3.8% (2.8 to 5.0)	3.7% (2.6 to 5.0)	-0.1 (-0.9 to 0.7)
French	WER	150	5.8% (4.5 to 7.0)	5.8% (4.7 to 7.1)	+0.1 (-0.9 to 1.0)
Hindi	WER	149	11.6% (9.8 to 13.5)	13.6% (10.1 to 17.6)	+2.0 (-0.8 to 5.4)
Indonesian	WER	150	5.1% (3.7 to 6.9)	16.2% (14.0 to 18.7)	+11.1 (9.3 to 12.9)
Italian*	WER	28	6.0% (2.1 to 12.1)	6.1% (2.1 to 12.3)	+0.2 (-1.6 to 1.9)
Japanese*	CER	29	8.6% (4.5 to 14.3)	7.3% (3.2 to 12.7)	-1.3 (-4.2 to 1.4)
Korean	CER	148	3.7% (2.1 to 5.8)	3.7% (2.5 to 5.1)	0.0 (-1.8 to 1.2)
Dutch*	WER	30	9.1% (6.8 to 11.5)	7.3% (4.9 to 10.1)	-1.8 (-4.3 to 0.8)
Portuguese	WER	149	6.2% (4.5 to 8.3)	5.8% (4.1 to 7.7)	-0.4 (-1.1 to 0.3)
Russian	WER	149	8.1% (6.4 to 9.7)	9.0% (7.1 to 11.0)	+0.9 (-0.3 to 2.4)
Vietnamese*	CER	28	5.0% (2.8 to 7.8)	3.9% (2.3 to 6.2)	-1.1 (-2.6 to 0.2)
Mandarin Chinese	CER	150	5.0% (3.6 to 6.9)	8.8% (6.9 to 10.8)	+3.7 (2.7 to 4.9)