Gladia

Speech Recognition Benchmark Report 2026

Open Methodology & Reproducible Evaluation

Table of contents

Executive summary

This report presents a benchmark of modern speech recognition systems evaluated across real-world conditions, including conversational, multilingual, noisy, and long-form audio. All models are tested under consistent conditions using a shared evaluation pipeline.

The results show that performance varies significantly depending on the domain. Systems that achieve low error rates on clean speech often degrade in conversational and noisy settings, where overlapping speech, disfluencies, and background noise introduce additional complexity.

Across datasets, two patterns stand out. First, conversational speech remains the most challenging setting, with substantially higher error rates driven by interruptions, short utterances, and speaker variability. Second, performance on multilingual and noisy audio is uneven, with some models showing strong robustness while others degrading sharply outside controlled conditions.

To ensure fair comparison, all transcripts are normalized before scoring to remove formatting differences that do not reflect actual recognition quality. This makes Word Error Rate (WER) more representative of true model performance.

The benchmark is designed to be reproducible. This allows results to be verified, extended, or re-run on custom audio. All datasets, evaluation steps, and scoring logic are defined explicitly. The full methodology is publicly available.

1. Introduction

Speech recognition systems have improved rapidly, with modern APIs achieving low error rates on standard benchmarks. However, these benchmarks rely on controlled datasets that fail to capture real-world conditions, such as noisy audio, multiple speakers, and multilingual conversations. This report evaluates leading speech recognition APIs on datasets designed to reflect real-world usage.

Motivation

Beyond dataset limitations, benchmark results are often difficult to interpret and compare. Providers use different evaluation pipelines, normalization rules, and reporting practices, making reported metrics inconsistent and sometimes misleading.

Without a reproducible and standardized methodology, reliable comparison across systems is not possible. The report addresses that gap by introducing an open evaluation framework that can be replicated and extended.

Goals of the benchmark

The goal of this benchmark is to provide a transparent and controlled comparison of production speech recognition APIs. All systems are evaluated under identical conditions, using a shared configuration and scoring pipeline.

Contributions

This report provides:

- A standardized evaluation pipeline, including dataset handling, normalization, and scoring
- A multi-domain evaluation setup covering varied audio conditions
- Public access to the methodology and tooling, enabling reproduction and extension

Together, these elements enable consistent, verifiable comparison of speech recognition systems in practice.

2. Benchmark scope

Systems evaluated

For each provider, the latest available model was evaluated. When a recent model did not support all languages, the previous model was also evaluated. The language is explicitly passed; code switching is disabled; and no additional provider-specific parameters are set unless noted below.

Provider	Model	API Mode	Specific Configuration
Gladia	solaria-1	Async	Code switching disabled
AssemblyAI	Universal 3 Pro	Async	—
AssemblyAI	Universal 2	Async	—
ElevenLabs	Scribe V2	Async	—
Deepgram	Nova-3	Async	—
Speechmatics	Enhanced	Async	—
Soniox	V4 Async	Async	Forced to single language due to language-detection issues
Mistral	Voxtral Mini Transcribe 2	Async	timestamp_granularities set to segment due to hallucination issues

Evaluation domains

Speech environments covered by the benchmark.

- Conversational telephone dialogue (Switchboard)
- Multilingual reading speech (Common Voice 24, Multilingual LibriSpeech)
- Financial and earnings call audio (Earnings22)
- Clean parliamentary speech recordings (VoxPopuli Cleaned)
- Real-time streaming transcription (Pipecat STT Benchmark)

Languages evaluated

Languages included in the benchmark.

- English
- German
- Spanish
- French
- Italian
- Portuguese

Datasets: Overview

Dataset	Domain	Languages	Source
Common Voice 24	Crowd-sourced multilingual speech	en, fr, it, pt, nl, es, de	Mozilla Common Voice
VoxPopuli Cleaned AA	Parliamentary speech (cleaned)	en	HuggingFace
Earnings22 Full	Corporate earnings calls (full-length)	en	HuggingFace
Earnings22 Cleaned AA	Corporate earnings calls (cleaned)	en	HuggingFace
Multilingual LibriSpeech	Read audiobooks	de, es, fr, it, pt	HuggingFace
Switchboard	Conversational telephone speech	en	HuggingFace

Dataset	Domain	Languages	Source
Pipecat STT Benchmark	Streaming speech evaluation	en	HuggingFace

Datasets: Descriptions

Common Voice 24

Common Voice is a large-scale, crowd-sourced speech dataset started by Mozilla. Volunteers record short sentences, producing diverse speaker demographics and recording conditions. The audio spans 7 hours in total, an hour per language. The benchmark uses the test split across seven languages (en, fr, it, pt, nl, es, de). Audio is 32 KHz mono MP3.

VoxPopuli (cleaned AA)

VoxPopuli is derived from European Parliament event recordings, spanning 1 hour and 58 mins and covering a variety of accents and speaking styles. The "Cleaned AA" variant was curated by Artificial Analysis to remove noisy or misaligned references. On average, model WER on VoxPopuli went down 3.5 percentage points after cleaning. English only.

Earnings22

Earnings22 consists of ~53 hours of corporate earnings calls and financial presentations. Two variants are evaluated: **Full** (entire meeting recordings, testing long-form transcription) and **Cleaned AA** (curated by Artificial Analysis; on average, model WER dropped by 5.6 percentage points after cleaning). Audio is 24 KHz stereo MP3. English only.

Multilingual LibriSpeech

Multilingual LibriSpeech (MLS) is a large multilingual corpus derived from read audiobooks from LibriVox. It covers 5 hours and 40 mins of audio in eight languages; this benchmark evaluates five non-English languages: German, Spanish, French, Italian, and Portuguese. Note: Polish was benchmarked separately because AssemblyAI Universal 3 Pro does not support it.

Switchboard

Switchboard is a collection of 3 hours and 53 mins of approximately 2,400 two-sided telephone conversations among 543 speakers (302 male, 241 female) from all areas of the United States. Audio is 16 KHz mono WAV. This dataset is challenging due to conversational disfluencies, overlapping speech, and telephony-bandwidth audio.

Pipecat STT benchmark

The Pipecat STT Benchmark uses 2 hours and 39 mins of audio published alongside the Pipecat real-time streaming benchmark. It tests short-utterance, low-latency transcription scenarios typical of voice AI applications. English only.

3. Audio sampling and excluded datasets

Audio selection criteria

Datasets are downloaded from Hugging Face using only the test split (never train or dev). Parquet files are processed to be made compatible with the benchmarking tool, in a flat files mode.

Audio is not re-encoded or converted. The original format is preserved. If the audio has multiple channels, it is flattened to mono with ffmpeg.

Datasets are sampled to control cost while maintaining statistical validity. For example, Multilingual LibriSpeech is 31 hours long; with 7 providers at approximately $0.50 per hour, a single benchmark run costs about $108. If one provider crashes or does not support a format or language, the benchmark must be restarted.

For a given dataset, the average WER converges to a stable value: adding more samples eventually no longer changes the average meaningfully. Convergence analysis was performed to determine adequate sample sizes for each dataset.

Out-of-scope evaluations

To avoid misinterpretation, several scenarios were intentionally excluded.

- **Languages not included in the benchmark.** Languages without specific normalization rules are not benchmarked.
- **Domain-specific fine-tuned models.** All models are the generic version — for example, it is always nova-3 (or nova-3-general) and never , never nova-3-medical.
- **Post-processing pipelines.** No parameter, such as custom prompt or vocabulary, has been set to correct the transcription.
- **Human correction.** No human has modified the results.
- **Streaming / real-time APIs.** The benchmark focused on asynchronous APIs only.
- **Old models.** We focused on the most recent models, with an exception when the language coverage of a new model is limited (for example, AssemblyAI Universal 3 Pro does not cover all languages, so Universal 2 was also evaluated).

4. Evaluation methodology

Benchmark design principles

- Fair comparison across providers
- Minimal provider-specific configuration
- Reproducibility and transparency

Input processing

Audio preprocessing

- **Sample rate**: Original sample rate preserved (no resampling)
- **Channel normalization**: Multi-channel audio flattened to mono via ffmpeg
- **Format**: Original codec preserved (WAV, MP3)

API invocation

- All benchmarks use **async (batch) mode**, each provider's asynchronous transcription API
- Language is passed explicitly for each request; code switching is disabled
- For Soniox V4, language configuration is forced with “Language hints” and “Strict language hints” options to avoid language-detection issues
- For Mistral Voxtral Mini Transcribe 2, timestamp_granularities is set to segment instead of word due to hallucination issues

Transcript normalization

Word Error Rate (WER) operates on raw strings and has no notion of meaning. Two transcriptions that say the same thing in different surface forms get penalized as errors:

Ground truth	STT output	Match without normalization
It's $50	it is fifty dollars	0/3 words match
3:00 PM	3 pm	0/2 words match
Mr. Smith	mister smith	0/2 words match

These are formatting differences, not transcription errors. Without normalization, WER scores become unreliable, and cross-engine comparisons are meaningless.

Both the ground truth and the STT output are reduced to a shared canonical form *before* WER is computed, so that only genuine recognition errors affect the score. Normalization is performed by gladia-normalization, an open-source, deterministic, language-aware pipeline published as a Python package (pip install gladia-normalization).

Pipeline stages

Every pipeline runs exactly three stages, always in this order:

1. **Stage 1 — Text pre-processing**: Full-text transforms such as protecting symbols, expanding contractions, converting numbers, casefolding, and removing symbols.
2. **Stage 2 — Word processing**: Per-token transforms such as word replacements and filler removal.
3. **Stage 3 — Text post-processing**: Full-text cleanup such as restoring placeholders, collapsing digits, formatting time patterns, and normalizing whitespace.

This ordering is a hard constraint — some steps depend on earlier steps having run (e.g., a placeholder protecting a decimal point in Stage 1 must be restored in Stage 3, so that symbol removal doesn't destroy it in between).

Example

None

```
Input:  "It's $50.9 at 3:00PM — y'know, roughly."
Output: "it is 50 point 9 dollars at 3 pm you know roughly"
```

Preset configuration

Pipelines are defined declaratively in YAML presets. Each preset lists the steps that run in each stage and their execution order. The benchmark uses the gladia-3 preset:

Python

```python
from normalization import load_pipeline

pipeline = load_pipeline("gladia-3", language="en")
pipeline.normalize("It's $50 at 3:00PM")
# => "it is 50 dollars at 3 pm"
```

Supported languages

Code	Language
en	English
fr	French (alpha)

Unsupported language codes fall back to a safe default that applies language-independent normalization only.

5. Evaluation metrics

Word Error Rate (WER)

Word Error Rate (WER) is the standard metric used to evaluate the quality of speech recognition systems. It measures the accuracy of the transcription by comparing the recognized words with a reference transcription.

WER is calculated as:

```
None
WER = (S + D + I) / N
```

Where:

- **S**: Number of substitutions (words replaced with incorrect words)
- **D**: Number of deletions (words missed by the system)
- **I**: Number of insertions (words added by the system that weren't spoken)
- **N**: Total number of words in the reference transcription

The lower the WER, the better the quality of the transcription. A perfect system would have a WER of 0%.

In real life, what that translates to is:

- A 5% WER is excellent / near human parity: 1 error in a 20-word sentence
- A 10% WER is good / highly usable: 1 error in a 10-word sentence
- A 20% WER is fair / needs manual review: 2 errors in a 10-word sentence

Diarization Error Rate (DER)

DER is the standard metric for speaker diarization evaluation. It measures the fraction of total reference speech time that is incorrectly diarized, combining three error types:

- **Missed speech (T_miss)**: Reference speech that the system failed to detect — a speaker is talking but the system outputs no speech activity for that interval, or two speakers overlap and only one is detected.
- **False alarm (T_fa)**: The system detected speech where none exists in the reference — typically background noise or non-speech sounds misclassified as speech.
- **Speaker confusion (T_conf)**: Speech was correctly detected but attributed to the wrong speaker. This is often the most damaging error for user experience, producing misleading transcripts rather than merely incomplete ones.

None

```
DER = (T_miss + T_fa + T_conf) / T_total
```

A DER of 10% means one-tenth of the total speech time contains errors. State-of-the-art systems achieve 5–8% on standard benchmarks and 15–25% on challenging real-world data. Lower is better.

Practical impact: for a 60-minute meeting, going from 10% to 5% DER eliminates roughly 3 minutes of incorrect speaker labels. User satisfaction drops sharply once DER exceeds 12–15%, as misattributions become frequent enough to undermine trust.

Real-Time Factor (RTFx)

RTFx measures processing speed as the ratio of audio duration to wall-clock processing time.

Therefore, an RTFx of 1 means a system processes speech in real-time (1 second of audio in 1 second of wall-clock time), while an RTFx of 2 means it processes the audio twice as fast (1 second of audio in 0.5 seconds of wall-clock time). Thus, a higher RTFx value indicates faster processing and lower effective latency.

6. Statistical analysis

Variance across samples

For a given dataset, the average WER converges to a stable value as more samples are added. Convergence was validated empirically to ensure that reported averages are statistically meaningful.

Outlier analysis

High-WER cases (WER > 100%) were investigated for each dataset. Common causes include:

- **Soniox V4** transcribing English audio in non-Latin scripts (Telugu, Hindi, Devanagari) due to language-detection failures
- **Reader instructions** in Multilingual LibriSpeech (e.g., "chapitre 8 de abc petits contes de jules maitre enregistre pour librivox point org") are being inconsistently included in both ground truth and hypothesis
- **Very short reference segments** (1–2 words) where any insertion produces extreme WER
- **Conversational disfluencies** in Switchboard, where overlapping speech from adjacent turns bleeds into the transcription

7. Results

Overall WER summary

Service	Common Voice 24	VoxPopuli Cleaned	Earnings 22 Full	Earnings 22 Cleaned	Multilingual LibriSpeech	Pipecat STT	Switchboard
gladia-solaria-1	6.70%	2.20%	11.80%	7.90%	5.80%	2.70%	35.80%
assemblyai-universal-3-pro	3.90%	2.10%	11.00%	7.00%	4.70%	2.00%	56.00%
mistralai-voxtral-mini-transcribe-2	5.10%	2.10%	11.60%	7.50%	—	2.60%	50.10%
elevenlabs-scribe_v2	3.90%	1.70%	9.40%	7.90%	3.70%	2.20%	62.50%

Service	Common Voice 24	VoxPopuli Cleaned	Earnings 22 Full	Earnings 22 Cleaned	Multilingual LibriSpeech	Pipecat STT	Switchboard
speechmatics	3.80%	3.00%	10.00%	7.70%	—	2.70%	56.00%
assemblyai-universal-2	5.20%	2.20%	11.10%	6.90%	6.20%	2.50%	63.10%
soniox-v4	7.20%	—	—	5.70%	5.60%	2.90%	62.90%
deepgram-nova-3	7.90%	3.20%	14.50%	12.70%	7.50%	3.10%	65.20%

"—" indicates the provider did not return results (timeout or unsupported).

Overall ranking by dataset

Dataset	1st	2nd	3rd
Common Voice 24	Speechmatics (3.80%)	AssemblyAI U3 Pro / ElevenLabs Scribe V2 (3.90%)	Mistral Voxtral (5.10%)
VoxPopuli Cleaned	ElevenLabs Scribe V2 (1.70%)	Mistral Voxtral / AssemblyAI U3 Pro (2.10%)	Gladia solaria-1 / AssemblyAI U2 (2.20%)
Earnings22 Full	ElevenLabs Scribe V2 (9.40%)	Speechmatics (10.00%)	AssemblyAI U3 Pro (11.00%)
Earnings22 Cleaned	Soniox V4 (5.70%)	AssemblyAI U2 (6.90%)	AssemblyAI U3 Pro (7.00%)
Multilingual LibriSpeech	ElevenLabs Scribe V2 (3.70%)	AssemblyAI U3 Pro (4.70%)	Soniox V4 (5.60%)
Pipecat STT	AssemblyAI U3 Pro (2.00%)	ElevenLabs Scribe V2 (2.20%)	AssemblyAI U2 (2.50%)

Dataset	1st	2nd	3rd
Switchboard	**Gladia solaria-1 (35.80%)**	Mistral Voxtral (50.10%)	Speechmatics / AssemblyAI U3 Pro (56.00%)

Dataset-level results

Common Voice 24

Service	Transcript	Average WER	RTFx	Perfect	High WER
speechmatics	100%	3.80%	2	1067	2
assemblyai-universal-3-pro	100%	3.90%	0	1041	1
elevenlabs-scribe_v2	100%	3.90%	6	1059	3
mistralai-voxtral-mini-transcribe-2	100%	5.10%	6	984	2
assemblyai-universal-2	100%	5.20%	1	968	2
gladia-solaria-1	100%	6.70%	1	904	2
soniox-v4	100%	7.20%	1	902	7
deepgram-nova-3	100%	7.90%	2	808	0

VoxPopuli cleaned

Service	Transcript	Average WER	RTFx	Perfect	High WER
elevenlabs-scribe_v2	100%	1.70%	5	418	0

Service	Transcript	Average WER	RTFx	Perfect	High WER
mistralai-voxtral-mini-transcribe-2	100%	2.10%	5	394	0
assemblyai-universal-3-pro	100%	2.10%	2	396	0
gladia-solaria-1	100%	2.20%	3	393	0
assemblyai-universal-2	100%	2.20%	3	377	0
speechmatics	100%	3.00%	6	326	0
deepgram-nova-3	100%	3.20%	7	363	0

Note: Soniox V4 did not return results for this dataset.

Earnings22 (full)

Service	Transcript	Average WER	RTFx	Perfect	High WER
elevenlabs-scribe_v2	100%	9.40%	35	0	0
speechmatics	100%	10.00%	17	0	0
assemblyai-universal-3-pro	100%	11.00%	71	0	0
assemblyai-universal-2	100%	11.10%	82	0	0
mistralai-voxtral-mini-transcribe-2	100%	11.60%	135	0	0
gladia-solaria-1	100%	11.80%	28	0	0

Service	Transcript	Average WER	RTFx	Perfect	High WER
deepgram-nova-3	100%	14.50%	348	0	0

Note: Soniox V4 timed out on this long-form dataset and did not return results.

Earnings22 (cleaned)

Service	Transcript	Average WER	RTFx	Perfect	High WER
soniox-v4	100%	5.70%	35	0	0
assemblyai-universal-2	100%	6.90%	59	0	0
assemblyai-universal-3-pro	100%	7.00%	64	0	0
mistralai-voxtral-mini-transcribe-2	100%	7.50%	57	0	0
speechmatics	100%	7.70%	24	0	0
elevenlabs-scribe_v2	100%	7.90%	32	0	0
gladia-solaria-1	100%	7.90%	39	0	0
deepgram-nova-3	100%	12.70%	234	0	0

Multilingual LibriSpeech

Service	Transcript	Average WER	RTFx	Perfect	High WER
elevenlabs-scribe_v2	100%	3.70%	15	565	3

Service	Transcript	Average WER	RTFx	Perfect	High WER
assemblyai-universal-3-pro	100%	4.70%	1	508	3
soniox-v4	100%	5.60%	6	378	3
gladia-solaria-1	100%	5.80%	3	367	3
assemblyai-universal-2	100%	6.20%	3	369	2
deepgram-nova-3	100%	7.50%	7	270	4

Note: Speechmatics and Mistral Voxtral did not return results for this multilingual dataset.

WER by language (multilingual LibriSpeech):

Service	DE	ES	FR	IT	PT
elevenlabs-scribe_v2	3.10%	3.20%	2.90%	6.10%	3.00%
assemblyai-universal-3-pro	3.50%	3.20%	2.60%	9.70%	4.40%
soniox-v4	5.40%	4.40%	5.00%	8.80%	4.30%
gladia-solaria-1	5.00%	4.00%	4.80%	9.90%	5.30%
assemblyai-universal-2	3.40%	4.00%	5.80%	11.90%	5.90%
deepgram-nova-3	6.90%	4.60%	6.20%	8.80%	11.30%

Pipecat streaming benchmark

Service	Transcript	Average WER	RTFx	Perfect	High WER
assemblyai-universal-3-pro	100%	2.00%	2	531	0
elevenlabs-scribe_v2	100%	2.20%	8	512	0
assemblyai-universal-2	100%	2.50%	1	494	0
mistralai-voxtral-mini-transcribe-2	100%	2.60%	5	485	0
speechmatics	100%	2.70%	0	476	0
gladia-solaria-1	100%	2.70%	4	482	0
soniox-v4	100%	2.90%	2	480	0
deepgram-nova-3	100%	3.10%	8	449	0

Switchboard conversational speech

Service	Transcript	Average WER	RTFx	Perfect	High WER
gladia-solaria-1	100%	35.80%	0	27	3
mistralai-voxtral-mini-transcribe-2	100%	50.10%	6	25	6
speechmatics	100%	56.00%	0	31	7
assemblyai-universal-3-pro	100%	56.00%	1	28	4

Service	Transcript	Average WER	RTFx	Perfect	High WER
elevenlabs-scribe_v2	100%	62.50%	0	31	6
soniox-v4	100%	62.90%	0	27	9
assemblyai-universal-2	100%	63.10%	1	20	8
deepgram-nova-3	100%	65.20%	1	26	7

8. Speaker diarization evaluation

DIHARD benchmark overview

Diarization is evaluated on the DIHARD III benchmark suite, which spans 10 diverse domains: broadcast audio, meetings, web video, socio-field recordings, court proceedings, clinical interviews, restaurant conversations, socio-lab recordings, conversational telephone speech (CTS), and map-task dialogues.

Gladia uses pyannoteAI "precision-2" for diarization. DIHARD datasets are provided during the "Speech Diarization Challenge" (https://dihardchallenge.github.io/dihard3/).

Diarization results

DER (Diarization Error Rate) comparison across providers. Lower is better.

Model	Broadcast	Meeting	Web Video	Socio Field	Court	Clinical	Restaurant	Socio Lab	CTS	Map task	Simple Avg	Weighted Avg
Gladia (solaria-1)	**9.4**	**29.9**	**44.4**	**12.3**	**3.9**	**13.3**	**41.3**	**5.5**	**7.7**	**4.5**	**17.2**	**16.6**

Model	Broadcast	Meeting	Web Video	Socio Field	Court	Clinical	Restaurant	Socio Lab	CTS	Map task	Simple Avg	Weighted Avg
Enhanced (Speechmatics)	17.2	55.6	55.6	28.9	15.0	24.9	58.4	18.6	20.1	23.4	31.8	30.1
Transcribe word-level (AWS)	16.4	51.4	60.3	25.2	16.7	27.3	63.1	20.2	31.2	22.9	33.5	33.8
OSS NeMo streaming sortformer (NVIDIA)	10.3	33.0	43.5	13.0	24.1	14.4	50.9	8.6	14.1	8.2	22.0	20.4
STT-async-preview-v1 (Soniox)	24.8	58.3	57.5	30.1	39.3	35.1	67.4	28.0	29.2	27.6	39.7	37.8
Scribe-v1 (ElevenLabs)	25.6	50.5	63.4	29.7	23.1	47.7	57.4	30.3	22.9	45.2	39.6	39.5
GPT-4o-transcribe-diarize (OpenAI)	26.4	57.8	64.1	28.8	30.0	40.8	59.7	26.5	N/A	34.8	41.0	42.8
Universal (AssemblyAI)	30.9	46.4	68.4	33.1	24.5	51.4	59.4	33.1	33.1	42.1	42.2	43.9

Model	Broadcast	Meeting	Web Video	Socio Field	Court	Clinical	Restaurant	Socio Lab	CTS	Map task	Simple Avg	Weighted Avg
Nova-3 (Deepgram)	27.0	59.7	83.0	35.5	25.6	44.8	75.2	32.2	35.5	45.9	46.4	46.9

Broadcast audio

Gladia solaria-1 achieves 9.4 DER, the best commercial result and competitive with open-source NeMo (10.3).

Meetings

Gladia solaria-1 achieves 29.9 DER, significantly outperforming all competitors (next best: NeMo at 33.0).

Web video

Gladia solaria-1 achieves 44.4 DER. NeMo is slightly better at 43.5, but all commercial providers are above 55.

Field recordings

Gladia solaria-1 achieves 12.3 DER on socio-field recordings, the best overall. NeMo is close at 13.0.

Telephony

Gladia solaria-1 achieves 7.7 DER on conversational telephone speech (CTS), the best result by a wide margin. The next-best commercial provider is Speechmatics at 20.1.

9. Error analysis

Representative transcription failures (WER > 100%) observed across datasets:

Common Voice 24:

- Soniox V4 transcribed English "Several deans and chairpersons work under this vice president" in Telugu script (277.8% WER)
- All providers failed on the short utterance "Chekhov." — producing outputs like "Check how", "Check out", or Chinese characters (200% WER)

Multilingual LibriSpeech:

- Reader instructions present in audio ("Jane Eyre Die Weise von Lowood von Charlotte Bronte Teil 1 Kapitel 3") are transcribed by all providers but absent from the ground truth, causing 185.7% WER for a 7-word reference

Switchboard:

- Very short segments like "oh" (1 word) where adjacent conversation bleeds in, causing 800–900% WER
- Laughter and non-speech tokens (<LAUGH>) in the reference that models expand into surrounding conversational context

Common failure modes

- **Overlapping speech**: All providers struggle with multi-party overlap, particularly in Switchboard
- **Accent variation**: Non-native accents in Common Voice 24 occasionally cause language-detection failures (Soniox V4 transcribing in wrong script)
- **Conversational disfluencies**: Short interjections ("uh", "uhhuh", "hm") are frequently expanded with surrounding context
- **Noisy environments**: ElevenLabs Scribe V2 is notably the weakest model on noisy audio (4.30% vs. 0.70–3.20% for others)

Language-specific errors

- **Italian** has the highest WER across all providers on Multilingual LibriSpeech (6.1–11.9%), likely due to less training data
- **Portuguese** shows high variance, with Deepgram Nova-3 at 11.30% vs. ElevenLabs Scribe V2 at 3.00%
- **German** and **French** are generally well-served by all providers (2.6–6.9% range)

10. Reproducibility

Benchmark results are straightforward to reproduce on a local setup:

- Download the dataset you want to benchmark from Hugging Face, make sure to use the test split.
- Implement the code that performs the transcription for the selected providers. Most provide an SDK to ease the implementation.
- Store the transcription result locally
- Compute the WER metrics by using the text normalization repository or the python package directly.
- Compare the results

Exact model versions

Provider	**Model Identifier**
Gladia	solaria-1
AssemblyAI	universal-3-pro
AssemblyAI	universal-2
ElevenLabs	scribe_v2
Deepgram	nova-3
Speechmatics	Enhanced
Soniox	v4
Mistral	voxtral-mini-transcribe-2

API configuration

Decoding parameters

Default parameters for all providers. Language is passed explicitly. Code switching is disabled.

Streaming vs batch usage

All evaluations use the asynchronous (batch) transcription API for each provider.

Timeout and retry policies

Benchmarks are run with a fair and similar timeout. Failed transcriptions are retried.

Execution environment

Hardware configuration

Benchmarks are executed via each provider's cloud API — no local GPU or CPU inference. Timing measurements reflect end-to-end API latency, including network round-trip and queue time.

Software dependencies

- Hugging Face for datasets
- metrics-api for WER/DER computation
- ffmpeg for audio channel normalization

Runtime environment

Gladiator runs on Gladia’s infrastructure. All provider accounts use pay-as-you-go billing with automatic credit top-ups.

Reproducing the benchmark

1. Download datasets from Hugging Face (test split only)
2. Flatten multi-channel audio to mono with ffmpeg
3. Configure provider API credentials
4. Run the benchmark against the target dataset and provider set
5. Retrieve WER/RTFx results from the scoring pipeline

11. Limitations

Dataset bias

Benchmark datasets may over-represent certain speaker demographics, recording conditions, or speech styles. Common Voice is crowd-sourced and may contain non-native accents labeled as native speech. Switchboard recordings date from the 1990s and may not reflect modern telephony conditions.

Model version drift

API models may change over time without notice. Results in this report reflect model versions available during the Q1 2026 benchmark campaign. Providers that update their models after the evaluation period may produce different results.

Domain coverage

Several speech domains are not covered: medical dictation, legal proceedings, broadcast news, code-switching within utterances, and languages beyond the six evaluated here.

12. Conclusion

This benchmark shows that speech recognition performance is not a single number. It is a function of domain, audio conditions, and evaluation methodology. Models that perform well on clean speech can degrade significantly in conversational, noisy, or multilingual settings. As a result, vendor-level comparisons without context are often misleading.

The primary takeaway is not which model ranks first, but how sensitive results are to evaluation methodology. Small differences in dataset selection, normalization, or scoring can significantly affect the outcomes. Without a shared, reproducible methodology, benchmark results are difficult to interpret and compare.

The implication is straightforward: benchmark results only matter if the evaluation setup is transparent and consistent. Otherwise, comparisons collapse under small methodological differences.

In practice, the most reliable way to evaluate speech recognition systems remains testing on your own audio, under your own conditions. This benchmark is a starting point, not a substitute for that process.

13. Future work

Planned extensions of the benchmark:

- Additional languages and CER (Character Error Rate) to cover non-Latin scripts
- Domain-specific benchmarks (medical, legal)
- Code-switching benchmark
- Key entities benchmarks
- Semantic WER / Embedding WER metrics
- Expanded streaming workloads
- Expanded diarization evaluation
- Additional providers (Google, Azure, etc.)