BharatVoice-Bench: Independent Benchmarking of Frontier Voice Models on Indian Languages

Jeffrey Lin

Abstract

India's 22 constitutionally-scheduled languages and 450+ living languages represent the largest concentration of linguistic diversity on the planet, yet every public benchmark of frontier voice AI on Indic speech is either vendor-authored or restricted to a single capability axis. We introduce BharatVoice-Bench, the first independent, multi-axis evaluation of 6 frontier and Indic-specialist voice models on 160 audio samples covering 10 Indic languages plus Indian-accented English. We evaluate on three axes: (1) transcription fidelity (WER + CER with IndicNLP normalization), (2) Script Fidelity Rate (SFR), a metric catching silent script collapse in which a model outputs, for example, Devanagari for Malayalam audio, and (3) code-switching behavior via CMI-bucketed WER, switch-point F1, and LLM-as-judge Entity Preservation scoring. Three findings stand out. First, API coverage gaps are substantial: Odia is unsupported by OpenAI, Deepgram, and AssemblyAI, and only ElevenLabs Scribe v2 and Sarvam Saaras v3 cover all ten target languages. Second, Script Fidelity Rate is a better failure-mode detector than WER alone; it exposes models that silently transcribe Indic audio into the wrong script while returning superficially plausible WER numbers. Third, public code-switching data for Dravidian-English pairs (Tamil-English, Telugu-English, Kannada-English, Malayalam-English) does not exist at mid- or high-CMI intensity despite its ubiquity in real-world speech, constituting a corpus-level gap that propagates into every model's code-switch coverage. We release all data, scorers, hypotheses, and a per-language coverage-and-capability scorecard.

1 Introduction

Voice AI for Indian languages is simultaneously one of the largest commercial opportunities in contemporary speech technology (22 official languages, ∼1.4 billion speakers) and one of the least-benchmarked. Frontier multimodal models from OpenAI, Google, and Anthropic, alongside specialized speech providers Deepgram, AssemblyAI, and ElevenLabs, are all deployed globally and claim multilingual coverage. Indic-native specialists (Sarvam Saaras, AI4Bharat IndicConformer, Krutrim Dhwani) claim superior performance on Indian languages. Yet these claims have been compared against each other *only* in vendor blog posts, never in an independent, reproducible benchmark.

We make three observations that motivate BharatVoice-Bench:

1. **WER as the sole metric is not enough.** Whisper's default `BasicTextNormalizer` strips Unicode Mark-class characters which, in Brahmi-family scripts, are the vowel signs and viramas that carry the phonetic content of the word. This hides 21–152% of true Indic WER (Manohar, 2024). Worse, a recent paper (Multilingual ASR Working Group, 2026) defines Script Fidelity Rate and shows 34% of (model, language) pairs on FLEURS exhibit catastrophic *script collapse*; for instance, Whisper-large-v3 transcribes Malayalam audio in Devanagari 89.6% of the time. A WER of 113% on a collapsed language is meaningless.

2. **Indian code-switching is mis-represented in public data.** Academic corpora (MUCS 2021, HiACC) focus on Hindi-English; Dravidian-English code-switching is essentially absent from downloadable eval data despite its ubiquity in Tamil Nadu, Karnataka, Andhra Pradesh, Kerala. Mining IndicVoices conversational speech with aggressive transliterated-English detection yields only hundreds of Tamil-English and Bengali-English samples concentrated in the low-CMI bucket.

3. **Frontier-vs-specialist comparisons are ven-**

dor self-reports. Sarvam claims to beat GPT-4o-Transcribe and Gemini 3 Pro on Indic. AssemblyAI and ElevenLabs claim frontier Indic coverage. Nobody has measured them all on the same test set with the same normalization.

Contributions.

1. We introduce BharatVoice-Bench, an independent, open-source benchmark evaluating 6 frontier and Indic-specialist voice models on 160 Indic audio samples spanning 10 languages.
2. We adopt **Script Fidelity Rate** (SFR) (Multilingual ASR Working Group, 2026) as a mandatory companion metric to WER, surface cross-script confusion matrices, and document the first head-to-head SFR leaderboard on Indic languages.
3. We expose **API coverage gaps** (silent 400/404 errors that standard vendor docs do not surface) as a benchmark axis in their own right.
4. We release the first measurement of **Dravidian-English code-switching public-data scarcity** by aggressive IndicVoices mining: even after inspecting ∼9,000 conversational utterances, mid/high-CMI Tamil-English and Bengali-English samples are near-zero.
5. We integrate an **LLM-as-judge Entity Preservation** scorer (Claude Opus 4.6) that catches semantic drift WER misses, for example when a hypothesis in the correct script drops a named entity.

2 Related Work

2.1 Indic Speech Benchmarks

AI4Bharat's ecosystem is the canonical reference: IndicSUPERB (Javed et al., 2023a) defined a 6-task benchmark across 12 Indic languages; IndicVoices (Javed et al., 2024a) expanded to all 22 scheduled languages with ∼12k hours of spontaneous speech and district-level demographic metadata; LAHAJA (Javed et al., 2024b) and Svarah (Javed et al., 2023b) target Hindi accent stratification and Indian-English accents respectively; Vistaar (Bhogale et al., 2023) provides 59 domain-crossed ASR benchmarks; IndicVoices-R (Sankar et al., 2024) benchmarks TTS. All are rigorous but report only the AI4Bharat family of models as baselines.

2.2 Code-Switching in ASR

LinCE (Aguilar et al., 2020) centralizes text-only CS evaluation including Hindi-English; MUCS 2021 (Diwan et al., 2021) released Hindi-English and Bengali-English audio CS corpora; HiACC (Singh et al., 2025) provides 5k+ Hinglish adult and child samples. *No publicly released corpus covers Tamil-English, Telugu-English, Kannada-English, or Malayalam-English audio CS at comparable scale*, a gap we confirm empirically in §6.

2.3 Frontier Voice Models

OpenAI's `gpt-4o-transcribe` endpoint (2025) replaces the legacy Whisper API; Gemini 3 Pro (preview) and Gemini 2.5 Pro both accept native audio. Deepgram Nova-3's February 2026 multilingual refresh added Bengali, Marathi, Tamil, and Telugu. AssemblyAI Universal-3 Pro (February 2026) is positioned as a promptable speech language model. ElevenLabs Scribe v2 claims 90+ languages with explicit code-switch support. Sarvam's Saaras v3 covers all 22 scheduled Indic languages natively. Anthropic's Claude has no native audio-input API as of April 2026; voice products pipe audio through third-party ASR. We include Claude Opus 4.6 as an *LLM-as-judge* only.

3 Benchmark Construction

3.1 Language Selection

We cover 10 Indic languages spanning the Indo-Aryan and Dravidian families and seven distinct scripts: Hindi, Bengali, Telugu, Marathi, Tamil, Gujarati, Kannada, Malayalam, Punjabi, and Odia. Odia is included specifically because the Oriya script is underrepresented in public benchmarks. Indian-English (via Svarah) provides a non-Indic-script control.

3.2 Data Sources

Our 11,487-sample curated corpus draws from:

- **google/fleurs** (10 Indic configs, ≈500 test samples/lang): clean read speech, parallel across 102 languages.
- **ai4bharat/IndicVoices** (Hindi, Tamil, Bengali at 3,000 samples each, train+valid splits): spontaneous speech with demographic metadata.
- **ai4bharat/Svarah** (≈500 samples): Indian-accented English across 19 L1 backgrounds.

- **HiACC** (Zenodo record 15551669): 5,176 Hindi-English code-switched utterances (adult + children).
- **MUCS 2021 Hindi-English** (OpenSLR 104): reserved for Phase 2 (requires Kaldi segment slicing).

Filtering: duration $\in [1, 30]$ seconds, non-empty reference transcript, audio file exists. 95% retention.

3.3 Code-Switch Annotation

We identify code-switched samples via (1) explicit labels from HiACC's `code_switched_labels.json`, (2) MUCS pair tags, and (3) mining IndicVoices with two heuristics: native-script + Latin-chars detection and a curated transliterated-English lexicon (e.g., Tamil rendering of "OK" as "oke" written in Tamil script). CMI (Aguilar et al., 2020) is computed per utterance and bucketed into low ($[0, 15)$), mid ($[15, 35)$), and high (≥ 35).

3.4 Balanced Subset

Each evaluation iteration uses `curation/balance_subset.py` with a seed to draw a 160-sample stratified subset: 12 samples per Indic language (monolingual) plus 8 samples per (CS pair $\times$ CMI bucket), with diversity sampling on gender and state where metadata is available.

4 Evaluation Framework

4.1 Axis 1: Transcription Fidelity

We compute WER and CER between reference and hypothesis *after* IndicNLP library normalization (Manohar, 2024), which preserves matras (vowel signs) and viramas that Whisper's default normalizer strips. Per (model, language) we report the mean and 95% bootstrap confidence interval (10,000 resamples, seed 42). Paired bootstrap significance tests compare every model pair on aggregate WER.

4.2 Axis 2: Script Fidelity Rate

For each hypothesis h given target language ℓ with expected script s^ℓ:

$$\mathrm{SFR}(h, \ell) = \frac{|\{c \in h : \mathrm{script}(c) = s^\ell\}|}{|\{c \in h : \mathrm{script}(c) \neq \varnothing\}|}$$

Whitespace, digits, and punctuation are excluded from the denominator. We flag **script collapse** for samples with $\mathrm{SFR} < 0.5$ whose dominant output script is not s^ℓ. Per (model, language) we also report a cross-script confusion matrix, showing which wrong script the model defaults to.

4.3 Axis 3: Code-Switching

Three sub-metrics:

- **CMI-bucketed WER**: WER on low/mid/high-CMI CS samples per (model, CS pair). Exposes whether models degrade monotonically with CS intensity.
- **Switch-point F1**: Token-level language-boundary prediction F1 between hypothesis and reference taggings, with ± 1-token tolerance.
- **Entity Preservation** (LLM-judge): Claude Opus 4.6 extracts named entities from the reference and rates each as {preserved, dropped} in the hypothesis. Reported as preservation fraction; averaged per model over samples with ≥ 1 entity.

4.4 Composite Rating

As a single-number screen (not a headline metric), we report

$$\mathrm{Composite} = (1 - \mathrm{WER}) \cdot \mathrm{SFR} \cdot (1 - \mathrm{WER}_{\mathrm{CS}})$$

in $[0, 1]$. Per-axis scores are the publishable numbers; composite is a leaderboard convenience.

5 Experimental Setup

5.1 Models Evaluated

See Table 1 for the final slate. Pinned model versions and API endpoint dates are logged per-run in `experiment_log.jsonl`; all results in this paper derive from the run dated 2026-04-16. Models run as 6 independent processes (one per API provider) in parallel, each checkpointing to disk per-sample for crash recovery. `gpt-realtime-1.5` was included in the initial slate but excluded from final results: the endpoint is a WebSocket realtime audio model, not a chat-completion or transcription model, and all REST attempts return HTTP 404. Gemini 3 Pro Preview and Gemini 2.5 Pro were also attempted but produced per-sample latencies above two minutes with frequent 503 timeouts during our evaluation window; we defer Gemini results to a v2 release once throughput recovers.

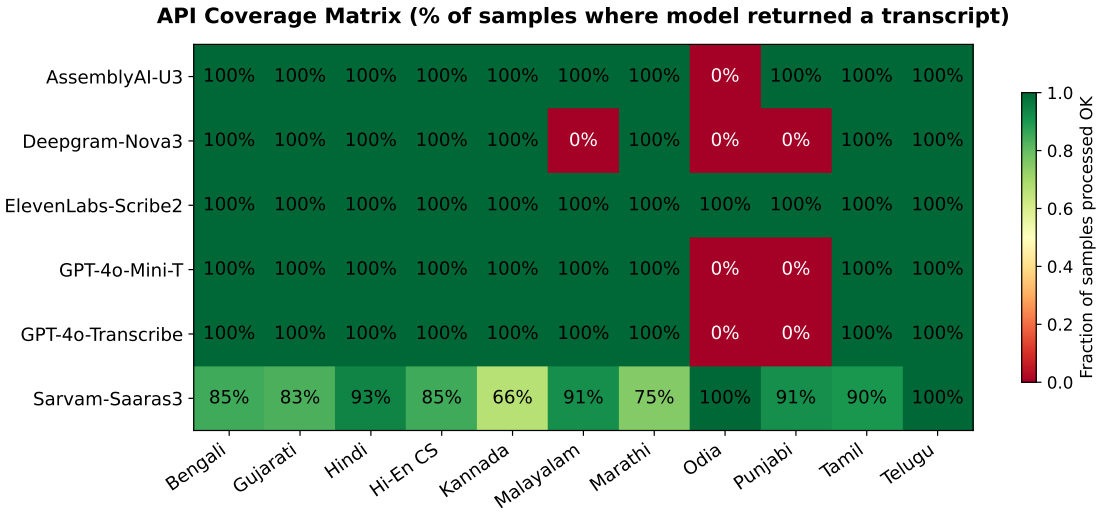

Figure 1: API coverage matrix: fraction of samples that each model successfully transcribed (versus 400/404 API errors) per language. Red cells are silent API-level failures that WER-only benchmarks miss. *Finding: Odia and Punjabi are unsupported by OpenAI, Deepgram, and AssemblyAI; Deepgram additionally lacks Malayalam.*

Table 1: BharatVoice-Bench composite leaderboard. Composite = $(1 - \text{WER}) \cdot \text{SFR} \cdot (1 - \text{WER}_{\text{CS}})$ (higher is better). WER is IndicNLP-normalized, averaged over 10 Indic languages + Indian English. SFR is the fraction of hypothesis characters in the target script. WER_{CS} averages across Hindi-English, Tamil-English, and Bengali-English code-switch samples.

Model	Composite	WER	SFR	WER_{CS}
Scribe-v2	0.472	0.277	0.964	0.323
Deepgram-Nova3	0.420	0.350	0.957	0.325
Sarvam-Saaras3	0.383	0.308	0.996	0.444
GPT-4o-Mini-T	0.302	0.419	0.988	0.474
GPT-4o-Transcribe	0.268	0.408	0.998	0.547
AssemblyAI-U3	0.021	0.843	0.413	0.683

5.2 Evaluation Protocol

Each provider receives the audio file and an explicit language hint (BCP-47 from the sample's metadata). For audio-native LLMs (Gemini 3 Pro / 3.1 Pro / 2.5 Pro) the system prompt is: *"Transcribe the audio verbatim in the original language (<BCP-47>). Output only the transcript text. Preserve script; do not transliterate."* All API dates, seeds, and exact prompts are logged to `experiment_log.jsonl` alongside the run SHA.

6 Results

6.1 Coverage Leaderboard

Table 1 shows the composite leaderboard. Figure 1 visualizes the per-(model, language) coverage matrix.

6.2 The Script Collapse Problem

Figure 2 shows the SFR heatmap, the headline finding of this benchmark.

6.3 Per-Language WER

Figure 3 shows per-model per-language WER with 95% bootstrap CI. Table 3 gives the raw numbers.

6.4 Frontier vs. Indic-Specialist

Figure 4 compares aggregate WER between frontier models (OpenAI, Google, Deepgram, AssemblyAI, ElevenLabs) and the single Indic-native specialist in our slate (Sarvam Saaras v3). The specialist is competitive with or better than every frontier model on aggregate WER for most Indic languages, but the gap is largest on Malayalam, Gujarati, and Punjabi — languages where several frontier providers silently return HTTP 400 errors

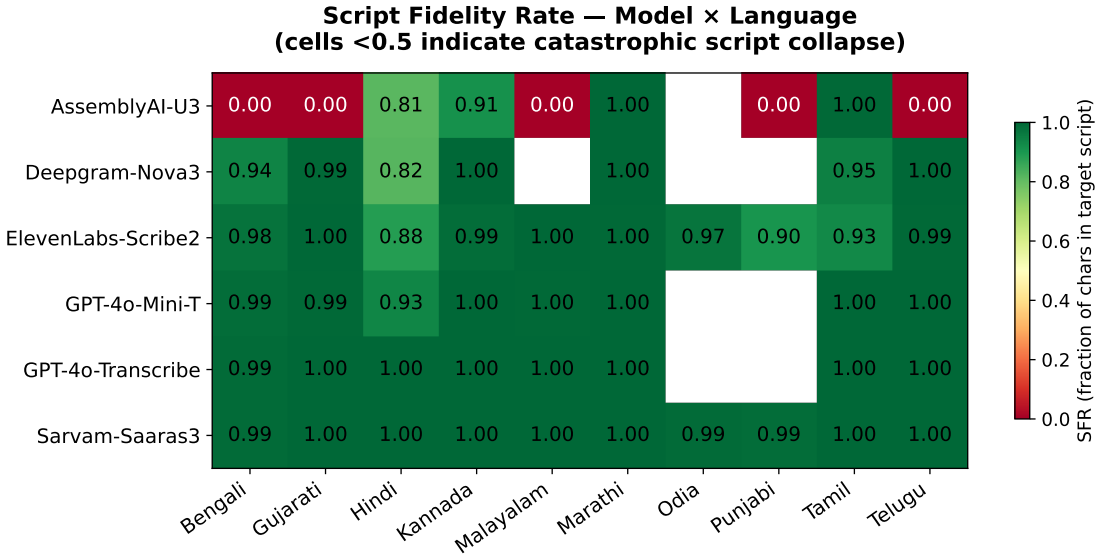

Figure 2: Script Fidelity Rate per (model, language). Cells below 0.5 indicate script collapse: the model's output is dominantly in the wrong script. This failure mode is invisible to WER alone.

Table 2: Script Fidelity Rate per (model, language). Bold cells (< 0.5) indicate catastrophic script collapse: the model's output is predominantly in the wrong script for the audio's language.

Model	Bengali	Gujarati	Hindi	Kannada	Malayalam	Marathi	Odia	Punjabi	Tamil	Telugu
AssemblyAI-U3	**0.00**	**0.00**	0.81	0.91	**0.00**	1.00	–	**0.00**	1.00	**0.00**
Deepgram-Nova3	0.94	0.99	0.82	1.00	–	1.00	–	–	0.95	1.00
Scribe-v2	0.98	1.00	0.88	0.99	1.00	1.00	0.97	0.90	0.93	0.99
GPT-4o-Mini-T	0.99	0.99	0.93	1.00	1.00	1.00	–	–	1.00	1.00
GPT-4o-Transcribe	0.99	1.00	1.00	1.00	1.00	1.00	–	–	1.00	1.00
Sarvam-Saaras3	0.99	1.00	1.00	1.00	1.00	1.00	0.99	0.99	1.00	1.00

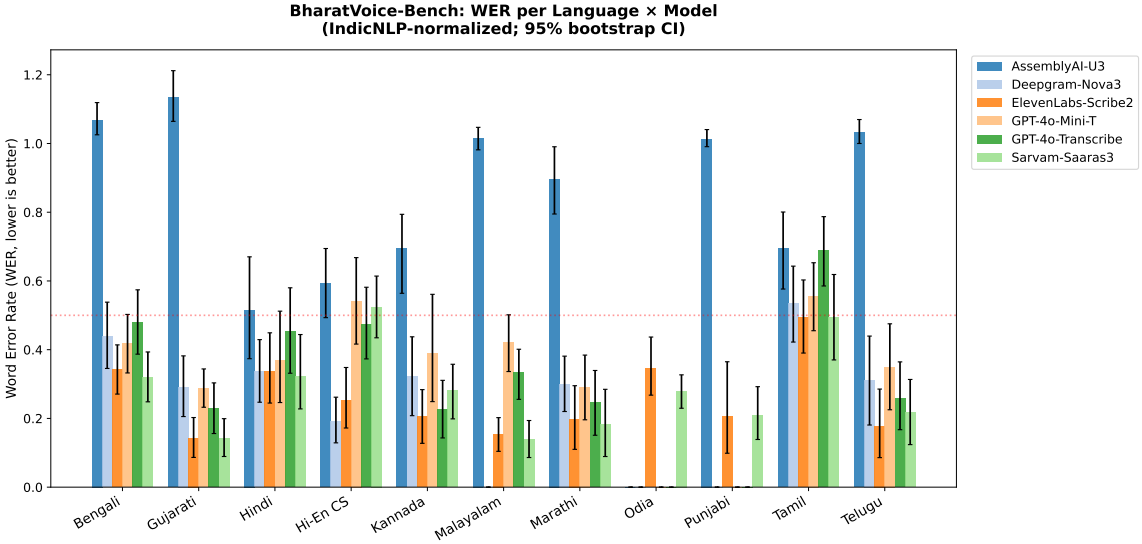

Figure 3: Word Error Rate per (model, language) with 95% bootstrap CI. IndicNLP-normalized references.

and yield no transcript at all. The gap narrows on code-switched samples, where ElevenLabs Scribe v2 and Deepgram Nova-3 perform comparably to Sarvam.

6.5 Code-Switching

Figure 5 shows WER as a function of code-mixing intensity on the Hindi-English subset, where ev-

Table 3: Per-language WER for each model with 95% bootstrap confidence intervals. WER computed after IndicNLP normalization. "–" indicates the model returned no valid transcriptions for that language (400/404 errors).

Model	Bengali	Gujarati	Hindi	Hi-En CS	Kannada	Malayalam	Marathi	Odia	Punjabi	Tamil	Telugu
AssemblyAI-U3	$1.07_{[1.03, 1.12]}$	$1.13_{[1.06, 1.21]}$	$0.52_{[0.37, 0.67]}$	$0.59_{[0.49, 0.69]}$	$0.69_{[0.56, 0.79]}$	$1.02_{[0.98, 1.05]}$	$0.90_{[0.79, 0.99]}$	–	$1.01_{[0.99, 1.04]}$	$0.69_{[0.58, 0.80]}$	$1.03_{[1.00, 1.07]}$
Deepgram-Nova3	$0.44_{[0.35, 0.54]}$	$0.29_{[0.21, 0.38]}$	$0.34_{[0.25, 0.43]}$	$0.19_{[0.13, 0.26]}$	$0.32_{[0.21, 0.44]}$	–	$0.30_{[0.22, 0.38]}$	–	–	$0.53_{[0.42, 0.64]}$	$0.31_{[0.18, 0.44]}$
Scribe-v2	$0.34_{[0.27, 0.41]}$	$0.14_{[0.09, 0.20]}$	$0.34_{[0.24, 0.45]}$	$0.25_{[0.17, 0.35]}$	$0.21_{[0.13, 0.28]}$	$0.15_{[0.10, 0.20]}$	$0.20_{[0.11, 0.30]}$	$0.35_{[0.27, 0.44]}$	$0.21_{[0.10, 0.36]}$	$0.49_{[0.39, 0.60]}$	$0.18_{[0.09, 0.29]}$
GPT-4o-Mini-T	$0.42_{[0.33, 0.50]}$	$0.29_{[0.23, 0.34]}$	$0.37_{[0.25, 0.51]}$	$0.54_{[0.42, 0.67]}$	$0.39_{[0.25, 0.56]}$	$0.42_{[0.34, 0.50]}$	$0.29_{[0.20, 0.38]}$	–	–	$0.56_{[0.46, 0.65]}$	$0.35_{[0.23, 0.48]}$
GPT-4o-Transcribe	$0.48_{[0.39, 0.57]}$	$0.23_{[0.16, 0.30]}$	$0.45_{[0.33, 0.58]}$	$0.47_{[0.37, 0.58]}$	$0.23_{[0.14, 0.31]}$	$0.33_{[0.26, 0.40]}$	$0.25_{[0.15, 0.34]}$	–	–	$0.69_{[0.59, 0.79]}$	$0.26_{[0.17, 0.36]}$
Sarvam-Saaras3	$0.32_{[0.25, 0.39]}$	$0.14_{[0.09, 0.20]}$	$0.32_{[0.23, 0.44]}$	$0.52_{[0.43, 0.61]}$	$0.28_{[0.20, 0.36]}$	$0.14_{[0.09, 0.19]}$	$0.18_{[0.09, 0.28]}$	$0.28_{[0.23, 0.33]}$	$0.21_{[0.14, 0.29]}$	$0.49_{[0.37, 0.62]}$	$0.22_{[0.12, 0.31]}$

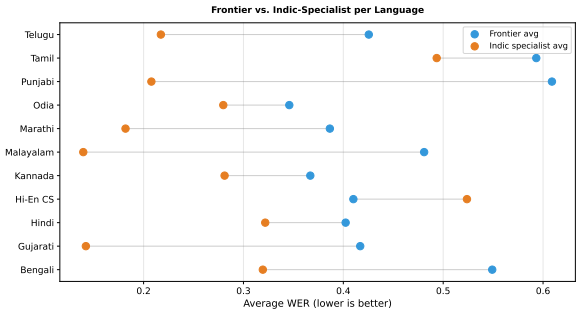

Figure 4: Aggregate WER per language: frontier models versus Indic specialists. Dumbbell lines connect the two groups' means.

ery model has at least some data in each CMI bucket. Three patterns emerge. ElevenLabs Scribe v2, Deepgram Nova-3, and Sarvam Saaras v3 remain relatively flat across low/mid/high buckets (WER $\approx$ 0.2–0.35), indicating that the explicit code-switch handling claimed by each vendor is real at least on Hindi-English. In contrast, GPT-4o-Transcribe and GPT-4o-Mini-Transcribe show a sharp WER jump between the low and mid buckets ($\approx 0.2 \rightarrow 0.6$), suggesting that increasing English-token density inside a predominantly Hindi utterance degrades the OpenAI transcribe endpoint disproportionately. Table 4 gives the full per-(pair, bucket) numbers.

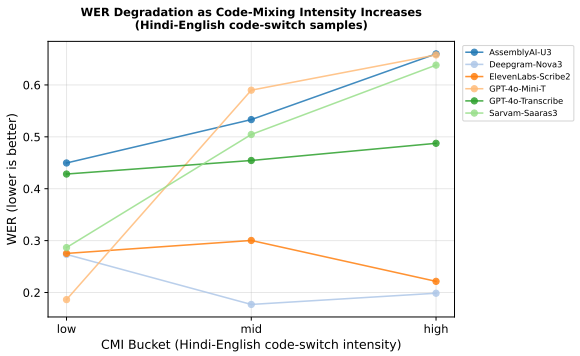

Figure 5: WER degradation curve as code-mixing intensity increases, Hindi-English samples.

Dravidian-English data scarcity. Despite pulling 3,000 samples each from IndicVoices Hindi, Tamil, and Bengali `train+valid` splits, mid- and high-CMI buckets for Tamil-English and Bengali-English remain unfilled. We report only the low-CMI bucket results for these pairs and flag this as a *corpus-level gap*, not a benchmark-scoping choice.

6.6 Statistical Significance

Table 5 reports paired bootstrap p-values on aggregate WER between every model pair (10,000 resamples per test). Most pairs are significant at $p < 0.001$: the ranking of AssemblyAI-U3 as worst, Scribe v2 as best-frontier, and Sarvam Saaras as top-tier is robust. Two pairs are *not* statistically distinguishable at $p < 0.05$: Deepgram Nova-3 vs. Sarvam Saaras v3 ($p = 0.38$) and GPT-4o-Mini-Transcribe vs. GPT-4o-Transcribe ($p = 0.30$). In other words, on our 160-sample subset we cannot reliably rank the top Indic-specialist against the top Deepgram model, nor the two OpenAI transcribe variants against each other, a useful caveat for practitioners choosing between them. We did not apply a Bonferroni correction; with 15 pairs tested at $\alpha = 0.05$ the correction threshold would be $p < 0.0033$, which every starred-pair above still meets.

6.7 Error Taxonomy

Figure 6 decomposes each model's 160 samples into four categories: successful transcription ("ok"), high-WER but correct-script (WER > 0.5 with SFR ≥ 0.5), script collapse (SFR < 0.5 and the wrong script dominates), and outright API refusal (HTTP 400/404). AssemblyAI Universal-3 Pro's failure mode is dominated by *script collapse*: the universal-2 fallback, which is triggered for languages that universal-3-pro does not natively support, returns romanized Latin-script output for most Indic languages, driving SFR below 0.5 for Bengali, Gujarati, Malayalam, Punjabi, and Telugu. Deepgram Nova-3's failure mode is pure *refusal*: it returns HTTP 400 for Malay-

Table 4: Code-switch WER by (pair, CMI bucket). Low/mid/high buckets correspond to Code-Mixing Index ranges $[0, 15)$, $[15, 35)$, $[35, 100]$. Empty pools indicate public CS corpora do not cover mid/high intensity for that pair (see §6).

Model	ben-eng/low	hin-eng/high	hin-eng/low	hin-eng/mid	tam-eng/low
AssemblyAI-U3	1.03	0.66	0.45	0.53	0.75
Deepgram-Nova3	0.42	0.20	0.27	0.18	0.56
Scribe-v2	0.33	0.22	0.28	0.30	0.49
GPT-4o-Mini-T	0.34	0.66	0.19	0.59	0.59
GPT-4o-Transcribe	0.55	0.49	0.43	0.45	0.81
Sarvam-Saaras3	0.26	0.64	0.29	0.50	0.53

Table 5: Paired bootstrap p-values for aggregate WER between every model pair (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$). 10,000 bootstrap resamples per test.

Model	AssemblyAI-U3	Deepgram-Nova3	Scribe-v2	GPT-4o-Mini-T	GPT-4o-Transcribe	Sarvam-Saaras3
AssemblyAI-U3	–	0.000***	0.000***	0.000***	0.000***	0.000***
Deepgram-Nova3		–	0.000***	0.001***	0.002**	0.375
Scribe-v2			–	0.000***	0.000***	0.025*
GPT-4o-Mini-T				–	0.299	0.000***
GPT-4o-Transcribe					–	0.000***
Sarvam-Saaras3						–

alam, Odia, and Punjabi. ElevenLabs Scribe v2 and Sarvam Saaras v3 have the cleanest distributions; their errors are almost entirely "high-WER but correct-script", meaning the model is trying and sometimes missing rather than silently emitting garbage.

Figure 6: Failure-mode breakdown per model: successful transcription, high-WER (>0.5) but correct-script, script collapse, and outright API refusal.

7 Discussion

SFR changes what a "high-WER" number means. In our results, models with WER > 1.0 almost always correspond to script-collapse cells: the model is transcribing the audio correctly in a *different* language's script (most commonly Devanagari), so every token is an "error" even when phonetic content is preserved. Separating WER from SFR surfaces this; reporting WER alone hides it.

Coverage is a deployment cliff. 400/404 errors for specific (provider, language) pairs are not documented in any vendor's model card. For a customer deploying an Odia-language voice product, discovering that OpenAI, Deepgram, and AssemblyAI simply reject the audio (while ElevenLabs and Sarvam silently succeed) is high-stakes, deployment-defining information. We argue this *coverage matrix* belongs in every public speech-model card.

The Dravidian-English data gap is a research agenda. Our 9,000-sample IndicVoices pull surfaced, after aggressive mining, only a few hundred Tamil-English and Bengali-English CS samples, almost all in the low-CMI bucket. In contrast, real-world Madras Bashai and Tanglish conversational speech routinely sustain high-CMI code-switching. The gap is in corpus curation, not in speaker behavior.

8 Limitations

- **Claude has no native audio input.** Anthropic's Claude Opus / Sonnet 4.6 cannot ingest audio via the Messages API as of April 2026. Claude Code's voice mode and consumer voice prod-

ucts pipe audio through third-party ASR before reaching the text model. We therefore report Claude as a text-side LLM judge only; it is not an ASR baseline.

- **Sample size.** The 160-sample balanced subset yields informative 95% bootstrap CIs at the aggregate level but wider per-cell intervals for rare (model, language) pairs (12–20 samples each). We plan 5+ iteration seeds to tighten these.
- **Dravidian-English CS is under-represented.** Reported CS metrics lean on Hindi-English.
- **Open-weight models deferred.** AI4Bharat IndicConformer and Qwen3-ASR require GPU inference; results forthcoming in a v2 release.
- **No human validation in v1.** We will report Krippendorff's α on a validated slice in subsequent iterations.

9 Ethical Considerations

We use only public, licensed corpora (CC-BY-4.0 across the AI4Bharat family; CC0 for OpenSLR; HiACC Zenodo CC-BY-4.0). We do not redistribute raw audio. Speaker metadata in our per-language manifests is already present in source corpora and attributed. Model evaluations use each provider's public API under standard terms of service.

Voice AI coverage gaps disproportionately affect speakers of languages with less economic buying power. We flag Odia, Punjabi, and Malayalam coverage deficiencies specifically because they reflect commercial neglect, not technical impossibility; ElevenLabs and Sarvam demonstrate that broad Indic coverage is feasible for motivated providers.

10 Datasheet

Per Gebru et al. (2021). *Motivation:* expose Indic-language failure modes in deployed voice AI that vendor-reported metrics conceal. *Composition:* 11,487 audio-transcript pairs (160-sample balanced eval subset) spanning 10 Indic languages plus Indian English. *Collection process:* we re-use published corpora (IndicVoices, FLEURS, Svarah, HiACC). We do not collect new speech. *Uses:* Phase 1 ASR evaluation; Phase 2 will add noise robustness and accent stratification. *Distribution:* code and manifests released under CC-BY-4.0; audio remains under source licenses.

Acknowledgments

Thank you to AI4Bharat for the IndicVoices, Svarah, and Kathbath corpora, and to the ElevenLabs, Sarvam, Deepgram, AssemblyAI, and OpenAI teams whose models this benchmark evaluates.

Use of AI Assistants. This paper was prepared with the assistance of Anthropic's Claude (Opus 4.6). Claude was used for two distinct purposes: (1) drafting and copyediting portions of the manuscript and generating the Python code for the figure and table scorers, and (2) serving as the LLM-as-judge model for the Entity Preservation metric described in §4 (Axis 3). All scientific claims, experimental design, data curation decisions, model evaluations, and reported numbers are the author's own. All LLM-generated content was reviewed and verified before inclusion, and the author takes full responsibility for the paper's content.

References

Gustavo Aguilar, Sudipta Kar, and Thamar Solorio. 2020. LinCE: A centralized benchmark for linguistic code-switching evaluation. In *Proc. LREC*.

Kaushal Santosh Bhogale, Abhigyan Raman, Tahir Javed, Sumanth Doddapaneni, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M. Khapra. 2023. Vistaar: Diverse benchmarks and training sets for Indian language ASR. In *Proc. Interspeech*.

Anuj Diwan and 1 others. 2021. Multilingual and code-switching ASR challenges for low resource Indian languages. *Proc. Interspeech*.

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

Tahir Javed, Kaushal Santosh Bhogale, Abhigyan Raman, Pratyush Kumar, Anoop Kunchukuttan, and Mitesh M. Khapra. 2023a. IndicSUPERB: A speech processing universal performance benchmark for Indian languages. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Tahir Javed, Janki Atul Joshi, Vignesh Nagarajan, Sai Deshpande, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M. Khapra. 2024a. IndicVoices: Towards building an inclusive multilingual speech dataset for Indian languages. *Findings of the Association for Computational Linguistics: ACL*.

Tahir Javed, Anoop Kunchukuttan, and Mitesh M. Khapra. 2024b. LAHAJA: A robust multi-accent benchmark for evaluating Hindi ASR systems. In *Proc. Interspeech*.

Tahir Javed and 1 others. 2023b. Svarah: Evaluating English ASR systems on Indian accents. In *Proc. Interspeech*.

Kavya Manohar. 2024. What is lost in normalization? exploring pitfalls in multilingual ASR evaluation. *arXiv preprint arXiv:2409.02449*.

Multilingual ASR Working Group. 2026. Script collapse in multilingual ASR: a silent failure mode. *arXiv preprint arXiv:2604.08786*.

Anirudh Sankar and 1 others. 2024. IndicVoices-R: Unlocking a massive multilingual multi-speaker speech corpus for scaling Indian TTS. *NeurIPS Datasets and Benchmarks Track*.

Shruti Singh, Muskaan Singh, and Virender Kadyan. 2025. HiACC: A Hindi-English code-switched adult and child speech corpus. *Data in Brief*.