OPEN ASR LEADERBOARD: TOWARDS REPRODUCIBLE AND TRANSPARENT MULTILINGUAL SPEECH RECOGNITION EVALUATION

Vaibhav Srivastav[1,*], *Steven Zheng*[1,*,†], *Eric Bezzam*[1,†],
Eustache Le Bihan[1], *Adel Moumen*[2], *Sanchit Gandhi*[3]

[1]Hugging Face, Inc., Paris, France
[2]University of Cambridge, United Kingdom, [3]Mistral AI, London, UK
[*]Equal contribution
[†] Corresponding authors: steven@huggingface.co, eric.bezzam@huggingface.co

ABSTRACT

Despite rapid progress, ASR evaluation remains saturated with short-form English, and efficiency is rarely reported. We present the *Open ASR Leaderboard*, a fully reproducible benchmark and interactive leaderboard comparing 60+ open-source and proprietary systems across 10 datasets, including a dedicated multilingual track. We standardize text normalization and report both word error rate (WER) and inverse real-time factor (RTFx), enabling fair accuracy–efficiency comparisons. For English transcription, Conformer encoders paired with LLM decoders achieve the best average WER but are slower, while CTC and TDT decoders deliver much better RTFx, making them attractive for long-form and offline use. Whisper-derived encoders fine-tuned for English improve accuracy but often trade off multilingual coverage. All code and dataset loaders are open-sourced to support transparent, extensible evaluation.

***Index Terms*—** Benchmarking, automatic speech recognition, reproducible, multilingual

1. INTRODUCTION

The expression "*jack of all trades, master of none*" is typically used to point out someone who has not gained expertise in a specific task. Interestingly, the second part of the original phrase – "*though oftentimes better than master of one*" – is usually omitted, which altogether changes its meaning. This tradeoff is highly relevant for machine learning systems: should a model excel at a single task, or perform competently across multiple tasks?

Automatic speech recognition (ASR) has seen remarkable progress in recent years, fueled in part by open-source contributions. Publicly-available datasets [1, 2] and pre-trained models [3, 4, 5] have enabled researchers across academia and industry to build on existing work. Yet, as the number of datasets and models grows, it becomes increasingly difficult for developers of new models to know which baselines to compare against and how. Similarly, users focused on inference may find it difficult to identify which model, whether open-source or proprietary, best meets their needs in terms of application and/or efficiency. Moreover, most existing benchmarks and evaluations overwhelmingly emphasize English.

Several efforts have sought to address parts of this problem, including benchmarks across multiple accents and diverse contexts in the French language [6], under noise and reverberation in far-field settings [7], comparing commercial and open-source models in English and Chinese [8]. Some common observations can be drawn from these efforts: (1) there is no "catch-all" model, (2) no single dataset is sufficient for evaluation, and (3) a single metric, *i.e.*, word error rate (WER), is not enough.

To address these challenges we introduce the *Open ASR Leaderboard*. Our contributions include:

1. An interactive leaderboard that compares 60+ open-source and proprietary models from 18 organizations, with evaluations over 10 datasets.[1]
2. A multilingual benchmark covering German, French, Italian, Spanish, and Portuguese.

For full transparency and to facilitate the addition of new models and datasets, the leaderboard's codebase is open-sourced.[2] The above model, dataset, and languages count are as of 8 Oct 2025, and will continue to increase with new additions to the leaderboard.

2. OPEN ASR LEADERBOARD

2.1. Overview

As part of the *Open ASR Leaderboard*, there are evaluations on two tasks, with the following tabs:

1. *Leaderboard*, which evaluates English transcription.

[1]Leaderboard: hf.co/spaces/hf-audio/open_asr_leaderboard
[2]Code: github.com/huggingface/open_asr_leaderboard

Table 1: Datasets used for the *Open ASR Leaderboard*. The *Task(s)* column indicates which tasks (as denoted in Section 2.1) the dataset is used for. The duration is of the test-split. The *Multilingual* datasets indicate the range of durations for the evaluated languages.

Dataset	Task(s)	Duration [h]	License	Source	Style	Transcriptions
AMI Meeting Corpus [9]	Leaderboard	9	CC-BY-4.0	Meetings	Spontaneous	Punctuated, cased
CoVoST-2 [10]	Multilingual (de/fr/it/es/pt)	5.3–23	CC-BY-NC-4.0	Open domain	Read	Punctuated, cased
Earnings22 [11]	Leaderboard	119	CC-BY-SA-4.0	Earnings calls	Oratory, spontaneous	Punctuated, cased
FLEURS [12]	Multilingual (de/fr/it/es/pt)	2.0–3.5	CC-BY-4.0	Wikipedia	Read	Punctuated, cased
GigaSpeech [2]	Leaderboard	40	apache-2.0	Audiobook, podcast, YouTube	Read, spontaneous	Punctuated
LibriSpeech (clean) [1]	Leaderboard	5.4	CC-BY-4.0	Audiobooks	Read	Normalized
LibriSpeech (other) [1]	Leaderboard	5.1	CC-BY-4.0	Audiobooks (noisier)	Read	Normalized
MLS [13]	Multilingual (fr/it/es/pt)	0.8–6.3	CC-BY-4.0	Audiobooks	Read	Normalized
SPGISpeech [14]	Leaderboard	100	User Agreement	Financial meetings	Oratory, spontaneous	Punctuated, cased
TED-LIUM v3 [15]	Leaderboard	3	CC-BY-NC-ND 3.0	TED Talks	Oratory	Normalized
VoxPopuli [16]	Leaderboard	5	CC0	European Parliament	Oratory	Punctuated

2. *Multilingual*, which currently evaluates German, French, Italian, Spanish, and Portuguese transcription.

The datasets used for evaluation are presented in Section 2.2. The models are evaluated according to WER, and can be dynamically sorted on the leaderboard page according to WER-performance on a particular dataset (more on metrics in Section 2.3). Section 2.4 presents models that are evaluated within the *Open ASR Leaderboard*, while Section 2.5 describes the process of adding a new model.

2.2. Datasets

Table 1 summarizes the datasets used for the *Open ASR Leaderboard*.

Dataset retrieval and usage is enabled through the *datasets* package [17]. The datasets themselves are hosted on the Hugging Face Hub,[3] which allows for interactive exploration, *e.g.*, listening to individual audio, viewing their metadata, and performing SQL queries, all from the browser and without having to download the datasets. With the above engineering choices, the datasets can be conveniently downloaded and used in Python, as shown below:

Listing 1: Example of loading dataset.

```
from datasets import load_dataset

ds = load_dataset("hf-audio/esb-datasets-test-only
    -sorted", "ami", split="test")
audio_sample = ds[0]
```

2.3. Metrics

We report results on two metrics: *word error rate* (WER) for comparing transcription quality, and *inverse real-time factor* (RTFx) for comparing inference speed.

To account for differences in punctuation and casing between model outputs and the dataset transcriptions (see Table 1), we normalize text before computing WER. Punctuation and casing are removed, and an English text normalization pipeline, closely following Whisper [3], is applied. This includes number normalization (*e.g.*, "0" to "zero"), spelling standardization, and filler word removal (*e.g.*, "uh", "mhm").

RTFx is defined as:

$$\text{RTFx} = \frac{\text{Total duration of audio}}{\text{Transcription time}}, \tag{1}$$

with a higher value being better (*i.e.*, lower latency). RTFx can be computed on a single audio or a batch. RTFx (and variants) is a useful metric for quantifying a model's efficiency in processing long audio samples [18].

[3] *Leaderboard*: hf.co/datasets/hf-audio/esb-datasets-test-only-sorted; *Multilingual*: hf.co/datasets/nithinraok/asr-leaderboard-datasets

Table 2: Distribution of encoder and decoder architectures for the open-source models in the *Open ASR Leaderboard*. Some models use hybrid architectures for the encoder or decoder, and are counted twice.

Enc ↓ / Dec →	Transformer	CTC	RNN-T/TDT	LLM	Total
Conformer-based	5	7	8	4	24
Whisper	18	0	0	2	20
Self-supervised	0	13	0	0	13
Custom	3	0	0	0	3
Total	26	20	8	6	60

2.4. Current models

Of the 64 models currently listed in the *Open ASR Leaderboard* (as of 8 Oct 2025), 57 are open-source. The 64 models come from 18 organizations: NVIDIA (18), Meta (12), OpenAI (8), Hugging Face (5), University of Washington (4), IBM (2), Rev AI (2), SpeechBrain (2), Useful Sensors (2), AssemblyAI (1), Aqua Voice (1), ElevenLabs (1), Kyutai (1), Microsoft (1), Mistral AI (1), Nyra Health (1), Speechmatics (1), and Ultravox (1).

Table 2 provides a breakdown of the encoder and decoder architectures of the models in the leaderboard, while some of the models are presented in our results in Tables 3 and 4. Conformer-based encoders [20, 19], Whisper-based encoders [3], self-supervised encoders (*i.e.*, wav2vec2 [4], HuBERT [5], data2vec [25]) and custom encoders [22, 26]

Table 3: Subset of *Open ASR Leaderboard* results on English transcription. WER is averaged over datasets corresponding to the *Leaderboard* in Table 1. *Whisper-FT* stands for Whisper-finetuned.

Model	Avg. WER ↓	Avg. WER rank	RTFx ↑	Open	Encoder	Decoder	# Lang.
NVIDIA Canary Qwen 2.5B	**5.63**	1	418.28	Yes	FastConformer [19]	LLM-based	1
IBM Granite Speech 3.3 8B	5.74	2	145.42	Yes	Conformer [20]	LLM-based	5
IBM Granite Speech 3.3 2B	6.00	3	259.57	Yes	Conformer [20]	LLM-based	5
Microsoft Phi 4 Multimodal Instruct	6.02	4	151.1	Yes	Conformer [20]	LLM-based	8
NVIDIA Parakeet TDT 0.6B v2	6.05	5	3386.02	Yes	FastConformer [19]	TDT [21]	1
Aqua Voice Avalon	6.24	6	-	No	-	-	1
NVIDIA Parakeet TDT 0.6B v3	6.32	7	3332.74	Yes	FastConformer [19]	TDT [21]	25
NVIDIA Canary 1B Flash	6.35	8	1045.75	Yes	FastConformer [19]	Transformer	4
Kyutai STT 2.6B en	6.40	9	88.37	Yes	Mimi codec [22]	Transformer	1
NVIDIA Canary 1B	6.50	10	235.34	Yes	FastConformer [19]	Transformer	4
Nyra Health CrisperWhisper	6.67	11	84.05	Yes	Whisper-FT [23]	Whisper-FT	1
ElevenLabs Scribe v1	6.88	12	-	No	-	-	99
Speechmatic Enhanced	6.91	13	-	No	-	-	55
Mistral AI Voxtral Mini 3B	7.05	16	109.86	Yes	Whisper-FT [24]	LLM-based	8
RevAI Fusion	7.12	18	-	No	-	-	1
NVIDIA Canary 1B v2	7.15	20	749	Yes	FastConformer [19]	Transformer	25
Distil-Whisper Large v3.5	7.21	21	202.03	Yes	Whisper [3]	Transformer	1
NVIDIA Parakeet CTC 1.1B	7.40	23	2728.52	Yes	FastConformer [19]	CTC	1
OpenAI Whisper Large v3	7.44	25	145.51	Yes	Whisper [3]	Whisper	99
NVIDIA FastConformer CTC Large	8.96	44	**6399.25**	Yes	FastConformer [19]	CTC	1
SpeechBrain ASR wav2vec2 Librispeech	14.4	52	451.18	Yes	wav2vec2 [4]	CTC	1

Table 4: Average WERs for different languages (German/French/Italian/Spanish/Portuguese) on the *Multilingual* datasets in Table 1.

Model	DE	FR	IT	ES	PT
Microsoft Phi 4 Multimodal Instruct	**4.50**	5.13	**4.80**	3.59	5.15
NVIDIA Canary 1B v2	4.96	**4.86**	5.66	**3.22**	6.23
OpenAI Whisper Large v3	4.97	6.59	5.14	3.32	**4.38**
NVIDIA Parakeet TDT 0.6B v3	4.90	5.38	5.58	3.72	5.95
Mistral AI Voxtral Mini 3B	5.36	5.96	5.88	3.81	4.80
ElevenLabs Scribe v1	10.8	15.1	7.71	9.28	22.8

are considered. Whisper-based encoders either use the encoder model as is [27], apply low-rank adaptation [28], or fine-tune [23, 24]. As decoders, the following are represented: transformed-based, CTC (Connectionist Temporal Classification), Recurrent Neural Network Transducer (RNN-T), Token-and-Duration Transducer (TDT) [21], and LLM (Large Language Model)-based. While LLM-based decoders are transformer-based, we denote a separate category for approaches that use a pre-trained LLM.

2.5. Adding a new model or dataset

All the evaluation code is open-sourced on GitHub.[4] As such, the process of adding a new model or dataset consists of opening a *pull request* (PR). To add a new model, a new folder should be created with: (1) a Python script for evaluating on a specific dataset and (optionally) with a specific model version;[5] and (2) a Bash script which calls the Python script for each dataset and model combination.

[4] github.com/huggingface/open_asr_leaderboard

[5] A template can be found in `transformers/run_eval.py`

3. RESULTS

All evaluation scripts, as described in Section 2.5, were conducted on an NVIDIA A100-SXM4-80GB GPU (driver 560.28.03, CUDA 12.6), using a batch size of 64 whenever memory allowed, and reduced adaptively (48, 32, 16, ...) when necessary to fit in device memory. Since the full results are continuously updated on the *Open ASR Leaderboard* and are too extensive to include here, we present a condensed version of the English leaderboard in Table 3 and the multilingual results in Table 4.

For English transcription (Table 3), models with a Conformer encoder and an LLM-based decoder achieve the best results, with the top four models using this architecture. However, these approaches are significantly slower compared to models using TDT or CTC decoders. While the latter can achieve superior RTFx, this comes at the cost of accuracy: *e.g.*, the best CTC-based model (*NVIDIA Parakeet CTC 1.1B*) ranks only 23rd in terms of WER.

Self-supervised learning (SSL) approaches have enabled ASR systems for 1K+ languages, yet the top SSL-based system for English transcription ranks only 52nd, *i.e.*, last row of Table 3 which uses *wav2vec2* [4]. As shown in Table 2, the leaderboard only covers SSL encoders with CTC decoders; combining such encoders with more performant decoders may help bridge this gap.

Another popular trend is to leverage Whisper's encoder (see Table 2), which has been trained on a large multilingual corpus. As shown in Table 3, models that fine-tune Whisper's encoder [23, 24] (or train a new decoder [27]) achieve better average WER than *OpenAI Whisper Large v3*.

The multilingual results of the *Open ASR Leaderboard* (Table 4) highlight an important tradeoff in ASR: specialization versus broad coverage. Namely, the improvement in English performance of fine-tuned Whisper models often comes at the cost of multilingual coverage: Whisper-derived models typically train on fewer languages (often just English), while Whisper itself supports 99 languages. Similarly, NVIDIA's models demonstrate this tradeoff more clearly: *Parakeet TDT 0.6B* v3 adds multilingual support compared to v2, and *Canary 1B* v2 expands from 4 to 25 languages. In both cases, broader language coverage comes at the cost of English transcription accuracy.

For Tables 3 and 4 (English and multilingual), open-source models show the strongest performance. The highest-ranking closed-source model (*Aqua Voice Avalon*) ranks 6th. Fairly computing RTFx for closed-source models is not possible due to upload latency and lack of GPU usage control, making direct efficiency comparisons infeasible.

4. CONCLUSION

We present the *Open ASR Leaderboard*, a reproducible benchmark covering 60+ systems and 10 datasets, including multilingual speech. Standardized text normalization enables a unified basis for comparing WER performance accuracy, and our RTFx evaluation allows for efficiency comparisons. Conformer–LLM models achieve the strongest English WER but at the cost of higher latency, whereas CTC/TDT decoders offer faster inference with only modest accuracy trade-offs, making them attractive for long-form transcription. All code and datasets are fully open-sourced to support transparent and extensible evaluation.

Future work include expanding language and domain evaluations (*e.g.*, far-field speech), incorporating additional metrics (*e.g.*, token error rate [8]), and exploring underrepresented encoder–decoder combinations (see Table 2). With the rise of LLMs and their ability for strong ASR performance (Table 3), we expect more approaches that employ them.

ACKNOWLEDGMENT

The authors would like to thank all the contributors to the *Open ASR Leaderboard*.

5. REFERENCES

[1] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 5206–5210.

[2] Guoguo Chen et al., "GigaSpeech: An Evolving, Multi-domain ASR Corpus with 10,000 Hours of Transcribed Audio," in *Proc. Interspeech*, 2021.

[3] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. International Conference on Machine Learning*, 2023.

[4] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," in *Advances in Neural Information Processing Systems*, 2020, vol. 33, pp. 12449–12460.

[5] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed, "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[6] Solène Evain et al., "Task Agnostic and Task Specific Self-Supervised Learning from Speech with LeBenchmark," in *Conference on Neural Information Processing Systems (Datasets and Benchmarks Track)*, 2021.

[7] Jon Barker, Ricard Marxer, Emmanuel Vincent, and Shinji Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding*, 2015, pp. 504–511.

[8] Jiayu Du, Jinpeng Li, Guoguo Chen, and Wei-Qiang Zhang, "SpeechColab leaderboard: An open-source platform for automatic speech recognition evaluation," *Computer Speech & Language*, vol. 94, 2025.

[9] Jean Carletta et al., "The AMI Meeting Corpus: A Pre-Announcement," in *Proc. International Conference on Machine Learning for Multimodal Interaction*, 2005, p. 28–39.

[10] Changhan Wang, Anne Wu, and Juan Pino, "CoVoST 2: A Massively Multilingual Speech-to-Text Translation Corpus," *arXiv:2007.10310*, 2020.

[11] Miguel Del Rio, Peter Ha, Quinten McNamara, Corey Miller, and Shipra Chandra, "Earnings-22: A Practical Benchmark for Accents in the Wild," *arXiv:2203.15591*, 2022.

[12] Alexis Conneau et al., "FLEURS: Few-Shot Learning Evaluation of Universal Representations of Speech," in *IEEE Spoken Language Technology Workshop*, 2023, pp. 798–805.

[13] Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert, "MLS: A Large-Scale Multilingual Dataset for Speech Research," *arXiv:2012.03411*, 2020.

[14] Patrick K. O'Neill et al., "SPGISpeech: 5,000 hours of transcribed financial audio for fully formatted end-to-end speech recognition," *arXiv:2104.02014*, 2021.

[15] François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Estève, "TED-LIUM 3: twice as much data and corpus repartition for experiments on speaker adaptation," in *International Conference on Speech and Computer*, 2018, pp. 198–208.

[16] Changhan Wang et al., "VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," in *Proc. Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, 2021, pp. 993–1003.

[17] Quentin Lhoest et al., "Datasets: A community library for natural language processing," in *Proc. Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2021, pp. 175–184.

[18] Nithin Rao Koluguri et al., "Investigating End-to-End ASR Architectures for Long Form Audio Transcription," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2024, pp. 13366–13370.

[19] Dima Rekesh et al., "Fast conformer with linearly scalable attention for efficient speech recognition," in *IEEE Automatic Speech Recognition and Understanding Workshop*, 2023, pp. 1–8.

[20] Anmol Gulati et al., "Conformer: Convolution-augmented transformer for speech recognition," *arXiv:2005.08100*, 2020.

[21] Hainan Xu, Fei Jia, Somshubra Majumdar, He Huang, Shinji Watanabe, and Boris Ginsburg, "Efficient sequence transduction by jointly predicting tokens and durations," in *Proc. International Conference on Machine Learning*, 2023.

[22] Neil Zeghidour et al., "Streaming sequence-to-sequence learning with delayed streams modeling," *arXiv:2509.08753*, 2025.

[23] Laurin Wagner, Bernhard Thallinger, and Mario Zusag, "CrisperWhisper: Accurate Timestamps on Verbatim Speech Transcriptions," *arXiv:2408.16589*, 2024.

[24] Alexander H. Liu et al., "Voxtral," *arXiv:2507.13264*, 2025.

[25] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli, "data2vec: A general framework for self-supervised learning in speech, vision and language," in *Proc. International Conference on Machine Learning*, 2022, vol. 162, pp. 1298–1312.

[26] Nat Jeffries, Evan King, Manjunath Kudlur, Guy Nicholson, James Wang, and Pete Warden, "Moonshine: Speech recognition for live transcription and voice commands," *arXiv:2410.15608*, 2024.

[27] Sanchit Gandhi, Patrick von Platen, and Alexander M. Rush, "Distil-Whisper: Robust Knowledge Distillation via Large-Scale Pseudo Labelling," *arXiv:2311.00430*, 2023.

[28] Keisuke Kamahori, Jungo Kasai, Noriyuki Kojima, and Baris Kasikci, "LiteASR: Efficient Automatic Speech Recognition with Low-Rank Approximation," *arXiv:2502.20583*, 2025.