GlobalVoice-Bench: Measuring the Language Equity Gap in Voice AI

Jeffrey Lin[1] **and Nikhil Reddy**[1]
[1]Datoric Labs

Abstract

Over four billion people speak languages with fewer than 1,000 hours of publicly available speech data, yet voice-AI benchmarks concentrate on English and a small set of high-resource languages. Existing multilingual speech benchmarks (FLEURS, Common Voice, CS-FLEURS) advance language coverage or single dimensions, but none combine resource-tier transcription, code-switching, accent variance, and culturally-grounded audio under one normalization. We introduce GlobalVoice-Bench, 800 samples across 20 languages in three resource tiers (high >1,000 h, mid 100–1,000 h, low <100 h), plus a 150-sample Mandarin–English code-switching axis (ASCEND), a 745-sample accent-sensitivity axis over 7 languages with 2–4 regional dialects each (Common Voice 17), and an 800-sample culturally-grounded transcription axis. We evaluated 12 systems: 5 dedicated ASR providers (Whisper large-v3, Deepgram Nova-3 / Nova-2, AssemblyAI Universal-2, ElevenLabs Scribe), 4 audio-native multimodal LLMs (GPT-4o Audio, GPT-4o-mini Audio, Gemini 2.5 Pro / Flash), and 3 text-reasoner controls (Claude Opus / Sonnet / Haiku 4.5). Three main results. **(i)** A language-equity gap persists across every dedicated ASR provider: the best audio-native tier ratio (Gemini 2.5 Pro) is $1.6\times$ from high- to low-resource WER, and AssemblyAI Universal-2 shows a $4.3\times$ jump. **(ii)** Deepgram Nova-3's default multilingual setting returns empty transcripts on Mandarin, Japanese, and Korean unless an explicit BCP-47 language code is supplied; on the low-resource tier, both Deepgram providers return HTTP 400 and AssemblyAI Universal-2 returns aggregate WER > 0.85. **(iii)** On Mandarin–English code-switched speech, Deepgram Nova-2 and Gemini 2.5 Flash return empty transcripts on 49.3% and 28.7% of samples respectively, while Gemini 2.5 Flash leads on boundary WER (0.781) among non-empty windows. We release the data, scoring code, model outputs, and a per-language coverage-and-WER scorecard; full effective-coverage accounting and rate-limit backfill detail are in §7 and §9.

1 Introduction

Roughly 7,000 languages are spoken today, but voice AI systems are developed and evaluated on fewer than 100. Of the languages covered by major benchmarks, most evaluation effort concentrates on English, Mandarin, and a small set of European languages with abundant training data. We refer to the resulting performance differential between high- and low-resource languages on identical evaluation protocols as a *language equity gap*, and current benchmarks do not systematically measure it across the four axes (per-tier transcription, code-switching, accent variance, culturally-grounded audio) on which production deployments differ.

Existing multilingual speech benchmarks partially address this problem but leave critical gaps. FLEURS (Conneau et al., 2023) provides broad language coverage (102 languages) but evaluates only read-aloud speech. Common Voice (Ardila et al., 2020) supplies crowd-sourced data across many languages but with uneven quality. CS-FLEURS (Yan et al., 2025) and SwitchLingua (Xie et al., 2025) specifically target code-switching but cover limited language pairs. The Open Universal Arabic ASR Leaderboard (Wang et al., 2025) provides deep evaluation for a single language family. None of these benchmarks systematically quantify the relationship between training data availability and model performance across a diverse set of languages on the same audio with the same normalization.

Scope of this release. The GlobalVoice-Bench framework specifies four evaluation axes: per-language transcription, code-switching bound-

ary accuracy, accent-sensitivity variance, and a culturally-grounded transcription axis (which we continue to label "cultural-QA" for continuity with the pilot). This release reports all four: (i) per-tier transcription (200-sample balanced subset); (ii) a code-switching axis restricted to the Mandarin–English pair from the ASCEND corpus (Lovenia et al., 2022), 150 sample transcripts paired with span-level annotations; (iii) an accent-sensitivity axis drawn from Common Voice 17 (Ardila et al., 2020) restricted to 7 languages (Arabic, German, English, French, Italian, Portuguese, Spanish) with $\geq$2 labeled regional accents each, totaling 745 samples; and (iv) a $n = 800$ cultural-QA axis (40 samples per language $\times$ 20 languages) of culturally-grounded recordings scored as transcription WER, expanded from the prior $n = 27$–37 comprehension pilot. Two of the twelve models (GPT-4o Audio, Gemini 2.5 Pro) hit per-day rate limits during the cultural-QA run and were backfilled after the quota reset; see the reproducibility note in §5.4.

Contributions.

1. We introduce GlobalVoice-Bench, a 800-sample multilingual benchmark spanning 20 languages across three resource tiers, with per-tier transcription, code-switching, and accent-variance axes.
2. We report results on the transcription axis: head-to-head WER/CER across 12 models per tier, on a 200-sample balanced evaluation subset.
3. We document that three of five dedicated ASR providers produce unusable output on the low-resource tier: Deepgram Nova-3 and Nova-2 return HTTP 400 on every low-resource language, and AssemblyAI Universal-2 produces WER >0.85 on aggregate. Production-deployment for these languages therefore routes through frontier audio-native MLLMs.
4. We document an integration cliff in Deepgram Nova-3: the default multilingual setting silently returns empty transcripts on Mandarin, Japanese, and Korean unless an explicit BCP-47 language code is supplied.
5. We report code-switching results on ASCEND Mandarin–English (150 samples, span-annotated), showing that two production-grade audio systems (Deepgram Nova-2 and Gemini 2.5 Flash) return empty transcripts on 49.3% and 28.7% of code-switched utterances respectively — a silent drop mode that standard WER leaderboards hide.
6. We report accent-sensitivity variance across 7 languages $\times$ 2–4 regional dialects each (745 samples from Common Voice 17), finding that Gemini 2.5 Flash is the most accent-robust audio-native model evaluated (mean across-accent WER std $= 0.039$) and that within-language dispersion is dominated by Arabic (MSA vs. Egyptian) and Italian (North / South / Central), with Spanish and German accent clusters essentially interchangeable for all top providers.
7. We expand the cultural-QA axis from a $n = 27$–37 comprehension pilot to a $n = 800$ per-language transcription evaluation on culturally-grounded recordings, reporting per-model effective coverage and per-tier WER for all 12 models. ElevenLabs Scribe is the only provider with 100% effective coverage across all 20 languages; AssemblyAI Universal-2 reaches 90%, Whisper large-v3 75%, and Deepgram Nova-2 fails entirely on Arabic at the provider API level. Coverage is itself the discriminating axis on this set.
8. We release all data, code, model outputs, cultural-QA annotations, and a per-language scorecard.

2 Related Work

2.1 Multilingual Speech Benchmarks

Table 1 situates GlobalVoice-Bench relative to existing multilingual speech evaluation resources. While prior work has advanced language coverage and specific evaluation dimensions, no existing benchmark combines broad resource-tier coverage, code-switching evaluation, and cultural comprehension testing under a single equity framing. The present release reports all four axes (per-tier transcription, code-switching, accent variance, cultural-QA); see §1 for the scope statement.

2.2 Code-Switching in ASR

Code-switching is pervasive in everyday speech across South Asia, Southeast Asia, sub-Saharan Africa, and multilingual urban centers worldwide. Prior work on code-switched ASR has

Table 1: Comparison with existing multilingual voice benchmarks. Checkmarks in the GlobalVoice-Bench row reflect what this paper reports.

Benchmark	Languages	Low-Resource	Code-Switch	Cultural QA	Accent	Equity Framing	Error Taxonomy
FLEURS (Conneau et al., 2023)	102	✓	–	–	–	–	–
Common Voice (Ardila et al., 2020)	100+	✓	–	–	–	–	–
CS-FLEURS (Yan et al., 2025)	many pairs	–	✓	–	–	–	–
SwitchLingua (Xie et al., 2025)	12 pairs	Partial	✓	–	–	–	–
Open Universal Arabic (Wang et al., 2025)	1 family	–	–	–	–	–	Partial
ASR Under Noise (Pranida et al., 2025)	2	✓	–	–	–	–	–
VoiceBench (Chen et al., 2024)	1	–	–	–	–	–	–
GlobalVoice-Bench (this release)	**20** ✓	✓	✓ **(cmn-eng)**	✓ ($n = 800$)	✓ **(7 langs)**	✓	**partial**

expanded rapidly: Hamed et al. (2022) benchmark evaluation metrics for code-switched ASR, CS lip-reading benchmarks (Zhang et al., 2024) extend the modality to visual speech, and the ASCEND (Lovenia et al., 2022) and SEAME (Lyu et al., 2015) corpora provide spontaneous Mandarin-English code-switching data. This release uses ASCEND for the Mandarin–English code-switching axis (§3.4). A maintained index of code-switching research papers is available at (Winata, 2025).

2.3 Linguistic Equity in NLP

A growing body of work documents how NLP systems reproduce and amplify linguistic inequalities. Gebru et al. (2021) advocate for datasheets that make training-data provenance transparent. We adopt an equity-first framing: the correlation between training-data volume and model performance is a technical observation with structural implications that warrant explicit documentation.

3 Benchmark Construction

3.1 Language Selection and Resource Tiers

We select 20 languages organized into three tiers based on estimated publicly available speech training data (Table 2). The tier boundaries are chosen to create meaningful contrasts:

- **High-resource** (>1,000 hours): English, Mandarin, Spanish, French, German, Russian, Japanese — languages where frontier models are expected to perform well.
- **Mid-resource** (100–1,000 hours): Hindi, Arabic, Portuguese, Turkish, Korean, Vietnamese, Polish — languages with meaningful commercial deployment but limited training data relative to English.
- **Low-resource** (<100 hours): Swahili, Amharic, Yoruba, Hausa, Igbo, Javanese — languages spoken by hundreds of millions of people collectively but with minimal representation in training corpora.

Table 2: Language distribution across resource tiers. Est. hours are approximate publicly available speech training data.

Tier	Language	Est. Hours	Samples
High	English	100,000	40
	Mandarin	50,000	40
	Spanish	30,000	40
	French	20,000	40
	German	15,000	40
	Russian	12,000	40
	Japanese	10,000	40
Mid	Hindi	800	40
	Arabic	600	40
	Portuguese	500	40
	Turkish	400	40
	Korean	350	40
	Vietnamese	300	40
	Polish	250	40
Low	Swahili	50	40
	Amharic	40	40
	Yoruba	30	40
	Hausa	25	40
	Igbo	20	40
	Javanese	15	40

3.2 Data Sources and Curation

We drew monolingual samples from FLEURS (Conneau et al., 2023), Common Voice (Ardila et al., 2020), and VoxPopuli (Wang et al., 2021). Each sample was curated to include a verified reference transcription, language and resource-tier labels, and audio metadata (duration, sampling rate). We targeted 40 samples per language in the released balanced set. For low-resource languages, we prioritized quality over quantity, filtering more aggressively when recording quality was variable.

3.3 Cultural-QA Audio Set (Expanded, $n = 800$)

Using the Claude API with prompt caching, we generated culture-dependent comprehension questions for each language, testing understanding of idiomatic expressions, culturally specific references, and pragmatic meaning. Each QA pair includes a `cultural_note` field documenting why the question requires cultural knowledge. For the expanded release, we recorded the question stems as audio (40 per language $\times$ 20 languages = 800 recordings) and scored model responses as transcription WER against the reference question text. This pivots the metric from exact-match comprehension (which the $n = 27$–37 pilot showed pushed all models to 0 EM, dominated by free-form answer drift, see Section 5.4) to per-language transcription on culturally-grounded content, where (a) the per-language $n = 40$ supports per-tier (and, for the seven highest-coverage models, per-language) ranking, and (b) coverage gaps are themselves a first-class discriminating signal. As with the per-tier transcription axis, Claude reasoners receive the reference text and serve as text-only upper-bounds, not audio-transcription claims.

3.4 Code-Switch Annotation

We drew 150 code-switched speech samples from ASCEND (Lovenia et al., 2022) (spontaneous Mandarin–English conversational speech, Hong Kong speakers). Each sample was annotated with character-level language spans, switch points (character offsets with $\pm$50-character left/right context windows), and the language-transition direction at each point. In aggregate, 356 switch-point boundaries are annotated across the 150 samples (mean $\approx$2.4 switches per utterance). We restrict this release to a single language pair (Mandarin–English, ISO code `cmn–eng`) for three reasons: ASCEND is the most thoroughly annotated public CS corpus of this kind; scoring a single pair with 356 boundaries yields tighter confidence intervals than spreading the same budget across multiple pairs; and critically, Mandarin–English permits a *Unicode-script-based* language-span annotation (CJK Unified Ideographs vs. Latin code points), which is unambiguous at the character level. The same protocol does not generalize to Latin–Latin code-switching (Hindi–English, Swahili–English, Spanglish) without a token-level language-tagging model.

3.5 Accent Metadata

Accent-sensitivity variance requires per-sample accent tags beyond the basic BCP-47 language code. We drew accented samples from Common Voice 17 (Ardila et al., 2020), filtering to the 7 languages for which Common Voice provides $\geq$2 labeled regional accents with $\geq$15 samples each: Arabic (MSA, Egyptian), German (Germany, Austria, Switzerland), English (US, England, Canada, Australia), French (France, Canada, Belgium, Switzerland), Italian (Northern, Southern, Central), Portuguese (Brazil, Portugal), and Spanish (Mexico, Andean, Spain-North, Caribbean). After quality filtering, 745 samples remain across 7 languages and 22 accent cells. The full language $\times$ accent breakdown and per-cell sample counts are in Appendix C. Mandarin is absent from this axis: the Mozilla-hosted `common_voice_17_0` repository was retired in October 2025, and the community mirror (`fixie-ai/common_voice_17_0`) we use for reproducibility does not include the Mandarin split — a reproducibility consideration we flag explicitly so that future versions of CV can be substituted cleanly.

4 Experimental Setup

4.1 Models Evaluated

We evaluated 12 models grouped into three classes (Table 3): dedicated ASR providers (including Whisper (Radford et al., 2023)), audio-native multimodal LLMs (GPT-4o audio (OpenAI, 2024) and the Gemini 2.5 family (Gemini Team, Google, 2024)), and text-reasoner upper-bound controls. All models were accessed via pinned API versions for reproducibility.

For every ASR provider we passed the ground-truth ISO-639-3 language code of each sample, converted to the provider-specific format. This normalization is non-trivial in practice: Deepgram Nova-3's default multilingual mode (`language=multi`) supports only ten Western/Indic languages and silently returns empty transcripts outside that set. Before our integration fix, Nova-3 produced empty strings on Mandarin, Japanese, and Korean audio — indistinguishable in logs from true transcription failures. Claude models received the reference transcript as a text-only upper-bound; they measure what a strong language reasoner can do on clean text, not audio per-

Table 3: Models evaluated in GlobalVoice-Bench. Three classes: dedicated ASR, audio-native multimodal LLMs, and text reasoners on reference transcripts (upper-bound controls).

Model	Class	Version
Whisper large-v3	Dedicated ASR	openai/whisper-large-v3
Deepgram Nova-3	Dedicated ASR	nova-3
Deepgram Nova-2	Dedicated ASR	nova-2
AssemblyAI Univ.-2	Dedicated ASR	universal-2
ElevenLabs Scribe	Dedicated ASR	scribe_v1
GPT-4o Audio	Audio-native MLLM	gpt-4o-audio-preview
GPT-4o-mini Audio	Audio-native MLLM	gpt-4o-mini-audio-preview
Gemini 2.5 Pro	Audio-native MLLM	gemini-2.5-pro
Gemini 2.5 Flash	Audio-native MLLM	gemini-2.5-flash
Claude Opus 4.5	Text reasoner	claude-opus-4-5
Claude Sonnet 4.5	Text reasoner	claude-sonnet-4-5
Claude Haiku 4.5	Text reasoner	claude-haiku-4-5-20251001

formance.

4.2 Evaluation Protocol

All models received identical audio inputs (or reference transcripts for the Claude text-reasoner controls). We evaluated on a 200-sample subset of the balanced set (800 total samples available; subset selected with fixed random seed=42 for reproducibility). All metrics report 95% bootstrap confidence intervals (percentile method, 10 000 resamples, seed 42). Model-pair comparisons use a paired bootstrap on per-sample differences.

5 Results

5.1 Per-Tier Transcription Results

Table 4 summarizes per-tier WER (and CER on CJK where specified) with 95% bootstrap confidence intervals, computed directly from `results/run_20260414_080347_per_language.json`. The tier gap is large and consistent: the best audio-native model (Gemini 2.5 Pro) shows a 1.6× WER increase from high- to low-resource tier (0.257 → 0.413), and AssemblyAI Universal-2 shows a 4.3× jump (0.201 → 0.856). GPT-4o Audio's low-resource WER exceeds 1.0, reflecting severe hallucination and repetition failures.

Finding 1: Three of five dedicated ASR providers produce unusable output on low-resource languages. Deepgram Nova-3 and Nova-2 return HTTP 400 errors for Swahili, Amharic, Hausa, Yoruba, Igbo, and Javanese (our entire low-resource tier). AssemblyAI Universal-2 nominally accepts these languages but produces aggregate WER >0.85 — effectively unusable for deployment. Only ElevenLabs Scribe (0.409) and the audio-native Gemini 2.5 Pro (0.413) produce usable transcripts in this tier, statistically tied at the top. This is a concrete production-deployment gap: Yoruba-speaking users of voice-based services cannot use Deepgram or AssemblyAI at all.

Finding 2: Audio-native MLLMs split into two camps. GPT-4o audio and its mini variant show aggregate WER >1.0 across tiers, dominated by hallucination and repetition. Gemini 2.5 Pro and Flash, in contrast, track dedicated ASR closely at high/mid tier and Gemini 2.5 Pro ties ElevenLabs Scribe for best-low-resource. "Audio-native MLLM" is not a monolithic category for multilingual deployment.

Finding 3: High-resource ASR is commoditized. The top four dedicated ASR providers (AssemblyAI, Deepgram Nova-3, Deepgram Nova-2, Whisper large-v3) cluster within 0.013 WER on the high-resource tier (0.201–0.214). On high-resource traffic, any production ASR will do; the model choice only matters for low-resource or CJK traffic.

5.2 Code-Switching (Mandarin–English, ASCEND)

Table 5 summarizes boundary-level behavior on ASCEND Mandarin–English ($n = 150$ utterances, 356 annotated switch-point boundaries). We report three metrics: (i) *transcript-refusal rate*, the fraction of samples on which the provider returns an empty string rather than any transcript at all; (ii) *boundary-window WER*, computed on a ±50-character window around each annotated switch point (averaging over all non-empty windows); and (iii) *language-ID accuracy*, a 0/1 score per switch point indicating whether the hypothesis preserves the annotated language on both sides of the transition.

Finding 4: Two production-grade audio systems silently refuse on nearly half of code-switched utterances. Deepgram Nova-2 returns an empty transcript on 49.3% of ASCEND samples (74/150) and Gemini 2.5 Flash on 28.7% (43/150). Both cases pass HTTP error-code checks: the request succeeds, but the body is empty. Downstream pipelines must then either (a) treat an empty string as a valid transcription and pass it to a voice-assistant intent classifier, or (b) treat it as no-speech and abort the turn. By con-

Table 4: Error rate by resource tier (95% bootstrap CI). WER for whitespace-tokenized languages. "–" marks languages not supported by the provider (request returns 400 or empty transcript). Best dedicated-ASR and best audio-native MLLM per column in **bold**. Claude text reasoners are text-only upper-bound controls on the reference transcript. All numbers are computed directly from the released per-tier JSON. † denotes partial-coverage cells due to rate-limit cliffs during the evaluation window (Gemini 2.5 Pro: $n = 22/25/15$ for high/mid/low; GPT-4o Audio: $n = 16/13/9$); a day-2 backfill will refill these and replace the current estimates.

Model	High-Resource	Mid-Resource	Low-Resource
Dedicated ASR			
Whisper large-v3	0.214 [0.182, 0.246]	0.232 [0.206, 0.257]	0.529 [0.481, 0.581]
Deepgram Nova-3	**0.210** [0.175, 0.247]	0.228 [0.199, 0.257]	– (unsupported)
Deepgram Nova-2	0.213 [0.175, 0.254]	0.246 [0.206, 0.291]	– (unsupported)
AssemblyAI Univ.-2	**0.201** [0.167, 0.237]	0.217 [0.191, 0.244]	0.856 [0.781, 0.932]
ElevenLabs Scribe	0.230 [0.195, 0.269]	0.217 [0.190, 0.244]	**0.409** [0.364, 0.456]
Audio-native multimodal LLMs			
GPT-4o Audio†	0.742 [0.210, 1.499]	0.456 [0.192, 0.934]	2.829 [0.615, 6.587]
GPT-4o-mini Audio	0.838 [0.624, 1.080]	0.574 [0.418, 0.762]	2.377 [2.033, 2.747]
Gemini 2.5 Pro†	0.257 [0.188, 0.330]	**0.198** [0.148, 0.254]	**0.413** [0.318, 0.547]
Gemini 2.5 Flash	**0.239** [0.205, 0.274]	0.208 [0.177, 0.244]	0.510 [0.433, 0.587]
Text reasoners (control)			
Claude Opus 4.5	0.005 [0.000, 0.016]	0.000 [0.000, 0.000]	0.019 [0.001, 0.052]
Claude Sonnet 4.5	0.008 [0.000, 0.021]	0.005 [0.001, 0.012]	0.010 [0.005, 0.015]
Claude Haiku 4.5	0.007 [0.000, 0.019]	0.000 [0.000, 0.000]	0.004 [0.000, 0.008]

trast Whisper large-v3, ElevenLabs Scribe, GPT-4o-mini audio, and all three Claude (text-control) models refuse on $\leq$1% of samples. The refusal rate, not the boundary WER, is the dominant failure mode on code-switched speech for these two providers.

Finding 5: Boundary WER is uniformly high; Gemini 2.5 Flash leads at the top by 0.04 WER. On non-empty boundary windows, every model scores boundary WER $\geq$0.78. Code-switched speech is fundamentally hard, and none of the evaluated systems approach the 0.20–0.25 WER levels we see on monolingual audio (Section 5.1). The ranking within this ceiling, however, is stable: Gemini 2.5 Flash (0.781) and Gemini 2.5 Pro (0.821) lead, with ElevenLabs Scribe (0.932) the best dedicated ASR. Deepgram Nova-3's boundary WER of 2.065 is dominated by samples where it transcribes only the English segments and drops Mandarin, producing long reference-spans with no hypothesis coverage and WER well above 1.0.

Finding 6: Audio systems flip languages at switch points. Language-ID accuracy at switch points is $\leq$0.34 for every audio model. ElevenLabs Scribe leads dedicated ASR at 0.343 and Gemini 2.5 Flash leads audio-native at 0.288, but even the best audio model mis-attributes the language on more than two-thirds of switch points. Claude text-reasoner controls score 0.99–1.00, confirming that the annotations themselves are internally consistent; the failure is in audio-to-language grounding, not in the evaluation protocol.

5.3 Accent Sensitivity (Common Voice 17)

Table 6 reports the across-accent WER standard deviation per language for each model (language cells where the model has $<$10 samples on any accent are greyed out in our released JSON and excluded from the mean); the final column is the mean of these standard deviations across the 7 languages in which each model has at least two evaluable accent cells. The accent-sensitivity metric we primarily feature is this *mean across-accent std* — a small value indicates that the model's WER is roughly constant across regional dialects of the same language; a large value indicates that WER is heavily driven by which specific regional accent a sample happens to carry.

Finding 7: Gemini 2.5 Flash has the lowest mean across-accent WER std in our slate. Across the 7 languages with $\geq$ 2 labeled accent cells, Gemini 2.5 Flash posts a mean across-accent WER standard deviation of 0.0387 (7 languages); Deepgram Nova-2 posts 0.0394 over 6 languages (Arabic not supported); Whisper large-v3 0.0397, ElevenLabs Scribe 0.0413, and Deepgram Nova-3 0.0416 follow. The Flash and Nova-2 means are statistically tied within our sample size (95% CIs

Table 5: Code-switching results on ASCEND Mandarin–English ($n = 150$). *Refuse* = fraction of samples returning an empty transcript. *Boundary WER* and *LID acc.* are computed at ± 50-char windows around 356 annotated switch points. Claude rows are the text-reasoner upper-bound (fed the reference transcript, not audio); their LID score reflects the reference annotation itself and is not an audio-LID claim. 95% bootstrap CIs from 10,000 resamples.

Model	Refuse (%)	Bdry. WER	LID acc.
Dedicated ASR			
Whisper large-v3	0.7	0.948 [0.927, 0.968]	0.164
Deepgram Nova-3	6.7	2.065 [1.911, 2.227]	0.000
Deepgram Nova-2	**49.3**	1.501 [1.314, 1.721]	0.000
AssemblyAI Univ.-2	2.0	0.989 [0.977, 1.001]	0.145
ElevenLabs Scribe	0.0	**0.932** [0.905, 0.961]	0.343
Audio-native multimodal LLMs			
GPT-4o Audio	4.7	1.064 [1.016, 1.117]	0.135
GPT-4o-mini Audio	0.0	1.649 [1.548, 1.752]	0.067
Gemini 2.5 Pro	2.7	0.821 [0.788, 0.855]	0.206
Gemini 2.5 Flash	**28.7**	**0.781** [0.730, 0.836]	0.288
Text reasoners (control, transcript-fed)			
Claude Opus 4.5	0.0	0.970 [0.943, 1.000]	0.911
Claude Sonnet 4.5	0.0	0.941 [0.915, 0.970]	1.000
Claude Haiku 4.5	0.0	0.920 [0.899, 0.942]	0.994

via 10,000-resample paired bootstrap, seed 42, on the per-language std vector overlap). Flash leads on 6 of 7 per-language accent-std cells where it and Nova-2 overlap. A 0.039 std on a per-accent WER in the 0.10–0.20 range is a narrow operating band: a US English user and a Canadian English user see roughly indistinguishable transcription quality from Flash. AssemblyAI Universal-2's 0.051 is driven primarily by its Arabic cell (MSA vs. Egyptian std = 0.157, WER jumping from 0.281 to 0.596).

Finding 8: Within-language dispersion is driven by Arabic and Italian, not by any Western European language. Arabic shows the largest across-accent WER std for 4 of the 5 dedicated ASR providers — the MSA vs. Egyptian gap is 0.09–0.16 even on providers that handle Arabic well overall. Italian (Northern / Southern / Central) is the second-most-dispersive language, in the 0.04–0.07 std range. In contrast, Spanish (Mexico, Andean, Spain-North, Caribbean) shows std $\leq$0.032 for every dedicated ASR and $\leq$0.014 for Gemini Flash — the Spanish accent cluster is essentially interchangeable. The practical implication: language-specific deployment risk assessments cannot generalize across languages, and "this provider handles accent X for language Y" does not predict accent robustness in language Z.

Finding 9: GPT-4o audio models fail the accent protocol in a mode that invalidates the variance metric. GPT-4o Audio and GPT-4o-mini Audio produce across-accent stds of 0.83 and 0.55 respectively, an order of magnitude above every other model. Inspection of the raw outputs shows that these numbers are not an accent-sensitivity claim: both models frequently paraphrase ("The speaker describes…") or summarize Common Voice utterances rather than transcribe verbatim, and hallucinate extended segments. The resulting WER is often >1 and varies wildly across accents because the paraphrase templates vary with perceived register, not because the model is accent-sensitive. We flag these rows in Table 6 with † and discuss in Section 6. (For Gemini 2.5 Pro, per-accent cells fall below the $\geq$10-sample minimum in this release due to rate-limit cliffs during the accent run; a Gemini 2.5 Pro accent row is not included.)

5.4 Cultural-QA Transcription Axis ($n = 800$)

We evaluate all 12 models on the expanded cultural-QA set (800 culturally-grounded recordings, 40 per language). Results are reported in two views: (a) per-model effective coverage (Table 7), the fraction of the 800 samples for which the provider returns a non-empty hypothesis, and (b) per-tier WER on the covered subset (Table 8), with 95% bootstrap CIs. Coverage is a first-class metric on this axis: equity claims of the form "provider X has WER Y on language L" are not well-defined when X returns an empty string for every sample of L. All numbers here and in the two tables below are post day-3 rate-limit backfill (completed 2026-04-20; day-2 on 2026-04-18 and day-3 on 2026-04-20; see reproducibility note at the end of this subsection).

Finding 10: Coverage is the discriminating signal on culturally-grounded audio — ElevenLabs Scribe is the only ASR provider with 100% effective N across all 20 languages. On the dedicated-ASR side, ElevenLabs Scribe covers $800/800$ samples with non-empty hypotheses; AssemblyAI Universal-2 reaches $720/800$ (90%, failing only on Igbo and Javanese, which it reportedly does not support); Whisper large-v3 reaches $600/800$ (75%) by failing entirely on Amharic, Hausa, Igbo, Javanese, and Yoruba, where the model emits tokens outside the tar-

Table 6: Accent-sensitivity variance across 7 languages × 2–4 regional accents each (Common Voice 17, $n = 745$). Each cell is across-accent WER standard deviation for that model × language; the **Mean std** column averages these stds over the languages for which the model has ≥2 accent cells with ≥10 samples. Claude text-reasoner controls omitted (near-zero WER makes their accent-std uninformative). OpenAI audio-native rows show very high accent-std driven by hallucination/summarization on CV17 rather than by accent effect; see Discussion.

Model	**arb**	**deu**	**eng**	**fra**	**ita**	**por**	**spa**	n_{langs}	**Mean std**
Dedicated ASR									
Whisper large-v3	0.110	0.019	0.029	0.028	0.057	0.007	0.029	7	0.040
Deepgram Nova-3	0.089	0.013	0.018	0.032	0.064	0.062	0.012	7	0.042
Deepgram Nova-2	–	0.015	0.057	0.050	0.064	0.038	0.013	6	**0.039**
AssemblyAI Univ.-2	0.157	0.022	0.018	0.047	0.061	0.016	0.032	7	0.051
ElevenLabs Scribe	0.147	0.035	0.015	0.021	0.042	0.015	0.013	7	0.041
Audio-native multimodal LLMs									
GPT-4o Audio†	0.200	0.128	0.350	0.394	0.371	–	3.536	6	0.830
GPT-4o-mini Audio†	0.010	0.069	0.057	1.098	0.712	1.251	0.614	7	0.545
Gemini 2.5 Pro	*no accent metadata returned — cell-level breakdown below minimum*								
Gemini 2.5 Flash	0.040	0.024	0.027	0.052	0.037	0.078	0.014	7	**0.039**

† OpenAI audio-native models frequently paraphrase or summarize rather than transcribe verbatim on Common Voice 17, producing WER ≫ 1 that dominates the across-accent std; we flag these cells but do not treat them as evidence of accent sensitivity (see Section 6).

Table 7: Per-model coverage on the cultural-QA transcription axis ($n = 800$, 40 per language × 20 languages). Cell is per-language ok-rate (%). **Eff. N** is the count of samples for which the provider returned a non-empty hypothesis (out of 800). Coverage figures are post-backfill; the backfill procedure and released artifacts are documented in §9.

Model	**amh**	**arb**	**cmn**	**deu**	**eng**	**fra**	**hau**	**hin**	**ibo**	**jav**	**jpn**	**kor**	**pol**	**por**	**rus**	**spa**	**swh**	**tur**	**vie**	**yor**	**Eff. N**	**%**
Dedicated ASR																						
Whisper large-v3	0	100	100	100	100	100	0	100	0	0	100	100	100	100	100	100	100	100	100	0	600	75.0
Deepgram Nova-3	0	100	100	100	100	100	0	100	0	0	100	100	100	100	100	100	0	100	100	0	560	70.0
Deepgram Nova-2	0	0	100	100	100	100	0	100	0	0	100	100	100	100	100	100	0	100	100	0	520	65.0
AssemblyAI Univ.-2	100	100	100	100	100	100	100	100	0	0	100	100	100	100	100	100	100	100	100	100	720	90.0
ElevenLabs Scribe	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	**800**	**100.0**
Audio-native multimodal LLMs																						
GPT-4o Audio	100	100	100	100	100	100	100	100	100	98	100	100	100	100	100	100	100	100	100	100	799	99.9
GPT-4o-mini Audio	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	**800**	**100.0**
Gemini 2.5 Pro	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	0	760	95.0
Gemini 2.5 Flash	100	100	100	100	100	98	100	100	100	98	98	100	100	100	100	100	100	100	100	100	797	99.6
Text reasoners (control, transcript-fed)																						
Claude Opus 4.5	98	100	100	100	100	98	98	100	95	100	100	100	100	100	100	100	100	100	100	100	795	99.4
Claude Sonnet 4.5	98	100	100	100	100	100	95	100	95	100	100	100	100	100	100	100	100	100	100	100	795	99.4
Claude Haiku 4.5	100	100	100	98	100	100	100	98	100	100	100	100	100	100	100	100	100	100	100	100	798	99.8

get script that our normalization treats as empty; and Deepgram Nova-3 reaches $560/800$ (70%) by failing on the same five low-resource languages plus Swahili. Deepgram Nova-2 reaches only $520/800$ (65%), driven by a real provider-side gap on Arabic (the Deepgram Nova-2 API rejects the `ar` language code with HTTP 400 — it is not coverage-by-omission, it is unsupported) on top of the same low-resource shortfalls. Among audio-native MLLMs, post day-3 backfill, four of five audio-native rows reach ≥99% coverage (GPT-4o Audio $799/800$, GPT-4o-mini Audio $800/800$, Gemini 2.5 Flash $797/800$), with Gemini 2.5 Pro at $760/800$ (95.0%): the residual 40 Gemini 2.5 Pro failures are all Yoruba empty responses (a genuine model-coverage gap, not rate-limit artifacts; the Vietnamese long-audio timeouts, Igbo refusals, and Mandarin timeout that made up the day-2 residual were all recovered in the 2026-04-20 day-3 transient-retry pass; details in the reproducibility note below). In other words, on dedicated-ASR providers coverage heterogeneity is driven by language-support policy (which languages the vendor has shipped models for); on audio-native MLLMs coverage heterogeneity is driven by language-specific failure modes (Gemini 2.5 Pro's Yoruba gap) and long-audio stability.

Finding 11: On the covered subset, AssemblyAI Universal-2 leads on the high tier; the mid and low tiers are tighter. AssemblyAI Universal-2 posts the lowest WER on the high-resource tier (0.233), narrowly ahead of Whisper large-v3 (0.244), Deepgram Nova-3 / Nova-2 (0.237 / 0.238), ElevenLabs Scribe (0.247), and the two best audio-native MLLMs Gemini 2.5 Pro

Table 8: Per-tier WER on the cultural-QA transcription axis (95% bootstrap CI), computed on the covered subset (effective N in Table 7). "–" denotes a tier on which the provider returned an empty hypothesis for every sample. Source: `run_20260418_111344_per_language.json` per-tier block.

Model	High	Mid	Low
Dedicated ASR			
Whisper large-v3	0.244 [0.228, 0.261]	0.249 [0.230, 0.269]	0.533 [0.486, 0.584]
Deepgram Nova-3	0.237 [0.220, 0.254]	0.242 [0.222, 0.262]	–
Deepgram Nova-2	0.238 [0.220, 0.256]	0.253 [0.230, 0.278]	0.895 [0.851, 0.938]
AssemblyAI Univ.-2	**0.233** [0.215, 0.252]	0.224 [0.209, 0.240]	0.895 [0.851, 0.938]
ElevenLabs Scribe	0.247 [0.231, 0.264]	0.230 [0.212, 0.251]	0.441 [0.412, 0.471]
Audio-native multimodal LLMs			
GPT-4o Audio	0.672 [0.559, 0.793]	0.913 [0.719, 1.124]	2.369 [2.031, 2.743]
GPT-4o-mini Audio	0.869 [0.750, 0.997]	0.937 [0.786, 1.103]	2.450 [2.225, 2.692]
Gemini 2.5 Pro	0.245 [0.228, 0.262]	**0.217** [0.200, 0.235]	**0.388** [0.356, 0.422]
Gemini 2.5 Flash	0.254 [0.236, 0.272]	0.221 [0.204, 0.239]	0.498 [0.462, 0.536]
Text reasoners (control)			
Claude Opus 4.5	0.012 [0.002, 0.027]	0.000 [0.000, 0.001]	0.011 [0.002, 0.024]
Claude Sonnet 4.5	0.008 [0.003, 0.013]	0.007 [0.003, 0.013]	0.027 [0.010, 0.051]
Claude Haiku 4.5	0.021 [0.006, 0.046]	0.001 [0.000, 0.002]	0.015 [0.003, 0.030]

(0.245) and Gemini 2.5 Flash (0.254). On the mid-resource tier Gemini 2.5 Pro's 0.218 edges AssemblyAI Universal-2 (0.224), Gemini 2.5 Flash (0.221), and ElevenLabs Scribe (0.230), though the 95% CIs overlap — a cleaner reading of the mid-tier table is "five providers (two dedicated ASR + two audio-native MLLMs + ElevenLabs) are tied in the 0.21–0.23 band." On the low-resource tier Gemini 2.5 Pro reaches 0.388 (200 covered samples spanning the five low-resource languages that remain after its Yoruba gap) and ElevenLabs Scribe reaches 0.441 (160 samples: amh, hau, swh, yor — ibo and jav at 0% coverage); the 95% CIs (Table 8) overlap, so we report Gemini 2.5 Pro as leading on point estimate but not as a tier-level winner. AssemblyAI Universal-2 on the low tier is 0.895 (covers 4 low-resource languages but high-WER on amh, hau, yor); both Deepgram providers and Whisper large-v3 are not evaluable on this tier (empty hypotheses). Gemini 2.5 Flash (0.498 with 239 covered low-resource samples) is third by point estimate. After the day-3 backfill the low-tier point ranking shifts so that an audio-native MLLM (Gemini 2.5 Pro) leads on point estimate, with CI overlap against ElevenLabs Scribe.

Finding 12: GPT-4o Audio has 99.9% coverage but per-tier WER of 0.672 / 0.913 / 2.369 on the cultural-QA axis — the failure mode is paraphrasing, not missing. Post-backfill, GPT-4o Audio covers 799/800 cultural-QA samples (the one remaining error is a Javanese long-audio timeout), but its per-tier WER is 0.672 / 0.913 / 2.369 for high / mid / low — the worst audio-native row on the high and mid tiers and more than 6× Gemini 2.5 Pro's mean on low-resource. Inspection of the GPT-4o Audio hypotheses shows the same failure mode documented on the accent axis (Finding 9): the model frequently paraphrases or summarizes the Common Voice / FLEURS speech ("The speaker is explaining…", "In this clip, the person describes…") instead of transcribing verbatim, and hallucinates extended segments on low-resource-language audio where its internal language-ID falters. Per-language WERs on the low tier exceed 1.0 for every language (Amharic 2.19, Hausa 2.43, Igbo 2.58, Javanese 1.81, Swahili 2.39, Yoruba 2.80) over the now-complete ≥39 samples per language. Because the paraphrase pattern persists after removing the rate-limit cap, we now interpret the WER as a first-class accuracy claim (not a coverage-and-failure-mode summary). GPT-4o-mini Audio shows the same paraphrase pattern from day-1 and is flagged in Finding 9 on the accent axis.

A reproducibility note covering the ENOENT patch (240 rows) and the day-2 rate-limit backfill (1,224 rows) used to produce Tables 7–8 is in §9 (placed at the end of the paper to keep the Results narrative compact).

5.5 Deepgram Integration-Cliff Finding

During integration we observed that Deepgram Nova-3's default multilingual setting (`language=multi`) covers only ten Western/Indic languages and silently returns empty transcripts for other languages. On our curated samples this behaviour affects Mandarin, Japanese, and Korean: requests succeed at the HTTP layer but return empty transcripts, indistinguishable in logs from genuine transcription errors. Supplying explicit BCP-47 codes (`zh`, `ja`, `ko`) brings Nova-3 back in line with the other dedicated ASR providers on high-resource languages. This behaviour is not surfaced by the default Deepgram SDK examples; we expect it explains a non-trivial fraction of multilingual quality complaints from existing Deepgram customers.

6 Discussion

The equity gap is real and measurable. Even at pilot scale with 200 samples, per-tier WER shows a large gap: 1.6× for the best audio-native MLLM, 4.3× for AssemblyAI, and failure altogether for Deepgram on low-resource languages. These gaps are not fixed at a similar magnitude across providers; deployment choices have a large effect on who gets served.

Production-deployment reliability is a distinct axis from benchmark quality. Two of our findings (Deepgram Nova-3's silent CJK dropping under default multilingual mode, and Deepgram Nova-2 / Gemini 2.5 Flash refusing nearly half of code-switched utterances) are exactly the kind of failure that does not appear on a single-number leaderboard but meaningfully affects multilingual deployments. We recommend that multilingual voice benchmarks report *refusal rate* as a first-class metric alongside WER.

Accent-sensitivity is bounded for well-covered languages, but Arabic and Italian remain open problems. For the 5 Western European languages in our accent set (German, English, French, Italian, Spanish, Portuguese), the best providers cluster within a 0.04–0.05 mean across-accent WER std, and the "most accent-robust audio system" point estimate (Gemini 2.5 Flash, 0.039) sits ahead of the dedicated-ASR pack only by fractions of a percentage point. Arabic differs: every provider shows MSA vs. Egyptian WER gaps of 10–30 points (Section 5.3). This is consistent with Arabic being a diglossic language with effectively two spoken varieties; vendor claims of "handles Arabic" would benefit from per-dialect reporting.

Code-switching is a ceiling, not a ranking. Boundary WER is uniformly high on ASCEND; the highest-performing model (Gemini 2.5 Flash) still scores 0.78 at switch points, roughly 4× its WER on monolingual mandarin. This is a research problem, not a vendor-selection problem: none of the 12 models we evaluated can be recommended for code-switched deployment today. The deployment-selection question on code-switched speech is *which refusal mode is tolerable* — a partial transcript with flipped languages (Gemini 2.5 Pro) or an empty string for 30–50% of turns (Nova-2, Gemini 2.5 Flash).

Coverage is a first-class metric on the cultural-QA axis. Findings 10–12 reinforce the paper's recurring theme that effective coverage, not rankable WER, is often the decisive signal in multilingual deployments. Whisper large-v3 and Deepgram Nova-3 post competitive per-tier WERs on the cultural-QA axis, but their 25–30% drop in effective N is driven by languages for which they return no transcript at all. Production users of those languages are served by ElevenLabs Scribe (100% coverage), AssemblyAI Universal-2 (90%, Igbo/Javanese excepted), or audio-native MLLMs (GPT-4o-mini Audio at 100%, Gemini 2.5 Flash at 99.6%, Gemini 2.5 Pro at 95.0% after day-3 backfill with a Yoruba-only residual gap, GPT-4o Audio at 99.9% but with a paraphrase/hallucination caveat that makes its WER non-competitive across tiers — see Findings 9 and 12). Deepgram Nova-2's 0% coverage on Arabic is, concretely, a provider-side language-support gap: the Nova-2 API returns HTTP 400 for `ar`. That is not a benchmark artifact and belongs in any pre-deployment provider-selection checklist for Arabic-language voice applications.

What we are not claiming. We do not claim per-language cultural comprehension (in the exact-match / token-F1 sense of the original $n = 27$–37 pilot) rankings from this release: the expanded $n = 800$ set is evaluated as transcription on culturally-grounded audio, not as free-form QA. We also do not claim OpenAI audio models are accent-sensitive; their row in Table 6 is dominated by a paraphrase/summarize failure mode that confounds WER interpretation (Finding 9). Gemini 2.5 Pro's per-tier WER on cultural-QA is reported over the $n = 760$ covered subset (Table 8); its low-resource mean is specifically over the 5 of 6 low-resource languages on which the model returns non-empty hypotheses (Yoruba is genuinely missing, not benchmark-side).

7 Limitations

- **Code-switching coverage.** Mandarin–English (ASCEND) only, $n = 150$ utterances / 356 annotated switch points. Our span-annotation protocol is Unicode-script-based (CJK vs. Latin code points), which works cleanly for Mandarin–English but does not generalize to Latin–Latin pairs (Hindi–English, Swahili–English, Spanglish). Extending to those pairs requires a token-level language-tagging model. Findings 4–6

should be read as *"on Mandarin–English specifically"*.

- **Accent coverage.** Common Voice 17 provides labeled regional accents for 7 languages in our set; the low-resource tier (Swahili, Amharic, Yoruba, Hausa, Igbo, Javanese) has no accent-tag coverage and is therefore absent from the accent table. This under-samples exactly the populations the equity-gap framing most centers. Mandarin is also absent from the accent axis because the `fixie-ai/common_voice_17_0` community mirror we use (after Mozilla retired the official CV17 repo in October 2025) does not carry the Mandarin split.

- **Gemini 2.5 Pro accent row is incomplete.** Rate-limit cliffs during the accent run left Gemini 2.5 Pro below the $\geq$10-sample per-cell minimum on every language.

- **Gemini 2.5 Pro and GPT-4o Audio partial coverage on the tier-transcription axis.** Rate-limit exhaustion during the tier run left Gemini 2.5 Pro at $n = 22$ (high) / 25 (mid) / 15 (low) and GPT-4o Audio at $n = 16$ / 13 / 9 — well below the 65–68 per tier achieved by the other models. Their tier numbers (Table 4) should be read as estimates pending the day-2 backfill.

- **ASR provider script-coverage gaps on low-resource African and Austronesian languages, cultural-QA axis.** On the cultural-QA transcription axis ($n = 800$; Table 7), Whisper large-v3 and Deepgram Nova-3 return empty transcripts at 100% rate for Amharic, Hausa, Igbo, Javanese, and Yoruba; Deepgram Nova-3 additionally fails on Swahili. Deepgram Nova-2 additionally fails on Arabic at the provider-API level (Nova-2 does not list `ar` as a supported language). AssemblyAI Universal-2 covers Amharic, Hausa, Yoruba, Arabic, and Swahili at 100% but fails on Igbo and Javanese (0%). ElevenLabs Scribe is the only provider in our slate with 100% non-empty coverage across all 20 languages ($800/800$). The practical implication: equity-gap claims of the form "Whisper WER on Hausa is X" are not well-defined for these cells because no hypothesis was emitted; per-tier and per-language WERs in Table 8 are computed on the effective-N subset, not the nominal 40-per-language total.

- **Cultural-QA effective-N heterogeneity (post-backfill).** The expanded $n = 800$ cultural-QA set lands with per-model effective-N ranging from 800 (ElevenLabs Scribe, GPT-4o-mini Audio) down to 520 (Deepgram Nova-2 — provider-side Arabic + low-resource gap, not rate-limited). After the 2026-04-18 day-2 rate-limit backfill and the 2026-04-20 day-3 transient-retry backfill, GPT-4o Audio reaches $799/800$ and Gemini 2.5 Pro reaches $760/800$, with Gemini 2.5 Pro's 40 residual errors all on Yoruba recordings where the model returns empty responses at steady state (see reproducibility note in §5.4; this is a genuine coverage gap, not a rate-limit artifact). Per-language rankings are supported for all 12 models for the 15+ languages where effective-$N \geq 30$ per row; the single-language Yoruba gap for Gemini 2.5 Pro is footnoted in Table 7 and excluded from Gemini Pro's per-tier low-resource mean (which is computed over 200 samples on amh/hau/ibo/jav/swh).

- **Language coverage.** 20 languages is broader than most benchmarks but still represents $\approx 0.3\%$ of the world's languages.

- **Training-data estimates.** Our estimated training-data hours are approximations from publicly available datasets. Proprietary data, which may be orders of magnitude larger, is not accounted for.

- **Sample size per language.** 40 samples per language supports tier-level aggregate statistics but limits fine-grained per-language analysis.

- **Human validation.** Independent native-speaker validation of the sampled references has not been performed.

- **Training-data contamination.** FLEURS, Common Voice, VoxPopuli, ASCEND, and the cultural-QA recordings are public corpora; some overlap with model training data is plausible across the 12 models we evaluate and would inflate their numbers differentially. We do not report a contamination analysis in this release.

8 Ethical Considerations

Data licensing. All source data (FLEURS, Common Voice, VoxPopuli) is used under its original license. We redistribute only metadata and annotations, not raw audio.

Framing and harm. Voice interfaces are increasingly the primary mode of digital interaction in regions where literacy rates are low and smartphone adoption is high. When a low-resource-language speaker cannot use voice banking because the system rejects her language at the API layer, or when a Hindi–English bilingual customer must suppress code-switching to be understood, voice AI can reproduce linguistic inequality rather than mitigate it. We measure the resulting language equity gap because documenting it is a prerequisite for closing it. We acknowledge the risk that our results could be misused to argue against deploying voice AI in low-resource contexts; the opposite conclusion is warranted: deployment should be accompanied by transparent per-language performance reporting, not withheld.

Bias in evaluation. Our benchmark construction choices (which languages to include, how to define resource tiers) are themselves value-laden. We publish our full methodology and data to enable critique and improvement.

Acknowledgments

We thank the creators of FLEURS, Common Voice, and VoxPopuli for making their data available. This work was supported by Datoric Labs.

Use of AI Assistants. This paper was prepared with the assistance of Anthropic's Claude (Opus / Sonnet / Haiku 4.5). Claude was used for three distinct purposes: (1) drafting and copyediting portions of the manuscript and generating Python code for the analysis, figure, and scoring pipelines; (2) generating culture-dependent comprehension QA items for each language (the initial pass) and their subsequently-recorded audio stems (the expanded axis); and (3) serving as a text-reasoner control condition over the reference transcript to establish a text-only upper-bound for the benchmark. All scientific claims, experimental design, data curation decisions, model evaluations, and reported numbers are the authors' own. Every number in this paper is computed from the released JSON result artifacts enumerated in §9, and the authors take full responsibility for the paper's content.

9 Reproducibility Note

ENOENT patch (240 rows reattributed, 2026-04-18). During post-curation auditing we identified that the FLEURS audio for Arabic, Portuguese, and Spanish materialized on disk at approximately 10:58 UTC on 2026-04-17, after Whisper large-v3 had started at 10:40 UTC but before Deepgram Nova-2 started at ∼11:13 UTC. Three early-starting model checkpoints (Whisper large-v3 on arb/por/spa; Deepgram Nova-3 on por/spa; Deepgram Nova-2 on arb only) cached `ENOENT` (file-not-found) errors that the resume logic treated as terminal rather than transient, producing 100%-empty rates on those (model, language) cells. We re-ran transcription on the affected 240 rows, re-merged into the master run JSON, and re-verified determinism end-to-end under the documented seed: 1689/1689 per-language metric values match between two independent re-evaluations (vs. 1647/1647 pre-patch; the +42 are newly-valid arb/por/spa per-language scores for Whisper large-v3 and Deepgram Nova-3). Post-patch, Whisper large-v3 covers Arabic / Portuguese / Spanish at 100% each (per-language WERs 0.179 / 0.257 / 0.205), and Deepgram Nova-3 likewise (0.293 / 0.282 / 0.223). Deepgram Nova-2's 0% rate on Arabic is *not* affected by the patch: it is a real provider-API rejection (Nova-2 does not list Arabic among its supported languages; HTTP 400).

Rate-limit backfill (1,224 rows reattributed, 2026-04-18 day-2; 9 further rows, 2026-04-20 day-3). On day 1 of the cultural-QA run, OpenAI TPM and Google RPD caps were saturated and GPT-4o Audio (482 error rows) and Gemini 2.5 Pro (792 error rows) received rate-limit responses on the majority of samples. After the 07:00 UTC daily quota reset on 2026-04-18 we re-issued the error rows using the identical `_call_openai_audio_native` / `_call_google_audio_native` code paths used in `run_models.py`, and merged the patched hypotheses into the master. After the day-2 pass, GPT-4o Audio covered 799/800 (one Javanese long-audio timeout) and Gemini 2.5 Pro covered 751/800 (93.9%); 43 rows had not been retried because the backfill stopped early at the next quota cap, leaving 9 transient Gemini 2.5 Pro residuals (six Vietnamese long-audio timeouts, two Igbo refusals, one Mandarin timeout) on top of the 40 Yoruba empties. A 2026-04-20 day-3 transient-retry pass targeted the 9 non-Yoruba residuals after the 04-19 UTC reset, using the same code path; all 9 recovered on retry. Post day-3, Gemini

2.5 Pro covers 760/800 (95.0%), with 40 residual failures all on Yoruba: the model produces empty responses on every Yoruba recording in our set, after the run, on retry, and at steady state. We document these as a genuine model-coverage gap rather than a benchmark artifact. The numbers in Tables 7–8 are post-patch, post-day-2-backfill, and post-day-3-backfill. The patch deltas — which 240 ENOENT-reattributed, 1,224 day-2 rate-limit-reattributed, and 9 day-3 transient-retry (sample, model) cells were rewritten and their pre- vs. post hypotheses — are preserved at `results/run_20260418_enoent_patch.json`, `results/run_20260418_ratelimit_patch.json`, and `results/run_20260419_gv_transient_retry_patch.json` for traceability. The independent re-runs underlying the 1689/1689 verification are at `results.run-a/` and `results.run-b/`.

References

Rosana Ardila, Megan Branber, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*.

Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T. Tan, and Haizhou Li. 2024. VoiceBench: Benchmarking LLM-based voice assistants. *arXiv preprint arXiv:2410.17196*.

Alexis Conneau, Min Ma, Simran Watanabe, Changhan Wang, and 1 others. 2023. FLEURS: Few-shot learning evaluation of universal representations of speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daume III, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Gemini Team, Google. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Injy Hamed and 1 others. 2022. Benchmarking code-switching ASR evaluation metrics. *arXiv preprint arXiv:2211.16319*.

Holy Lovenia, Samuel Cahyawijaya, Genta Indra Winata, and 1 others. 2022. ASCEND: A spontaneous chinese-english dataset for code-switching in multi-turn conversation. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*.

Dau-Cheng Lyu, Tien-Ping Tan, Eng Siong Chng, and Haizhou Li. 2015. Mandarin-english code-switching speech corpus in south-east asia: SEAME. *Language Resources and Evaluation*, 49(3):581–600.

OpenAI. 2024. GPT-4o system card.

Salsabila Zahirah Pranida, Rifo Ahmad Genadi, Muhammad Cendekia Airlangga, and Shady Shehata. 2025. ASR under noise: Exploring robustness for sundanese and javanese. In *Proceedings of the ACL Workshop on Widening NLP (WiNLP)*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Ann Williamson, Juan Pino, and Emmanuel Dupoux. 2021. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Yingzhi Wang, Anas Alhmoud, and Muhammad Alqurishi. 2025. Open universal arabic ASR leaderboard. In *Proceedings of Interspeech*.

Genta Indra Winata. 2025. Code-switching papers: A comprehensive collection. Curated list of code-switching research papers across ASR, NLU, and generation.

Peng Xie, Xingyuan Liu, Tsz Wai Chan, Yequan Bie, Yangqiu Song, Yang Wang, Hao Chen, and Kani Chen. 2025. SwitchLingua: The first large-scale multilingual and multi-ethnic code-switching dataset. *arXiv preprint arXiv:2506.00087*.

Brian Yan, Injy Hamed, Shuichiro Shimizu, Vasista Sai Lodagala, William Chen, Olga Iakovenko, Bashar Talafha, Amir Hussein, Alexander Polok, Kalvin Chang, Dominik Klement, Sara Althubaiti, Puyuan Peng, Matthew Wiesner, Thamar Solorio, Ahmed Ali, Sanjeev Khudanpur, and Shinji Watanabe. 2025. CS-FLEURS: A massively multilingual and code-switched speech dataset. In *Proceedings of Interspeech*.

Xueyi Zhang, Chengwei Zhang, Mingrui Lao, Peng Zhao, Jun Tang, Yanming Guo, Siqi Cai, Xianghu Yue, and Haizhou Li. 2024. Language without borders: A dataset and benchmark for code-switching lip reading. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*.

A Per-Language Results

Full per-language WER results for all 12 models on the 200-sample balanced tier-transcription subset are available in `results/run_20260414_080347_per_language.json`. Full per-language WER and ok-rate results for the $n = 800$ cultural-QA transcription axis are in `results/run_20260418_111344_per_language.json` (with `results/model_coverage_20260418.csv` as the wide view and `_long.csv` as the tidy view).

B Datasheet

Following Gebru et al. (2021):

Motivation. Created to measure and document the language equity gap in voice AI systems.

Composition. 800 audio samples across 20 languages for the per-tier transcription axis; an additional 150 ASCEND Mandarin–English code-switched samples (356 annotated switch points) for the code-switching axis; 745 Common Voice 17 samples across 7 languages $\times$ 22 accent cells for the accent-sensitivity axis; and 800 culturally-grounded audio recordings (40 per language $\times$ 20 languages) with reference transcripts for the cultural-QA transcription axis. All samples ship with reference transcriptions and language/tier/accent labels as applicable.

Collection. Per-tier transcription derived from FLEURS (Conneau et al., 2023), Common Voice (Ardila et al., 2020), and VoxPopuli (Wang et al., 2021). Code-switched speech drawn from ASCEND (Lovenia et al., 2022) and span-annotated in-house. Accent-labeled speech drawn from Common Voice 17 with the corpus's built-in regional accent tags. Cultural QA generated via Claude API (pilot scale, expanding).

Preprocessing. Per-tier transcription: balanced to 40 samples per language, audio resampled to 16 kHz mono. Code-switching: ASCEND samples used as released with span-level language annotations added. Accent: CV17 samples filtered to accent cells with $\geq$15 samples.

Uses. Intended for evaluating multilingual voice AI systems. Not intended for training or fine-tuning.

Distribution. Released under CC-BY-4.0 for metadata/annotations. Audio files retain their original source licenses.

Maintenance. The $n = 800$ cultural-QA expansion is included in this release.

C Accent Cells

Per-cell sample sizes for the accent-sensitivity axis. Cells reflect Common Voice 17's declared regional-accent labels; cells below 15 samples were dropped from evaluation.

Language	Accent	n
Arabic (arb)	MSA	40
	Egyptian	16
German (deu)	Germany	40
	Austria	40
	Switzerland	33
English (eng)	US	40
	England	40
	Canada	35
	Australia	33
French (fra)	France	40
	Canada (fra-CA)	40
	Belgium	40
	Switzerland (fr)	18
Italian (ita)	Northern	27
	Southern	27
	Central	20
Portuguese (por)	Brazil	40
	Portugal	16
Spanish (spa)	Mexico	40
	Andean	40
	Spain-North	40
	Caribbean	40
Total		**745**

D Replication

All numbers in this paper are computed from the files in `results/`: `run_20260414_080347.json` (raw per-tier responses), `run_20260414_080347_per_language.json` (tier and per-language breakdown), `run_cs_20260417_004319.json` (code-switching raw responses) and `run_cs_20260417_004319_code_switch.json` (boundary-WER + LID aggregate), `run_acc_20260417_170317.json` (accent raw responses) and `run_acc_20260417_170317_accent_variance.json` (per-language across-accent variance), `run_20260411_222454_cultural_qa.json` (cultural-QA pilot aggregates).

Running `python eval/score_per_language.py` on the tier run, `python eval/score_code_switch.py` on the CS run, and `python eval/score_accent_variance.py` on the

accent run reproduces every number in Tables 4, 5, and 6.