ChatGPT

Open-Source ASR and Speaker-Diarization Models (fine-tunable)

Automatic Speech Recognition (ASR) and speaker-diarization models evolve quickly. The tables below summarise notable open-source models that support fine-tuning, sorted **from newest to oldest** based on their release year and then by model size (smallest at the top of each year). Only models with publicly available code/models (GitHub/Hugging Face) were included. Word Error Rate (WER) is used for ASR accuracy, Diarization Error Rate (DER) for diarization models and Real-Time Factor (RTF×) for throughput (higher RTF× means faster). "—" indicates that a metric was not reported.

ASR models (latest → oldest)

Year ↓ / Model	Language(s) / size	Fine-tuning & repository	Key performance (WER, RTF× & notes)
2026			
Canary Qwen 2.5B	English only; **2.5 B parameters**	Fine-tuned via NVIDIA NeMo with LoRA; checkpoints on HuggingFace (`nvidia/canary-qwen-2.5b`)	Tops the Open ASR leaderboard (average **5.63 % WER**) and achieves **418×** real-time throughput. Trained on 234 k h of English and uses a hybrid Fast-Conformer + Qwen LLM decoder [1]. SALM architecture allows switching between transcription and language-model modes.

Year ↓ / Model	Language(s) / size	Fine-tuning & repository	Key performance (WER, RTF× & notes)
IBM Granite Speech 3.3 (≈9 B)	English + French, German & Spanish; **≈9 B parameters**	Fine-tuned with LoRA; available via IBM Watson GitHub and HuggingFace (`ibm/granite-speech-3_3-8b`)	Average **5.85 % WER** on Open ASR leaderboard; micro-batch throughput unspecified. Multilingual training with modality alignment and LoRA fine-tuning provides robust noise resilience [2].
Moonshine (Useful Sensors)	Multilingual; **Tiny 27 M / Base 62 M parameters**	Fully open-source with Transformers and ONNX support (`UsefulSensors/moonshine-*` on HuggingFace). Designed for edge devices; fine-tuning via `transformers`/ ONNX.	Moonshine Tiny (27 M) and Base (62 M) process audio **5–15× faster** than Whisper while matching or improving WER on several languages. For English the Tiny model reaches **12.66 % WER** and the Base **10.07 % WER**, outperforming Whisper Tiny and Base variants of similar size [3]. The models support Arabic, Chinese, Japanese, Korean, Spanish, Ukrainian and Vietnamese with comparable CER/WER [4].

Year ↓ / Model	Language(s) / size	Fine-tuning & repository	Key performance (WER, RTF× & notes)
Distil-Whisper Large V3	English (currently); **756 M parameters**	HuggingFace `distil-whisper` models; fine-tuning via LoRA or full fine-tuning.	Compact distilled version of Whisper Large V3: within **≈1 % WER** of the full model while providing **≈6×** faster inference; uses two decoder layers and knowledge distillation [6]. Noise robustness improved (2.1 % lower insertion errors) [6].
Whisper Large V3 Turbo	99 + languages; **809 M parameters**	Fine-tuned variant of Whisper V3 (OpenAI); available via `openai/whisper-large-v3-turbo` on HuggingFace; can be fine-tuned using PEFT/LoRA.	Reduces decoder layers from 32 to 4, yielding **6× speedup** and **216× RTF** while keeping WER within **1–2 %** of Whisper Large V2 [7]. Suitable for fast multilingual transcription.
Whisper Large V3	99 + languages; **1.55 B parameters**	OpenAI's MIT-licensed model on GitHub (`openai/whisper`) and HuggingFace. Fine-tuning via LoRA or full fine-tuning using the HuggingFace transformer library.	Average **7.4 % WER** on mixed benchmarks; strong multilingual performance [8]. Handles punctuation, capitalization and 128-bin mel-spectrograms. On modern GPUs it processes audio faster than real time [8].

Year ↓ / Model	Language(s) / size	Fine-tuning & repository	Key performance (WER, RTF× & notes)
Parakeet TDT (1.1 B)	English; **1.1 B parameters**	NVIDIA's streaming RNN-Transducer; open weights on HuggingFace (`nvidia/parakeet-tdt-*`). Supports LoRA fine-tuning via NeMo.	Designed for ultra-fast streaming: RTF× > **2 000** (up to 2 000× real-time) while maintaining around **8 % WER** on Open ASR leaderboard [9]. Prioritizes throughput over contextual accuracy.
FireRedASR	Mandarin, Chinese dialects & English; **LLM 8.3 B / AED 1.1 B**	GitHub: `FireRedTeam/FireRedASR` with scripts for fine-tuning. Requires Qwen2-7B for the LLM variant.	On public Mandarin ASR benchmarks, FireRedASR-LLM achieves average **3.05 % character-error rate** and the AED variant 3.18 % [10]. On English LibriSpeech the LLM model obtains **1.73 % WER (test-clean)** and **3.67 % WER (test-other)** [11].

Year ↓ / Model	Language(s) / size	Fine-tuning & repository	Key performance (WER, RTF× & notes)
Reverb ASR (Rev)	English; **≈280 M parameters**	Open-sourced by Rev (GitHub: `revdotcom/reverb`), built on WeNet; supports fine-tuning. Turbo variant provides int8 quantization.	On long-form earnings-call benchmarks, Reverb Verbatim records **7.64 % WER (Earnings21)** and **11.38 % WER (Earnings22)**, outperforming Whisper Large V3 (13.67 % and 18.53 %) and Canary-1B [13]. Turbo variant slightly increases WER but runs faster. On Rev16 dataset the model achieves **7.99 % WER (verbatim) / 7.06 % (non-verbatim)** [14]. On GigaSpeech it records **11.05 % WER** (compared to Whisper's 10.02 %) [15].
Distil-Whisper Large V2 / Whisper Large V2	2024 (distilled) / 2023; **0.75–1.6 B parameters**	LoRA fine-tuning via HuggingFace; widely used baseline; large-V2 preceded V3.	WER ≈ 10 % on Open ASR; widely used as baseline (see Whisper Large V3 for improvements). Distilled variant is ~2× faster while maintaining similar WER [17].
2023			
WavLM	Multilingual; pre-trained representation model (**Base ≈94 M params**, Large ≈316 M)	Released by Microsoft; fine-tuning via `torch.hub` or HuggingFace for ASR, speaker verification and diarization.	Self-supervised model pre-trained on 94 k h of speech. Evaluations on SUPERB show WavLM Large improves ASR performance by **2.4 WER points** over HuBERT and reduces diarization error rate by **12.6 %** [18] [19].

Year ↓ / Model	Language(s) / size	Fine-tuning & repository	Key performance (WER, RTF× & notes)
Whisper (Tiny/Base/Small/Medium)	2019-2022; **39 M–769 M parameters**	OpenAI's Whisper models on GitHub. Fine-tuning via LoRA/PEFT for domain-specific tasks.	Provide strong multilingual ASR with WER ≈ 10–20 % depending on size; slower throughput than newer models. Widely used for baseline and accessible.
2020			
Wav2Vec 2.0	English (XLSR variant 53 languages); **95 M params**	Meta's self-supervised model (HuggingFace `facebook/wav2vec2-*`); fine-tune with small labeled data.	Fine-tuning on the full LibriSpeech dataset yields **1.8 % WER (clean)** and **3.3 % WER (test-other)**; with only **10 min** of labeled data it achieves **4.8 %/8.2 % WER** [20]. Often used as a feature extractor for downstream ASR tasks.
2019			
DeepSpeech (Mozilla)	English; **RNN with ~50 M parameters**	Archived project on GitHub (`mozilla/DeepSpeech`); fine-tuning on custom data via TensorFlow training scripts.	Achieved **7.06 % WER** on the LibriSpeech clean corpus and processed audio **≈30 % faster** than Whisper Large [22]. Focused on low-resource hardware and offline use; now deprecated but still used for edge devices.

Speaker-diarization models (latest → oldest)

Year ↓ / Model	Fine-tuning & repository	DER / WDER & notes	Inference speed / latency	C
2025 – 2026				

Year ↓ / Model	Fine-tuning & repository	DER / WDER & notes	Inference speed / latency	C
pyannote Precision-2	`pyannote/speaker-diarization-precision-2` on HuggingFace (requires pyannote AI token). Fine-tuning via pyannote audio.	Provides the lowest error rates among pyannote models: on AISHELL-4 it achieves **11.4 % DER**, AliMeeting 15.2 %, AMI-IHM 12.9 %, AMI-SDM 15.6 %, CALLHOME-part2 16.6 % and DIHARD 3 full 14.7 % [23]. Improvements come from refined speaker assignment and counting.	Offline pipeline (GPU or CPU). Latency depends on chunk size; not reported but supports batch processing.	p c r
pyannote Community-1	`pyannote/speaker-diarization-community-1` (open after accepting usage conditions).	Aimed at community research; DER values slightly higher than Precision-2 but lower than legacy 3.1: e.g., **AISHELL-4: 11.7 %; AliMeeting: 20.3 %; AMI-IHM: 17.0 %; AMI-SDM: 19.9 %; CALLHOME-part2: 26.7 %; DIHARD 3: 20.2 %; VoxConverse: 11.2 %** [23].	CPU-friendly pipeline; inference time depends on hardware.	p c r

Year ↓ / Model	Fine-tuning & repository	DER / WDER & notes	Inference speed / latency	C
DiariZen Large-s80-v2 (2024)	GitHub `roy13k/DiariZen` (open-source). Models can be fine-tuned using PyTorch; includes streaming support.	Across benchmark datasets DiariZen models significantly outperform pyannote v3.1: on AMI-SDM the Large-s80-v2 variant reduces DER from **22.4 % (pyannote v3.1)** to **13.9 %**, on AISHELL-4 from 12.2 % to **10.1 %**, and on VoxConverse from 11.3 % to **9.1 %** [25].	Real-time factor (RTF) not explicitly given; the DiariZen paper highlights streaming capability using efficient attention.	D C
Sortformer v2 / streaming Sortformer (NVIDIA)	HuggingFace `nvidia/diar_streaming_sortformer_4spk-v2` and `nvidia/diar_sortformer` (requires NeMo). Supports full fine-tuning via NeMo.	In the 2025 benchmark by ETH Zürich, Sortformer v2 achieved DER **7.0 %** on the AliMeeting dataset and outperformed other models in that scenario [26]. Self-reported DER values on HuggingFace show for streaming v2 with optimized post-processing: **DIHARD III ≤4 spk: 13.24 %**, **CALLHOME-2spk: 6.05 %**, **CALLHOME-3spk: 9.88 %** and **CALLHOME-4spk: 11.72 %** [27].	ETH benchmark reports Sortformer v2 as the **fastest diarization model** with average RTF≈**214 ×** (audio processed 214× faster than real-time) [26]. Streaming version allows latencies down to 0.32 s with slight DER increases [27].	E b p H r

Year ↓ / Model	Fine-tuning & repository	DER / WDER & notes	Inference speed / latency	C
Reverb Diarization v2	GitHub `revdotcom/reverb` (pyannote-based).	Evaluated using Word-level DER (WDER): on Earnings21 test set WDER **0.046**, on Rev16 **0.078**, outperforming pyannote 3.0 (0.051 and 0.090) [28]. Fine-tuned on 26 k h of labeled data [28].	Integrated into Reverb ASR pipeline; inference speed unspecified.	R
pyannote v3.1 (legacy 2023)	Early open-source diarization pipeline widely used in research (`pyannote/speaker-diarization-3.1`).	Serves as baseline in many studies. For example, on AMI-SDM dataset it yields **22.4 % DER**, on AISHELL-4 **12.2 %**, on VoxConverse **11.3 %**, on CALLHOME-part2 **28.5 %**, and on DIHARD 3 full **21.4 %** [23].	Inference typically runs slower (RTF≈2.5 on GPU) [29].	p l c
Multi-Scale Diarization Decoder (MSDD)	Part of NVIDIA NeMo toolkit (`nvidia/ne-mo`); uses TitaNet embeddings and multi-scale segmentation; fine-tuning supported.	Implements neural refinement over initial clustering for improved diarization. Detailed DER benchmarks are not published; the CHiME-7 challenge team reported DER <15 % in simulations.	Inference latency depends on segment lengths; designed for streaming with latencies down to ~0.5 s.	N g (c

Year ↓ / Model	Fine-tuning & repository	DER / WDER & notes	Inference speed / latency	C
WhisperX (ASR + Diarization)	GitHub `m-bain/whisperX`. Combines OpenAI's Whisper with pyannote diarization; can be fine-tuned with LoRA.	Provides **70× real-time** batched transcription using `faster-whisper` backend and attaches pyannote diarization to produce speaker-attributed transcripts [31]. Accuracy depends on the underlying Whisper model; diarization quality similar to pyannote v3.x.	Highly optimized: 70× faster than Whisper Large V2 and runs with <8 GB GPU memory [31].	W R
2019 – 2022				
WavLM / Wav2Vec 2.0 diarization	Pre-trained models such as WavLM and Wav2Vec 2.0 can be fine-tuned for diarization tasks (e.g., by training an LSTM classifier on speaker embeddings). The WavLM team reports a **12.6 %** relative reduction in DER compared to prior systems [19].	Fine-tuning requires custom scripts; speeds vary.	WavLM article [19]	

Guidance for selecting models

1. **Choose models based on accuracy vs speed trade-off.** Canary Qwen 2.5B and Granite Speech 3.3 achieve the lowest WER among open-source ASR models but require high-end GPUs. Distil-Whisper and Whisper Turbo sacrifice <2 % accuracy for dramatic speed improvements (RTF ≈ 216×). Parakeet TDT and Moonshine are preferred for real-time or edge deployment.
2. **Model size matters for deployment.** Moonshine (27–62 M parameters) and DeepSpeech (~50 M) run on edge devices, while modern SALM and Conformer-LLM hybrids (Granite, Canary) require billions of parameters.
3. **Fine-tuning is essential for domain-specific tasks.** Most ASR models listed (Whisper, Distil-Whisper, Parakeet, Reverb ASR, Wav2Vec2, WavLM) support fine-tuning via HuggingFace, NeMo or PyTorch frameworks. For diarization, pyannote audio and NeMo provide training scripts to adapt models to new languages or environments.
4. **Latest diarization models improve both accuracy and speed.** Pyannote Precision-2 and DiariZen significantly reduce DER compared with legacy pyannote 3.1. Sortformer v2 sets a new bar for inference speed (RTF ≈ 214×) but still trails Precision-2 in accuracy on some datasets [26]. Reverb's diarization models focus on WDER (word-level diarization) to better integrate with ASR [28].

5. **Consider license and deployment constraints.** Most modern models use permissive licenses (CC-BY-4.0, Apache 2.0 or MIT), making them suitable for commercial use. However, some (e.g., pyannote community models) require registration and impose API usage conditions.

The tables above summarize the currently available fine-tunable ASR and diarization models as of February 2026. When selecting a model, consider domain specificity, available computing resources, speed requirements and whether speaker diarization is needed alongside transcription.

---

[1] [2] [6] [7] [8] [9] [20] [21] Best open source speech-to-text (STT) model in 2026 (with benchmarks) | Blog — Northflank
https://northflank.com/blog/best-open-source-speech-to-text-stt-model-in-2026-benchmarks

[3] [4] [5] GitHub - moonshine-ai/moonshine: Fast and accurate automatic speech recognition (ASR) for edge devices
https://github.com/moonshine-ai/moonshine

[10] [11] [12] GitHub - FireRedTeam/FireRedASR: Open-source industrial-grade ASR models supporting Mandarin, Chinese dialects and English, achieving a new SOTA on public Mandarin ASR benchmarks, while also offering outstanding singing lyrics recognition capability.
https://github.com/FireRedTeam/FireRedASR

[13] [14] [15] [16] [28] Reverb: Open-Source ASR and Diarization from Rev
https://arxiv.org/html/2410.03930v3

[17] The Top Open Source Speech-to-Text (STT) Models in 2025
https://modal.com/blog/open-source-stt

[18] [19] WavLM: Universal Self-Supervised Speech Model
https://www.emergentmind.com/topics/wavlm-model

[22] DeepSpeech Statistics 2026
https://www.aboutchromebooks.com/deepspeech-statistics/

[23] [24] pyannote/speaker-diarization-community-1 · Hugging Face
https://huggingface.co/pyannote/speaker-diarization-community-1

[25] GitHub - BUTSpeechFIT/DiariZen: A toolkit for speaker diarization.
https://github.com/BUTSpeechFIT/DiariZen

[26] Benchmarking Diarization Models
https://arxiv.org/pdf/2509.26177

[27] nvidia/diar_streaming_sortformer_4spk-v2 · Hugging Face
https://huggingface.co/nvidia/diar_streaming_sortformer_4spk-v2

[29] Top 8 speaker diarization libraries and APIs in 2025
https://www.assemblyai.com/blog/top-speaker-diarization-libraries-and-apis

[30] Models — NVIDIA NeMo Framework User Guide
https://docs.nvidia.com/nemo-framework/user-guide/24.09/nemotoolkit/asr/speaker_diarization/models.html

[31] raw.githubusercontent.com
https://raw.githubusercontent.com/m-bain/whisperX/main/README.md