HORN-ASR: A SPONTANEOUS-SPEECH EVALUATION BENCHMARK FOR LANGUAGES OF THE HORN OF AFRICA

Anonymous authors
Paper under double-blind review

ABSTRACT

Evaluation datasets for multilingual automatic speech recognition (ASR) on African languages are constructed in ways that diverge from the spontaneous speech these systems will transcribe in use: translated sentences read aloud by fewer than ten speakers per language, held-out splits of single-speaker Bible recordings, or image-prompted elicitation. FLEURS, the most widely used of these benchmarks, covers three languages of the Horn of Africa. For each of them, our audit finds that every split contains only one gender, 3 to 7 speakers per language, and every development speaker also in the test split. We introduce HORN-ASR, an evaluation benchmark of *naturally-occurring* spontaneous speech (broadcasts and interviews recorded for a public audience) for more realistic evaluation of Amharic, Oromo, Somali, and Tigrinya, languages spoken by more than 100 million people. It consists of 1,000 curated segments per language (15.43 h in total) from 1,710 publicly available interviews, balanced for gender and diverse across dialect and topical domain. We evaluate six multilingual ASR systems from three providers under a single evaluation protocol. HORN-ASR is harder than FLEURS and far from saturated: the best word error rate (WER) on any language is 0.275, and where a system is scored on the same language in both benchmarks, its WER is on average 0.16 higher on HORN-ASR. It supports disaggregated evaluation by gender and dialect, which single-gender read-speech datasets cannot, and reveals significant gender gaps in every language as well as dialect gaps of up to 0.22 WER that favor the dominant variety. An inter-annotator agreement study measures the reliability of the reference transcripts and labels. We release the transcripts, labels, and guidelines (CC BY-SA 4.0), the audio (under terms of use), the curation methodology, and all system outputs and scoring code.

1 INTRODUCTION

Evaluation datasets are how we measure progress in automatic speech recognition (ASR) for a language: they shape what gets optimized and show where systems still fail. For African languages, two kinds dominate. Multilingual systems claiming coverage of hundreds of languages (Whisper (Radford et al., 2023), MMS (Pratap et al., 2024), USM (Zhang et al., 2023), SeamlessM4T v2 (Seamless Communication Team, 2023), and most recently Omnilingual ASR (Omnilingual ASR Team et al., 2025)) report results on FLEURS (Conneau et al., 2023), where fewer than ten speakers per language read aloud sentences translated from English Wikipedia. The claims of coverage for thousands of languages, in MMS and Omnilingual ASR, rest instead on held-out splits of the same single-speaker Bible-recording corpora used for training. Neither resembles how these languages are spoken in use: read translationese narrows the speaker pool and the vocabulary, and held-out splits from the training source measure consistency within one domain. When most evaluation is read speech, architectures, decoding choices, and the threshold for "good enough" are tuned to that register, and aggregate word error rate (WER) on a single-gender split cannot show which speakers a system fails.

Amharic, Oromo, Somali, and Tigrinya, four Afroasiatic languages of the Horn of Africa, have a combined first- and second-language speaker population above 100 million (Eberhard et al., 2024). Even FLEURS coverage is limited: it covers three of the four (Tigrinya is absent) with fewer than ten readers each; Common Voice (Ardila et al., 2020) adds small amounts of Amharic and Somali

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

read aloud by volunteers, and the concurrent WAXAL release (Diack et al., 2026) of image-prompted speech excludes Somali and has no dialect or domain labels.

We introduce HORN-ASR, a curated evaluation benchmark of naturally-occurring spontaneous speech (broadcasts and interviews recorded for a public audience) for these four languages: 1,000 segments per language (15.43 h) selected from a pool of 59,288 utterances (229.1 h) from community radio, television, podcasts, and online media, with gender, dialect, and domain labels on every segment so that error can be reported for the speakers it affects. It is evaluation-only, with no training split, to keep it a measurement instrument. Six current systems are far from saturating it, FLEURS can rank them differently, and the disparities it reveals are invisible on single-gender read speech; for Tigrinya it is the first public independent evaluation. We make four contributions:

1. The first public ASR evaluation benchmark of naturally-occurring spontaneous speech for Amharic, Oromo, Somali, and Tigrinya, balanced for gender and diverse across dialect and domain, for more realistic evaluation, with an inter-annotator agreement study of its references and labels.
2. A replicable, language-agnostic four-stage curation methodology (source gathering, annotation, sampling, and transcription with review and editing) with per-language guidelines.
3. An audit of FLEURS for the three languages it covers: single-gender development and test splits, 3 to 7 speakers per language, and every development speaker also in the test split.
4. An evaluation of six current systems showing that HORN-ASR is harder than FLEURS (0.16 higher WER on average) and revealing gender gaps, and dialect gaps that favor the dominant variety, that no existing benchmark for these languages could show, with all system outputs and scoring code released.

2 RELATED WORK

2.1 NATURALLY-OCCURRING SPEECH CORPORA

Switchboard (Godfrey et al., 1992) and Fisher (Cieri et al., 2004) established that conversational speech exposes failure modes that read-speech evaluation does not, and CHiME (Barker et al., 2018) and AMI (McCowan et al., 2005) did the same for far-field and multi-party conditions. Corpora built from speech that already existed, broadcast and web audio in particular, followed for well-resourced languages: GigaSpeech (Chen et al., 2021) and VoxPopuli (Wang et al., 2021) for English and European languages, and the MGB challenges (Ali et al., 2017; 2019) for Arabic. For low-resource languages the closest analogue is IARPA BABEL (Gales et al., 2014), conversational telephone speech in 26 languages that are not freely redistributable, and for African speech AfriSpeech-200 (Olatunji et al., 2023), which covers accented English rather than African languages. None of these covers Amharic, Oromo, Somali, or Tigrinya, for which, to our knowledge, no naturally-occurring evaluation dataset existed before this work.

2.2 DISAGGREGATED ASR EVALUATION

Aggregate WER hides who a system fails. Koenecke et al. (2020) document error-rate disparities across speaker groups in commercial systems; Tatman (2017) and Garnerin et al. (2019) document gender gaps in captioning and broadcast ASR, and Attanasio et al. (2024) find that the direction of the gap in multilingual models varies by language and model; multi-dialectal Arabic (Ali et al., 2017) and Indic dialect (Javed et al., 2023) evaluations extend disaggregation to other language families; and recent benchmarks make subgroup reporting the primary output (Junczyk, 2024; Rai et al., 2025). All of these require per-utterance metadata that the evaluation sets available for our languages lack. We annotate dialect at the interview level from observed linguistic markers rather than speaker geography, and no prior public corpus exposes gender, dialect, and domain labels for any of the four languages.

2.3 ASR RESOURCES FOR THE FOUR LANGUAGES

FLEURS (Conneau et al., 2023) extends the FLoRes translation benchmark to speech by recording native readers of translated English-Wikipedia sentences for roughly 100 languages; it is the benchmark on which Whisper, MMS, USM, SeamlessM4T v2, and Omnilingual ASR report results,

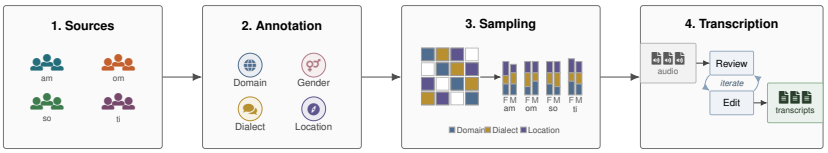

Figure 1: Curation pipeline: (1) source gathering across native-speaker communities, (2) interview-level annotation of domain, gender, dialect, and region, (3) segmentation and joint-balance sampling, with a feedback loop to stage 1 for under-filled strata, and (4) multi-round transcription, review, and editing.

and it covers three of our four languages (Tigrinya is absent) with the audit findings of Section 5.2. MMS (Pratap et al., 2024) reaches ∼1,100 languages through single-speaker New Testament recordings, and AfriHuBERT (Alabi et al., 2025) pretrains on broadcast monologue and religious audio. Public datasets are read-aloud crowdsourcing (Common Voice (Ardila et al., 2020): some Amharic, ∼2 h of Somali, effectively no Oromo or Tigrinya), image-prompted elicitation (WAXAL (Diack et al., 2026), March 2026: Amharic, Oromo, and Tigrinya, no Somali, no dialect or domain labels), and single-language read-speech corpora built for Amharic and Oromo ASR (Abate et al., 2005; Gauthier et al., 2016; Abu et al., 2025). Across these, the register is read or formal, speaker pools are limited, and none exposes dialect or topical-domain labels at segment level.

2.4 Scoring and validity

WER assumes a fixed orthography; for Ge'ez-script languages, where several graphemes are pronunciation-identical writer choices, homophone normalization changes WER by 0.009 to 0.035 on our data (Section 6.7), and Nigatu et al. (2025) argue it belongs at evaluation time rather than in training data, which is how the homophone-aware scheme and the CER reporting of Section 6.7 apply it. Test-set contamination through web-scale training data is a recognized threat to benchmark validity (Magar & Schwartz, 2022; Sainz et al., 2023); Section 7 states what is and is not structurally excluded for a benchmark whose audio is public but whose references are new.

3 Approach

This section describes the language-agnostic design principles and pipeline stages; Section 4 operationalizes them for the four languages.

3.1 Design principles

Three principles set the scope. *First, serve as an evaluation benchmark*: we release neither a train nor a development split, to keep the artifact aligned with its purpose and to avoid contamination via corpus aggregation. *Second, sample real-world spontaneous speech.* A candidate is *naturally-occurring* if the recording predates contributor outreach, was disseminated to a public audience by an independent entity (broadcaster, publisher, host), and the speaker's primary task at recording time was not contributing to a research corpus. Within this constraint we target conversational and interview formats, preserving disfluencies, code-switching, and dialect variation in the audio at source selection (the references omit fillers, Section 4.5). *Third, be representative across speakers and content*: balance across gender, dialect, and topical domain, annotated before sampling and exposed as per-segment metadata, with sampling that rewards under-represented *joint* combinations of the axes, since marginal balance alone lets joint strata stay empty.

3.2 Four-stage curation pipeline

The pipeline (Figure 1) has four stages with a feedback edge from stage 3 to stage 1. *Source gathering*: native-speaker community members return URLs of publicly available recordings, each paired with evidence for inclusion (dialect markers, domain signals, place keywords) that every later stage audits against; per-broadcaster, per-channel, and per-year caps bound any single source. *Annotation*: interview-level domain, gender, dialect, and region, inherited by every segment of the interview, with

dialect attributed by observed linguistic markers rather than geography or self-report. *Segmentation and sampling*: voice-activity segmentation into a duration window, within-interview stratification across positional thirds, and cross-interview sampling that greedily fills a fixed budget with the candidate whose dialect, region, domain, and interview are least represented, under a gender floor and a per-interview cap. *Transcription, review, and editing*: from-scratch transcription by native speakers, then iterative rounds of review and editing by different editors; segments whose disagreements are unresolved at the final editor are excluded. Stages 1 to 3 loop: collection ran in phases targeting under-filled strata (Section 4.2), and the coverage limits that remain are stated in Section 7.

4 METHODOLOGY

This section operationalizes Section 3 for the four languages: dialect targets, source gathering, annotation, sampling, transcription and editing, and the resulting corpus.

4.1 LANGUAGES, SCRIPTS, AND DIALECT TARGETS

Dialect labels, named as they appear in the released metadata and in Section 6: Amharic (Addis Ababa, Shewa, Gojam, Gonder, Wollo, an L2 label for non-native speakers, and an Unknown residual), Oromo (West Central, the Mecha-Tulama varieties; Eastern, Hararghe; Southern, Borana-Arsi-Guji), Somali (Northern Somali, pooling the Somaliland, Ogaadeen, and Puntland labels; Benadiri Somali; Other), and Tigrinya (Z, L, D). Amharic and Tigrinya use the Ge'ez script; Oromo and Somali use Latin. Code-switching with English (all four) and Arabic (Somali) is preserved. Typology references are in Appendix J.

4.2 SOURCE GATHERING

Source URLs were gathered by paid native-speaker contributors per language from a shared guideline. A pilot showed that most Horn-of-Africa community radio and broadcaster content is posted on YouTube (with a smaller share on TikTok) rather than on broadcaster websites. Per-broadcaster, per-channel, and per-year caps bound any single source, and collection ran in phases per language, each targeting the coverage gaps of the previous one. The inventories contain 1,004 / 1,013 / 958 / 922 interviews for Amharic / Oromo / Somali / Tigrinya (Appendix I).

4.3 ANNOTATION

Each interview is annotated for gender, dialect, region, and topical domain. Dialect is attributed by linguistic markers (phonological, lexical, intonational), not by location, which is recorded separately as `region` and is not part of the release. Per-language teams used heterogeneous label vocabularies that were normalized post hoc to common values.

4.4 WITHIN-INTERVIEW SAMPLING

Within-interview sampling is human-driven: contributors review voice-activity candidates (`inaSpeechSegmenter` (Doukhan et al., 2018)) and select up to 20 single-speaker interviewee segments of 5 to 30 s per interview, spread across positional thirds (one interviewee per interview). Algorithmic sampling would conflate interviewer and interviewee speech. The cap provides pool depth for the cross-interview down-selection (up to $k = 6$ per interview; Section 4.6).

4.5 TRANSCRIPTION, EDITING, AND QC

Transcription is from scratch by native-speaker transcribers. Transcripts capture linguistic content, including code-switched tokens and dialect-specific lexical variation, but not disfluencies or fillers; numbers are digits; punctuation follows each script's convention; foreign tokens are transliterated into Ge'ez for Amharic and Tigrinya and kept in Latin orthography for Oromo and Somali. An editor reviews a sample of each transcriber's work and returns feedback, the transcriber fixes the flagged issues, and at least one senior editor performs a complete pass before QC, with different editors at different stages. QC re-edits all single-speaker audio and excludes multi-speaker segments.

4.6 Cross-interview diversity-aware sampling

After within-interview sampling, transcription, and editing, $\sim$3,000 single-speaker, edited segments per language remain. We down-sample to $T = 1{,}000$ per language under a soft female-share floor ($\tau = 0.5$), a hard per-interview cap ($k = 6$), and a joint diversity reward over {dialect, region, domain, interview}: the pool is capped at k segments per interview sorted by QC rating, then the budget is filled greedily with the candidate whose attributes are least represented across the joint axes (inverse-frequency reward), restricted to female speakers while the running female share is below τ (the floor is soft: the restriction applies per step and lifts when no female candidate remains). Pseudocode and a reference implementation are in Appendix L. We down-sample rather than release the pool with sampling weights because hard constraints guarantee the per-interview cap and the female-share floor that weights only approximate in expectation, a fixed set keeps evaluation cost and subgroup sizes identical across languages and systems, and the remainder has not received the final QC re-edit and is not released. The resulting gender shares are in Section 5.1.

4.7 Resulting corpus

The source pool comprises 3,897 distinct interviews, 59,288 utterances, and 229.1 hours of audio (per-language counts in Table 15). The down-selection yields 4,000 segments (1,000 per language), 15.43 hours (Amharic 4.19, Oromo 4.11, Somali 3.61, Tigrinya 3.50), drawn from 1,710 source interviews (524 / 572 / 294 / 320), retaining roughly 6.7% of source utterances: a deliberate choice of edited, axis-balanced segments over raw scale. We hold no speaker identities: segment counts are exact, and the interview count is a proxy for speakers because one interviewee is selected per source interview, but a speaker appearing in several interviews is not merged and nothing is identity-verified (Section 7 states the consequence for the bootstrap).

4.8 Fair compensation and ethics

All contributors (collectors, transcribers, editors) are paid hourly at a rate benchmarked above the highest reported local salary for skilled office work in their region, to reflect their linguistic expertise. The project was assessed under a documented internal ethics protocol (Ethics Statement). Content containing explicit violence or hate speech is flagged at acquisition by contributors, reviewed by the per-language final editors, and excluded if confirmed.

5 Dataset Analysis

5.1 Horn-ASR audits

Each released segment carries gender, dialect, and topical-domain fields inherited from its interview (Section 4.3); acoustic conditions are as captured by the source broadcasters and are not labeled. This subsection gives the subgroup definitions that Section 6 reports on.

Duration. Segment durations span 5 to 30 s (overall mean 13.89 s, median 10.08 s); Amharic and Oromo utterances are longer on average than Somali and Tigrinya ones (Table 13, Appendix H).

Gender. Speaker gender is near-balanced in all four languages (48.2 / 49.7 / 47.7 / 49.8% female for Amh / Orm / Som / Tir) against 31 / 21 / 28 / 39% in the source pool; the residual deviation reflects female-speaker availability in the target dialects, the binding constraint.

Dialect. Per-language label counts are in Table 7. Amharic has seven labels (Addis Ababa 222 to Shewa 85); L2 (196) marks non-native speakers, reflecting the language's lingua-franca role, and Unknown (112) is an annotator residual reported as its own subgroup. Somali's three Northern labels (Somaliland, Ogaadeen, Puntland; 525 utterances) are pooled and contrasted with Benadiri (291); Other (184, diaspora and Maay-area varieties) is excluded from that contrast. Oromo and Tigrinya have three labels each. The dominant variety, against which Section 6.5 measures the dialect gap, is Addis Ababa Amharic (the prestige variety), West Central Oromo (the basis of the written standard), and Tigrinya Z (the majority variety in the benchmark, 542 of 1,000); no such variety is designated

Table 1: HORN-ASR alongside FLEURS and WAXAL on the four target languages. Labels: axes exposed per segment (G = gender, D = dialect, T = topical domain). Design: Eval = evaluation-only, Both = train and evaluation splits. FLEURS hours cover all splits; n/r = not reported per language. ✓ = covered, × = absent. *For HORN-ASR, source interviews (one speaker selected per interview), a proxy for speakers.

	Coverage								
Corpus	Amh	Orm	Som	Tir	Hrs/lang	Speakers*	Register	Labels	Design
HORN-ASR	✓	✓	✓	✓	4.2 / 4.1 / 3.6 / 3.5	1,710	Natur.-occ.	G, D, T	Eval
FLEURS	✓	✓	✓	×	∼10 (all)	3–7/lang	Read	G	Both
WAXAL	✓	✓	×	✓	n/r	n/r	Elicited	G	Both

for Somali. Gender is highly reliable across annotators, dialect moderately, and domain weakly (Appendix F). Coverage limits are stated in Section 7.

Domain. Fifteen topical-domain labels are present (Table 14); the three largest (Politics 573, Social 527, Entertainment 493) cover roughly 40% and reflect broadcaster output. Per language, labels with fewer than 100 utterances are pooled as Other(k) (Table 6).

5.2 FLEURS AUDITS

Two annotators assigned speaker identities to the FLEURS development and test rows that are repeat readings of a prompt (Amharic 170 of 223 development and 394 of 516 test rows; Somali all 432 development and 223 of 1,019 test rows; Oromo all 19 development and 41 test rows), so the speaker counts below are lower bounds (method and per-split table in Appendix O). Three patterns emerge. Each language's development and test splits are single-gender (female for Amharic, male for Oromo and Somali). Every development-split speaker also appears in the test split, which adds at most one unseen speaker: the (development, test, both) speaker counts are $(3, 3, 3)$ for Amharic, $(4, 5, 4)$ for Oromo, and $(6, 7, 6)$ for Somali. The Oromo splits contain only 19 and 41 rows in total, and annotators flagged recurring read-fluency issues in them (non-native pronunciation, omissions, incorrect pauses).

5.3 COMPARISON WITH EXISTING CORPORA

Table 1 situates HORN-ASR alongside FLEURS (read translationese, three of four languages) and WAXAL (Diack et al., 2026) (image-prompted elicitation; no Somali). To our knowledge, HORN-ASR is the only naturally-occurring evaluation set, and the only one with dialect and domain labels, for any of the four languages.

6 EXPERIMENTS AND RESULTS

6.1 SETUP

We evaluate six publicly accessible multilingual ASR systems from three providers that list at least one of the four languages: MMS-1B-all (Pratap et al., 2024), omniASR-LLM-7B and omniASR-CTC-7B from the Omnilingual ASR release (Omnilingual ASR Team et al., 2025), Google's Chirp 2 and Chirp 3 (Cloud Speech-to-Text v2, built on USM (Zhang et al., 2023); outputs collected in July 2026, since hosted models can change over time), and Whisper-large-v3 (Radford et al., 2023). All systems are run as is: the language is set explicitly and no external language model (LM) is used, except in one control reported in Section 6.3. Coverage is uneven: only Meta's systems transcribe Tigrinya, Chirp 3 covers only Amharic among the four, and Whisper supports Amharic and Somali with only 32 and 21 hours of training audio.

All systems are scored with one evaluation protocol (Appendix D). Before scoring, references and system outputs are normalized identically. The *full* scheme removes punctuation, converts Ge'ez-script digits to ASCII, lowercases Latin script, and collapses whitespace; the *phon* scheme also merges Ge'ez characters that sound the same and differ only by writer choice (Section 6.7). Tables report the

Table 2: WER / CER on HORN-ASR and WER on the FLEURS test split for the same system (full normalization; lower is better). Bold marks the best system per language on HORN-ASR; "—" marks a language the system does not support; † marks collapse (WER ≥ 1.0); Tigrinya has no FLEURS split. Confidence intervals: Table 4.

	Amharic			Oromo			Somali			Tigrinya	
	HORN-ASR		FLEURS	HORN-ASR		FLEURS	HORN-ASR		FLEURS	HORN-ASR	
System	WER	CER	WER	WER	CER	WER	WER	CER	WER	WER	CER
MMS-1B-all	0.582	0.228	0.282	0.575	0.179	0.665	0.663	0.260	0.443	0.786	0.351
omniASR-LLM-7B	0.554	0.159	0.177	**0.548**	**0.172**	0.580	**0.493**	**0.181**	0.351	0.784	0.385
omniASR-CTC-7B	0.501	0.205	0.415	0.710	0.375	—	0.750	0.384	—	**0.637**	**0.342**
Chirp 2	**0.275**	**0.107**	0.178	0.728	0.312	0.516	0.693	0.328	0.406	—	—
Chirp 3	0.309	0.138	0.155	—	—	—	—	—	—	—	—
Whisper-large-v3	2.401†	1.461	3.654†	—	—	—	1.004†	0.430	—	—	—

full scheme. The same systems are run on the FLEURS (Conneau et al., 2023) test splits for contrast (Tigrinya has none). Confidence intervals (95%) come from a bootstrap (2,000 resamples) over whole interviews rather than single segments, because segments from one interview are not independent (Section 4.7), the interview-clustered bootstrap. Two systems on the same language are compared with a paired bootstrap over the same interviews. All system outputs, scoring code, and result tables are released (`https://anonymous.4open.science/r/horn-asr-baselines-C765/`).

6.2 OVERALL RESULTS

Table 2 reports WER and character error rate (CER) per system and language on HORN-ASR, with the same system's FLEURS WER beside it. No system is close to saturating the benchmark. The best result on any language is 0.275 WER (Chirp 2 on Amharic) and, apart from Chirp 3 on Amharic (0.309), every other model–language pair lies between 0.49 and 0.79 WER. The best system differs by language: Chirp 2 on Amharic, omniASR-LLM-7B on Oromo and Somali, omniASR-CTC-7B on Tigrinya, for which, to our knowledge, this is the first public independent evaluation. Whisper-large-v3 collapses on both languages it supports (WER at least 1.0 on HORN-ASR, more errors than reference words, and 3.65 on FLEURS Amharic), so we exclude it from the averages below. The rankings are not due to chance: 30 of the 34 pairs of systems scored on the same language differ significantly in WER (paired bootstrap, 95%, no correction for multiple comparisons).

6.3 HORN-ASR VERSUS FLEURS

Table 2 places the two benchmarks side by side for every model–language pair both cover. Over the eleven shared pairs (Whisper excluded), the same systems average 0.539 WER on HORN-ASR against 0.379 on FLEURS, an absolute gap of 0.16, and 0.206 against 0.111 CER; over the eight Amharic and Somali pairs alone the gap is 0.21 (0.509 against 0.301), and HORN-ASR WER is 1.2 to 3.1 times the FLEURS WER per pair. On Oromo the ratios are 0.9 to 1.4, but the FLEURS test split is 41 read sentences with the fluency problems noted in Section 5.2, so we do not interpret the Oromo comparison. The two benchmarks also disagree about which system is best on Amharic: Chirp 3 leads FLEURS but Chirp 2 leads HORN-ASR, where the difference is significant (we did not test significance on FLEURS). A benchmark of read, translated sentences therefore does not just under-estimate error on spontaneous broadcast and interview speech; it can also rank systems differently.

Two controls address alternative explanations. First, adding an LM does not close the gap. For the one system with a released LM, Meta's Somali LM (trained on Common Crawl) lowers the verbatim (unnormalized) WER of MMS-1B-all on HORN-ASR Somali only from 0.716 to 0.684, even with the LM weight tuned on the test set itself; the default MMS weight ($\alpha = 1.0$) gives 0.863. With the same LM, MMS reports 0.428 on FLEURS (Table 12). Second, our scoring reproduces the published number where one exists: verbatim FLEURS Somali WER for MMS-1B-all is 0.538, as reported. Developers' results on their own held-out data look better still: the Omnilingual ASR release reports 0.057 to 0.098 CER for the four languages, while omniASR-LLM-7B measures 0.159 to 0.385 CER on HORN-ASR, 2.2 to 3.9 times higher.

Table 3: WER on HORN-ASR by speaker gender (F / M, full normalization). Bold marks the better side; * marks gaps whose interview-clustered 95% interval excludes zero.

	Amharic		Oromo		Somali		Tigrinya	
System	F	M	F	M	F	M	F	M
MMS-1B-all	**0.565**	0.602*	0.590	**0.556***	0.667	**0.659**	**0.769**	0.805*
omniASR-LLM-7B	**0.526**	0.587*	**0.514**	0.589*	0.498	**0.489**	**0.767**	0.804
omniASR-CTC-7B	**0.496**	0.507	**0.700**	0.721	0.778	**0.721***	**0.619**	0.657
Chirp 2	**0.254**	0.299*	**0.683**	0.782*	0.694	**0.693**	—	—
Chirp 3	**0.291**	0.330*	—	—	—	—	—	—
Whisper-large-v3	**2.225**	2.600*	—	—	**0.995**	1.012	—	—

6.4 BREAKDOWN BY GENDER

Every FLEURS split for these languages is single-gender (Section 5.2), so FLEURS cannot measure a gender gap at all. Table 3 reports WER on female and male speakers and tests whether the female-minus-male gap differs from zero, using the interview-clustered bootstrap. At least one system shows a significant gap in every language, but direction is model-dependent: on Amharic four of the five non-collapsed systems (and Whisper) are significantly more accurate on female speakers; on Oromo omniASR-LLM-7B and Chirp 2 favor female and MMS-1B-all male speakers, all significantly; on Somali the only significant gap favors male speakers (omniASR-CTC-7B); on Tigrinya MMS-1B-all favors female speakers. A single-gender benchmark shows none of this, and conclusions that models are unbiased, drawn from high-resource languages, do not carry over.

6.5 BREAKDOWN BY DIALECT

Every HORN-ASR utterance carries a dialect label (Section 4.1). Table 8 reports, per language and system, the easiest and hardest dialect and the gap between them, and Table 9 the full per-dialect grid with confidence intervals. For Somali we compare two groups, Northern Somali (three released labels pooled) and Benadiri Somali, following the taxonomy in Appendix J. To our knowledge this is the first dialect-disaggregated WER report for any of the four languages. The dominant variety is easiest in three of four languages: Addis Ababa Amharic for every system, West Central Oromo for three of four, and Tigrinya dialect Z for every system. The hardest are Shewa, Wollo, or Gojam Amharic, Eastern or Southern Oromo, and Tigrinya dialect D. The WER gap between the easiest and hardest dialect is 0.08 to 0.15 on Amharic, 0.04 to 0.15 on Oromo, and 0.10 to 0.22 on Tigrinya. On Somali, the Northern and Benadiri confidence intervals overlap for every non-collapsed system. The hardest dialect is consistent within a language but the size of the penalty is system-specific, and the advantage of the dominant variety is consistent with its over-representation in web and broadcast training data. Dialect labels have moderate inter-annotator reliability (Appendix F), so we report each axis separately, not their combinations; every dialect subgroup has at least 85 utterances from 42 interviews.

6.6 BREAKDOWN BY DOMAIN

Domain labels have the lowest inter-annotator agreement of the three axes (Appendix F) and several domain subgroups have fewer than 100 utterances, so we treat per-domain results as exploratory, report them only in the released result files, and make no claims comparing systems by domain.

6.7 METRIC CHOICES

Three findings shape how these languages should be scored (Tables 10 and 11). First, published per-language numbers are scored verbatim, without normalization. For languages without an enforced spelling standard, this counts spelling choices as recognition errors: on HORN-ASR Amharic, averaged over the non-collapsed systems, full normalization lowers WER by 0.070 compared with verbatim scoring, and homophone collapsing lowers it by a further 0.035 (0.018 to 0.045 across systems); on Tigrinya, 0.037 and 0.009. Second, CER changes less under these choices (a 9.0% relative reduction from verbatim to phon, against 12.7% for WER), so it should be reported alongside

WER. Third, WER and CER ignore whether a system splits its output into sentences. Scoring Ge'ez end-of-sentence markers (፡ and ?) as tokens shows that no Meta system produces any on HORN-ASR, that Whisper still punctuates its Somali output (sentence-marker F1 0.55) even while its words collapse, and that the two systems that produce Ge'ez markers recall 0.83 to 0.99 of them on single-sentence FLEURS rows but only 0.10 to 0.38 on multi-sentence HORN-ASR segments (Table 11).

7 LIMITATIONS

Coverage. We designed the dialect coverage of each language deliberately (Section 4.1), but it is not exhaustive: Maay is not treated as a separate Somali dialect (Maay-area speech falls under Other), and varieties beyond the labeled buckets may be under-represented. Broadcast sourcing over-represents urban, educated, and prominent speakers within each dialect, contains no children, and likely contains few elderly speakers. Gender is a label assigned by annotators from the audio (female, male, or unknown), not a claim about the speaker's identity. Because each interview stands in for one speaker (Section 4.7), a person who appears in several interviews is treated as several independent clusters in the bootstrap, so confidence intervals may be too narrow. With 4,000 segments, subgroups that combine gender, dialect, and domain are too small to test, so we report each label on its own, with at least 30 utterances and 10 interviews per subgroup.

Metrics. Transcribing spontaneous speech in these languages consistently is hard: none has an enforced orthographic standard, and dialect variation adds further spelling and word-boundary choices. In our agreement study, three annotators who transcribed the same 100 utterances per language from scratch, in a single pass and without review, differed from each other by 0.26 to 0.48 WER under the homophone-aware scheme (Appendix F). This first-pass disagreement is what the multi-round review-and-editing pipeline (Section 4.5) was built to solve, and the released references are closer to each annotator than the annotators are to each other. Even so, WER against a single reference cannot resolve system differences smaller than the variation among careful transcribers; homophone-aware normalization and CER narrow this without removing it and are the metrics we recommend. Fillers and disfluencies are not transcribed in the references because annotators disagreed strongly on them; a consistent transcription is left for future work.

Contamination. The references were produced for this benchmark and were not public before the evaluated systems were released, so no system can have trained on this audio paired with these transcripts. We cannot rule out that the audio alone was used in pretraining, or that captions of the same broadcasts made by others appear in web-scraped training sets; the uniformly high error and the developers' own-data results, 2.2 to 3.9 times better than ours, argue against memorization without proving its absence. Going forward, the Terms of Use prohibit training on HORN-ASR, papers reporting on it are asked to state that their training data excludes it, the broadcast-year field supports evaluation on segments broadcast after a training cutoff, and a refresh release is committed within 24 months (Ethics Statement).

8 CONCLUSION

We introduced HORN-ASR, a benchmark dataset built as an evaluation instrument for ASR on four Horn of Africa languages. It complements read-speech benchmarks with what they cannot provide: naturally-occurring spontaneous broadcast and interview speech, gender balance, dialect and domain labels on every segment, and one evaluation protocol under which we measure six current systems. Three findings follow. Error on read translationese does not predict error on spontaneous speech, and the two can rank systems differently. Aggregate error hides two kinds of gap: gender gaps whose direction depends on the system, so fairness conclusions drawn from high-resource languages do not carry over, and dialect gaps that consistently favor the dominant variety. And first-pass transcriptions of this speech by independent annotators differ by 0.26 to 0.48 WER, so system differences near that level should be reported with CER, a homophone-aware scheme, and interview-clustered intervals rather than as rankings. The pipeline is replicable: a team can apply it to build a private test set for its own languages. Closing these gaps requires better training data; measuring them comes first.

ETHICS STATEMENT

HORN-ASR is intended as a measurement instrument for ASR evaluation on Amharic, Oromo, Somali, and Tigrinya, including disaggregated per-gender and per-dialect reporting. The artifacts we authored are released under open licenses: transcripts, labels, metadata, and guidelines under CC BY-SA 4.0, and the scoring code under the MIT license. The audio is not relicensed: it is redistributed under binding Terms of Use through a gated dataset repository (access on request from an identified account that accepts the Terms of Use; access is revocable and logged), alongside a Gebru-style datasheet (Gebru et al. 2021; Appendix A), a Croissant metadata export with the Responsible AI extension, the per-language annotation guidelines, and the evaluation protocol and scoring code (Appendix D). The dataset card, guidelines, scoring code, and transcripts are public; the audio and the per-segment metadata sit behind the gate.

The audio consists of short (5 to 30 s), selective excerpts of publicly available interviews and speech, downloaded from the broadcasters' own public postings under per-broadcaster, per-channel, and per-year collection caps, and used as evaluation instances. Speakers were recorded for public broadcast, not for research-corpus inclusion; we do not claim that public availability implies consent, and we rely only on the narrower premise that research evaluation of ASR on broadcast speech is consistent with the speaker's expectation that the broadcast would be heard, transcribed, and discussed. The copyright position is a research-use argument (short, non-substitutable excerpts), not a legal determination. The release contains no names, source URLs, or broadcast dates beyond the calendar year; the internal provenance mapping that links segments to sources is held only for takedown and errata and is not released. Removal requests from any speaker or broadcaster are honored through the takedown protocol in Appendix G.1, which includes requester verification and a public removal log. The Terms of Use prohibit voice cloning, surveillance, speaker identification, training-data reuse on HORN-ASR or near-duplicates, and demographic profiling derived from per-segment metadata. The contributor right of post-hoc labor withdrawal and the minors-exclusion policy are documented in Appendix G.2; the versioning, errata cadence, and 24-month refresh-release commitment in Appendix G.3. The project was assessed under the documented internal ethics protocol in Appendix M. All contributors were paid hourly at rates benchmarked above skilled local office work (Section 4.8).

REPRODUCIBILITY STATEMENT

Scoring is fully reproducible: the anonymized evaluation repository (`https://anonymous.4open.science/r/horn-asr-baselines-C765/`) contains the outputs of every system on every model–language pair (one JSON file per system, benchmark, and language, plus the LM-decoding control), the self-contained scoring code (normalization schemes, homophone tables, WER/CER/chrF, the interview-clustered and paired bootstraps with seed 0), the result CSVs, and a manifest with checksums. Reference transcripts are never copied into that repository; the scripts fetch them at scoring time from pinned revisions of the released HORN-ASR transcripts (public) and FLEURS, so every number in Section 6 and Appendix E regenerates deterministically from the published references. Inference is reproducible for the open-weight systems with the versions and decoding settings in Appendix D; the Chirp results are API captures from July 2026 and cannot be re-run identically, which is why the outputs themselves are released. The paper's LaTeX tables are produced from the CSVs by a single script in the same repository. The curation methodology is specified at the operational level in Sections 3 and 4, with sampler pseudocode and a reference implementation in Appendix L, the per-language annotation and transcription guidelines in Appendix B, and the source inventory in Appendix I. The inter-annotator agreement study is reported in full in Appendix F.

AI USE STATEMENT

Large language models were used as writing and coding assistants: to edit and condense prose drafted by the authors, to draft LaTeX table generation and scoring scripts that the authors reviewed and ran, and to check the manuscript for consistency between the text and the released result files. All system outputs reported here were produced by the evaluated systems themselves; all reference transcripts were produced from scratch by paid native-speaker contributors and edited in rounds by human editors, and all gender, dialect, and domain labels and the inter-annotator agreement study

were produced by human annotators. The scoring pipeline was checked against a published number: it reproduces the FLEURS Somali WER reported by MMS. No language model contributed to the design of the benchmark, the annotation of the data, or the interpretation of the results.

REFERENCES

Solomon Teferra Abate, Wolfgang Menzel, and Bairu Tafila. An Amharic speech corpus for large vocabulary continuous speech recognition. In *Proc. Interspeech*, pp. 1601–1604, 2005. doi: 10.21437/Interspeech.2005-467.

Turi Abu, Ying Shi, Thomas Fang Zheng, and Dong Wang. Sagalee: An open source automatic speech recognition dataset for Oromo language. In *Proc. IEEE ICASSP*, pp. 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10890761. arXiv:2502.00421.

Jesujoba O. Alabi, Xuechen Liu, Dietrich Klakow, and Junichi Yamagishi. AfriHuBERT: A self-supervised speech representation model for African languages. In *Proc. Interspeech*, 2025. arXiv:2409.20201.

Ahmed Ali, Stephan Vogel, and Steve Renals. Speech recognition challenge in the wild: Arabic MGB-3. In *Proc. ASRU*, 2017.

Ahmed Ali, Suwon Shon, Younes Samih, Hamdy Mubarak, Ahmed Abdelali, James Glass, Steve Renals, and Khalid Choukri. The MGB-5 challenge: Recognition and dialect identification of dialectal Arabic speech. In *Proc. ASRU*, 2019.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. Common Voice: A massively-multilingual speech corpus. In *Proc. LREC*, pp. 4218–4222, 2020. URL https://aclanthology.org/2020.lrec-1.520/.

Giuseppe Attanasio, Beatrice Savoldi, Dennis Fucci, and Dirk Hovy. Twists, humps, and pebbles: Multilingual speech recognition models exhibit gender performance gaps. In *Proc. EMNLP*, pp. 21318–21340, 2024. doi: 10.18653/v1/2024.emnlp-main.1188.

Jon Barker, Shinji Watanabe, Emmanuel Vincent, and Jan Trmal. The fifth CHiME speech separation and recognition challenge: Dataset, task and baselines. *Proc. Interspeech*, pp. 1561–1565, 2018.

Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, et al. GigaSpeech: An evolving, multi-domain ASR corpus with 10,000 hours of transcribed audio. In *Proc. Interspeech*, pp. 3670–3674, 2021.

Christopher Cieri, David Miller, and Kevin Walker. The Fisher corpus: A resource for the next generations of speech-to-text. In *Proc. LREC*, pp. 69–71, 2004. URL https://aclanthology.org/L04-1500/.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. FLEURS: Few-shot learning evaluation of universal representations of speech. In *Proc. 2022 IEEE Spoken Language Technology Workshop (SLT)*, pp. 798–805, 2023. doi: 10.1109/SLT54892.2023.10023141.

Abdoulaye Diack, Perry Nelson, Kwaku Agbesi, Angela Nakalembe, MohamedElfatih Mohamed-Khair, et al. WAXAL: A large-scale multilingual African language speech corpus. *arXiv preprint arXiv:2602.02734*, 2026.

David Doukhan, Jean Carrive, et al. inaSpeechSegmenter: An open-source cnn-based audio segmentation toolkit for voice activity detection, music detection, and gender detection, 2018.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig (eds.). *Ethnologue: Languages of the World*. SIL International, Dallas, Texas, twenty-seventh edition, 2024. URL https://www.ethnologue.com.

Mark J. F. Gales, Kate M. Knill, Anton Ragni, and Shakti P. Rath. Speech recognition and keyword spotting for low-resource languages: BABEL project research at CUED. In *Proc. 4th International Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU)*, pp. 16–23, 2014. URL https://www.isca-archive.org/sltu_2014/gales14_sltu.html.

Mahault Garnerin, Solange Rossato, and Laurent Besacier. Gender representation in French broadcast corpora and its impact on ASR performance. In *Proc. 1st International Workshop on AI for Smart TV Content Production, Access and Delivery (AI4TV), ACM Multimedia*, pp. 3–9, 2019. doi: 10.1145/3347449.3357480.

Elodie Gauthier, Laurent Besacier, Sylvie Voisin, Michael Melese, and Uriel Pascal Elingui. Collecting resources in sub-Saharan African languages for automatic speech recognition: a case study of Wolof. In *Proc. LREC*, 2016.

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

John J Godfrey, Edward C Holliman, and Jane McDaniel. Switchboard: Telephone speech corpus for research and development. In *Proc. ICASSP*, pp. 517–520, 1992.

Catherine Griefenow-Mewis. *A Grammatical Sketch of Written Oromo*. Rüdiger Köppe Verlag, Köln, 2001.

Asfaw Gedamu Haileslasie, Asmelash Teka Hadgu, and Solomon Teferra Abate. Tigrinya dialect identification. In *AfricaNLP Workshop at ICLR 2023*, 2023.

Kebede Hordofa Janko. *Towards the Genetic Classification of the Afaan Oromoo Dialects*. Phd dissertation, University of Oslo, 2009.

Tahir Javed, Kaushal Bhogale, Abhigyan Raman, Pratyush Kumar, Anoop Kunchukuttan, and Mitesh M. Khapra. IndicSUPERB: A speech processing universal performance benchmark for Indian languages. In *Proc. AAAI Conference on Artificial Intelligence*, volume 37, pp. 12942–12950, 2023. doi: 10.1609/aaai.v37i11.26521.

Michał Junczyk. BIGOS V2 benchmark for Polish ASR: Curated datasets and tools for reproducible evaluation. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024), Datasets and Benchmarks Track*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/69bddcea866e8210cf483769841282dd-Abstract-Datasets_and_Benchmarks_Track.html.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689, 2020. doi: 10.1073/pnas.1915768117.

Marcello Lamberti. *Map of Somali Dialects in the Somali Democratic Republic*, volume 5 of *Studia linguarum Africae orientalis*. Helmut Buske Verlag, Hamburg, 1986.

Inbal Magar and Roy Schwartz. Data contamination: From memorization to exploitation. In *Proc. ACL*, 2022.

Iain McCowan, Jean Carletta, Wessel Kraaij, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, et al. The AMI meeting corpus. In *Proc. Measuring Behavior*, 2005.

Niguss Weldezgu Mehari. *A Grammar of Rayya Tigrinya*. Phd dissertation, Addis Ababa University, 2021.

Tsehaye Kiros Mengesha. A comparison of Wajerat Tigrigna vs. Standard Tigrigna, 2009.

Hellina Hailu Nigatu, Atnafu Lambebo Tonja, Henok Biadglign Ademtew, Hizkiel Mitiku Alemayehu, Negasi Haile Abadi, Tadesse Destaw Belay, and Seid Muhie Yimam. A case against implicit standards: Homophone normalization in machine translation for languages that use the Ge'ez script. In *Proc. EMNLP*, pp. 10309–10320, 2025. doi: 10.18653/v1/2025.emnlp-main.523.

Tobi Olatunji, Tejumade Afonja, Aditya Yadavalli, Chris Chinenye Emezue, Sahib Singh, Bonaventure F. P. Dossou, Joanne Osuchukwu, Salomey Osei, Atnafu Lambebo Tonja, Naome Etori, and Clinton Mbataku. AfriSpeech-200: Pan-african accented speech dataset for clinical and general domain ASR. *Transactions of the Association for Computational Linguistics*, 11:1669–1685, 2023. doi: 10.1162/tacl_a_00627.

Omnilingual ASR Team, Gil Keren, Artyom Kozhevnikov, Yen Meng, Christophe Ropers, et al. Omnilingual ASR: Open-source multilingual speech recognition for 1600+ languages. *arXiv preprint arXiv:2511.09690*, 2025.

Jonathan Owens. *A Grammar of Harar Oromo (Northeastern Ethiopia)*. Helmut Buske Verlag, Hamburg, 1985.

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, 2021.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52, 2024. URL https://jmlr.org/papers/v25/23-1318.html.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proc. ICML*, pp. 28492–28518, 2023.

Anand Rai, Satyam Rahangdale, Utkarsh Anand, and Animesh Mukherjee. ASR-FAIRBENCH: Measuring and benchmarking equity across speech recognition systems. In *Proc. Interspeech*, 2025. URL https://www.isca-archive.org/interspeech_2025/rai25_interspeech.html.

John I. Saeed. *Somali*, volume 10 of *London Oriental and African Language Library*. John Benjamins, Amsterdam, 1999.

Oscar Sainz, Jon Ander Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of EMNLP*, 2023.

Seamless Communication Team. SeamlessM4T: Massively multilingual and multimodal machine translation. *arXiv preprint arXiv:2308.11596*, 2023.

Harry Stroomer. *A Grammar of Boraana Oromo (Kenya)*. Rüdiger Köppe Verlag, Köln, 1995.

Mengistu Tadesse. *Documentation and Description of Amharic Dialects*. Phd thesis, Addis Ababa University, Addis Ababa, Ethiopia, 2018. URL https://etd.aau.edu.et/items/a93786a4-49e4-491b-9a2f-8008b9a7466b. Department of Linguistics: Documentary Linguistics and Culture Program, College of Humanities, Language Studies, Journalism and Communication.

Rachael Tatman. Gender and dialect bias in YouTube's automatic captions. In *Proc. First ACL Workshop on Ethics in Natural Language Processing (EthNLP)*, pp. 53–59, 2017. doi: 10.18653/v1/W17-1606.

Mauro Tosco. The unity and diversity of Somali dialectal variants. In *The Languages of Africa and the Diaspora: Educating for Language Awareness*. Multilingual Matters, 2012.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proc. ACL*, 2021.

Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, et al. Google USM: Scaling automatic speech recognition beyond 100 languages. *arXiv preprint arXiv:2303.01037*, 2023.

A Datasheet for Datasets

We follow the Gebru et al. datasheets-for-datasets template (Gebru et al., 2021). The full seven-section datasheet (Motivation, Composition, Collection Process, Preprocessing / Cleaning / Labeling, Uses, Distribution, Maintenance) is released as a standalone PDF with the dataset; we summarize the essential answers here.

Motivation. HORN-ASR was created to fill a gap: there is no publicly available naturally-occurring spontaneous-speech ASR evaluation benchmark for Amharic, Oromo, Somali, or Tigrinya (Section 2). Existing public datasets for these languages are read or elicited and have small speaker pools.

Composition. Each instance is a single-speaker audio segment (5 to 30 s) extracted from a publicly broadcast interview, paired with a human-edited transcript and per-segment metadata (gender, dialect, domain). The benchmark contains 4,000 segments (1,000 per language), 15.43 hours of audio total, drawn from 1,710 source interviews (one speaker selected per interview; a proxy for distinct speakers, not an exact count). Splits: HORN-ASR v1 is released as a single evaluation split (no train, no dev).

Collection. The four-stage curation pipeline (Section 3, Section 4) was operated by paid native-speaker contributors per language: source URLs gathered from publicly disseminated speech recordings (broadcaster channels, podcasts, online media), interview-level annotation, voice-activity-based segmentation, and from-scratch transcription with a multi-role edit cycle.

Preprocessing, cleaning, labeling. Audio is converted to 16 kHz mono PCM 16-bit WAV; source-level encoder metadata is stripped. Transcripts capture linguistic content per the per-language annotation guideline (Appendix B); disfluencies and filler words are not transcribed. Multi-speaker segments are excluded.

Uses. HORN-ASR is a measurement instrument, intended for *evaluation* of multilingual ASR systems. The Terms of Use prohibit voice cloning, surveillance, training-data reuse, and demographic profiling (see the Ethics Statement).

Distribution. Authored artifacts (transcripts, labels, metadata, guidelines) are released under CC BY-SA 4.0; the audio is redistributed under binding Terms of Use through a gated dataset repository (link in the README of the anonymized evaluation repository).

Maintenance. Versioned on the Hub with an errata cadence and a 24-month refresh release commitment (Appendix G.3).

B Annotation Guidelines

The full per-language transcription and editing guidelines used by the contributor teams will be released alongside the dataset upon acceptance. They cover the common conventions (matrix-language script rule for code-switching; numbers written as digits; abbreviation expansion; punctuation; non-transcribed filler words and non-speech sounds) and the per-language specifics for Amharic, Tigrinya, Oromo, and Somali (Ge'ez-script homophone unification for Amh/Tig; Latin-script Qubee conventions for Oromo; Ethiopia / Somalia / Somaliland / Djibouti orthographic variants for Somali).

C Per-Segment Metadata Schema

The per-segment metadata file released with the dataset is a flat-tabular file (one row per segment) with the fields below. All segments are 16 kHz mono PCM 16-bit WAV; per-segment audio files are referenced by `segment_id`. The dialect and domain bucket sets are described in Appendix J and Section 5.1 respectively.

Field	Type	Allowed values / notes
`segment_id`	string	Unique per segment; basename of the WAV file.
`interview_id`	string	Stable identifier for the source interview; multiple segments share an interview.
`language`	enum	{`am`, `om`, `so`, `ti`}.
`dialect`	enum	Per-language bucket set (Appendix J).
`gender`	enum	{`F`, `M`, `U`} (`U` = unknown; rare, only Tigrinya v1).
`domain`	enum	One of fifteen labels (Table 14).
`start_offset_s`	float	Segment start time within the source interview, in seconds.
`end_offset_s`	float	Segment end time within the source interview, in seconds.
`duration_s`	float	End minus start; 5 to 30 s.
`transcript`	string	UTF-8 transcript per the per-language guideline (Appendix B).
`source_outlet`	string	Hashed identifier for the source broadcaster / channel (Appendix I).
`broadcast_year`	int	Calendar year of the source broadcast.

D Evaluation Protocol

The evaluation protocol is released as self-contained scoring code with the system hypotheses (`https://anonymous.4open.science/r/horn-asr-baselines-C765/`). It fixes (i) the text normalization applied identically to references and hypotheses, (ii) the metrics, and (iii) the uncertainty procedure.

Normalization schemes. Four schemes are defined. *Verbatim* applies no processing: WER tokens are the whitespace-separated strings as published and as decoded, and CER is computed over characters with spaces retained. This matches the scoring protocol of the MMS paper's per-language FLEURS tables, and our verbatim FLEURS Somali WER for MMS-1B-all reproduces the reported 0.538. *Full*, the main scheme of the result tables, strips all punctuation (Ge'ez and Latin), transliterates Ge'ez-script digits to ASCII, lowercases Latin-script text, and collapses whitespace. *Phon* adds language-specific collapsing of Ge'ez homophone graphemes that are pronunciation-identical writer choices: for Amharic the ሐ-, ዐ-, and ኸ-series onto ሀ, ሠ onto ሰ, ፀ onto ጸ, ዀ onto ኰ, and the vowel-order merger ሃ onto ሀ and ኻ onto ኸ; for Tigrinya ሠ onto ሰ, ዐ onto ሀ, and ጸ onto ፀ (canonical ፀ, the reverse of Amharic), while ሐ /ħ/, ኸ /x/, and ዀ /ʕ/ are kept apart because they are distinct phonemes in Tigrinya. Each merged set is one word written differently by different transcribers within the same released references (for example, 27 and 21 occurrences of the two spellings of Amharic "country"). Phon equals full for the Latin-script languages. *EOS* is full with the sentence-end markers (። and ? for Ge'ez script, . and ? for Latin) kept and scored as standalone tokens, with terminal-marker precision, recall, and F1 reported alongside. Whisper's own normalizer sits closest to full; MMS scores verbatim; USM reports language-averaged CER, and Omnilingual ASR per-language CER on its own held-out data, neither with a per-language normalization recipe, so the schemes bracket every published protocol.

Metrics. WER is computed at the token level (whitespace split) and CER at the character level after Unicode NFC, both micro-averaged within each subgroup. chrF (sacreBLEU, on full-scheme text) is reported for subgroup tables. Fillers and disfluencies are absent from the references by the transcription guideline; hypothesis-side filler stripping is not applied in this version.

Uncertainty. Confidence intervals are interview-clustered bootstraps: for each subgroup, the interview-level speaker identifiers are resampled with replacement, all utterances of each resampled identifier are retained, and the micro-averaged metric is recomputed; 2,000 resamples, seed 0, percentile intervals at $[2.5, 97.5]\%$. Same-language system pairs are compared with a paired bootstrap over the same clusters, and a difference is called significant when the interval of the paired difference excludes zero. Subgroups below 30 utterances or 10 interviews are suppressed rather than reported (Appendix E).

Systems. MMS-1B-all is run with its per-language adapters, greedy CTC decoding, and no LM (`target_lang` forced); omniASR-LLM-7B with beam 5 and the language forced; omniASR-CTC-7B with greedy CTC and no LM; Chirp 2 and Chirp 3 through the Cloud Speech-to-Text v2 API with defaults and the language forced (Chirp 2 accessed 2026-07-24, Chirp 3 2026-07-01); Whisper-large-v3 with default transcription and the language forced. The LM-decoding control runs MMS-1B-all with Meta's released Common-Crawl Somali LM across LM weights (Table 12).

E Full Evaluation Results

All tables in this appendix are generated from the released result CSVs (`https://anonymous.4open.science/r/horn-asr-baselines-C765/`) by a single script; the main-text tables are drawn from the same files.

Confidence intervals and pairwise tests. Table 4 gives interview-clustered bootstrap 95% intervals (2,000 resamples) for every model–language pair on both benchmarks. Of the 34 same-language system pairs on Horn-ASR, the four whose WER difference is not significant under the paired bootstrap are: Oromo, Chirp 2 versus omniASR-CTC-7B ($+0.018$); Somali, Chirp 2 versus MMS-1B-all ($+0.030$); Somali, Chirp 2 versus omniASR-CTC-7B (-0.056); Tigrinya, MMS-1B-all versus omniASR-LLM-7B ($+0.001$).

Table 4: WER and CER (full normalization) with interview-clustered bootstrap 95% confidence intervals for every model–language pair.

System	Language	Benchmark	n	WER	95% CI	CER	95% CI
MMS-1B-all	Amharic	Horn-ASR	1000	0.582	[0.569, 0.596]	0.228	[0.219, 0.238]
MMS-1B-all	Amharic	FLEURS	516	0.282	[0.267, 0.297]	0.079	[0.073, 0.084]
MMS-1B-all	Oromo	Horn-ASR	1000	0.575	[0.562, 0.589]	0.179	[0.172, 0.186]
MMS-1B-all	Oromo	FLEURS	41	0.665	[0.622, 0.707]	0.189	[0.173, 0.206]
MMS-1B-all	Somali	Horn-ASR	1000	0.663	[0.650, 0.677]	0.260	[0.249, 0.269]
MMS-1B-all	Somali	FLEURS	1019	0.443	[0.433, 0.453]	0.133	[0.127, 0.140]
MMS-1B-all	Tigrinya	Horn-ASR	1000	0.786	[0.774, 0.797]	0.351	[0.342, 0.360]
omniASR-LLM-7B	Amharic	Horn-ASR	1000	0.554	[0.531, 0.579]	0.159	[0.151, 0.169]
omniASR-LLM-7B	Amharic	FLEURS	516	0.177	[0.165, 0.189]	0.051	[0.047, 0.055]
omniASR-LLM-7B	Oromo	Horn-ASR	1000	0.548	[0.534, 0.563]	0.172	[0.164, 0.180]
omniASR-LLM-7B	Oromo	FLEURS	41	0.580	[0.539, 0.619]	0.164	[0.148, 0.181]
omniASR-LLM-7B	Somali	Horn-ASR	1000	0.493	[0.477, 0.511]	0.181	[0.171, 0.193]
omniASR-LLM-7B	Somali	FLEURS	1019	0.351	[0.340, 0.362]	0.116	[0.109, 0.124]
omniASR-LLM-7B	Tigrinya	Horn-ASR	1000	0.784	[0.761, 0.810]	0.385	[0.364, 0.407]
omniASR-CTC-7B	Amharic	Horn-ASR	1000	0.501	[0.481, 0.522]	0.205	[0.188, 0.222]
omniASR-CTC-7B	Amharic	FLEURS	516	0.415	[0.390, 0.441]	0.079	[0.075, 0.084]
omniASR-CTC-7B	Oromo	Horn-ASR	1000	0.710	[0.690, 0.729]	0.375	[0.347, 0.404]
omniASR-CTC-7B	Somali	Horn-ASR	1000	0.750	[0.731, 0.768]	0.384	[0.350, 0.414]
omniASR-CTC-7B	Tigrinya	Horn-ASR	1000	0.637	[0.616, 0.660]	0.342	[0.317, 0.367]
Chirp 2	Amharic	Horn-ASR	1000	0.275	[0.265, 0.287]	0.107	[0.100, 0.114]
Chirp 2	Amharic	FLEURS	516	0.178	[0.166, 0.190]	0.059	[0.053, 0.066]
Chirp 2	Oromo	Horn-ASR	1000	0.728	[0.699, 0.763]	0.312	[0.286, 0.342]
Chirp 2	Oromo	FLEURS	41	0.516	[0.474, 0.554]	0.154	[0.138, 0.171]
Chirp 2	Somali	Horn-ASR	1000	0.693	[0.636, 0.764]	0.328	[0.286, 0.379]
Chirp 2	Somali	FLEURS	1019	0.406	[0.396, 0.416]	0.143	[0.138, 0.150]
Chirp 3	Amharic	Horn-ASR	1000	0.309	[0.297, 0.323]	0.138	[0.128, 0.150]
Chirp 3	Amharic	FLEURS	516	0.155	[0.143, 0.167]	0.049	[0.043, 0.056]
Whisper-large-v3	Amharic	Horn-ASR	1000	2.401	[2.302, 2.508]	1.461	[1.401, 1.525]
Whisper-large-v3	Amharic	FLEURS	516	3.654	[3.483, 3.822]	2.360	[2.275, 2.443]
Whisper-large-v3	Somali	Horn-ASR	1000	1.004	[0.963, 1.053]	0.430	[0.404, 0.461]

Subgroup sizes and minimum-size rule. Table 6 lists utterance and interview counts for every reported subgroup. We report single-axis subgroups (gender, dialect, domain) and never three-

Table 5: WER on HORN-ASR versus FLEURS for every model–language pair both benchmarks cover (full normalization), with the HORN-ASR/FLEURS ratio. Tigrinya has no FLEURS split. † marks collapse.

	Amharic ($n_{\text{FLEURS}} = 516$)			Oromo ($n_{\text{FLEURS}} = 41$)			Somali ($n_{\text{FLEURS}} = 1019$)		
System	Horn	FLEURS	ratio	Horn	FLEURS	ratio	Horn	FLEURS	ratio
MMS-1B-all	0.582	0.282	2.1	0.575	0.665	0.9	0.663	0.443	1.5
omniASR-LLM-7B	0.554	0.177	3.1	0.548	0.580	0.9	0.493	0.351	1.4
omniASR-CTC-7B	0.501	0.415	1.2	—	—	—	—	—	—
Chirp 2	0.275	0.178	1.5	0.728	0.516	1.4	0.693	0.406	1.7
Chirp 3	0.309	0.155	2.0	—	—	—	—	—	—
Whisper-large-v3	2.401†	3.654	0.7	—	—	—	—	—	—

way intersections: 314 of the 348 gender × dialect × domain intersections contain fewer than 30 utterances. Subgroups below 30 utterances or 10 interviews are suppressed (marked ^{s}) rather than reported; the released Somali dialect labels (Table 7) are pooled into the Northern Somali versus Benadiri Somali contrast for the same reason.

Table 6: Utterances / interviews per reported subgroup. ^{s} marks a subgroup suppressed from the analysis by the minimum-size rule; domain "Other(k)" pools the k smallest domain labels of that language.

Language	Gender (utt/interviews)	Dialect (utt/interviews)	Domain (utt/interviews)
Amharic	female 482/265; male 518/311	Addis Ababa 222/117; Shewa 85/42; Gojam 162/107; Gonder 87/55; Wollo 136/109; L2 196/91; Unknown 112/91	Social 146/110; Culture 125/59; Business 116/60; Health 112/67; Politics 104/58; Agriculture 98/80; Entertainment 89/44; Other(8) 210/114
Oromo	female 497/275; male 503/360	West Central 300/161; Eastern 372/240; Southern 328/219	Politics 144/110; Culture 109/71; Business 106/64; Social 103/67; Health 85/35; Education 81/49; Entertainment 80/55; Other(7) 292/189
Somali	female 477/150; male 523/198	Northern 525/170; Benadiri 291/106; Other 184/71	Politics 151/44; Social 133/48; Public Awareness 116/43; History 103/53; Entertainment 102/48; Other 100/49; Other(9) 295/113
Tigrinya	female 498/191; male 500/185; unknown 2/1^s	Z 542/206; L 239/101; D 219/89	Entertainment 222/86; Politics 174/68; Social 145/71; Culture 134/48; Health 118/42; Business 83/27; Other(7) 124/53

Table 7: Released dialect labels with utterance and interview counts, before pooling.

Language	Released dialect label	Utterances	Interviews
Amharic	Addis Ababa	222	117
Amharic	L2	196	91
Amharic	Gojam	162	107
Amharic	Wollo	136	109
Amharic	Unknown	112	91
Amharic	Gonder	87	55
Amharic	Shewa	85	42
Oromo	Eastern Oromo	372	240
Oromo	Southern Oromo	328	219
Oromo	West Central Oromo	300	161
Somali	Northern Somali (Somaliland)	355	131
Somali	Benadiri Somali	291	106
Somali	Other	184	71
Somali	Northern Somali (Ogaadeen)	86	31
Somali	Northern Somali (Puntland)	84	23
Tigrinya	Z	542	206
Tigrinya	L	239	101
Tigrinya	D	219	89

Per-dialect results. Table 9 gives WER per dialect subgroup with interview-clustered 95% intervals.

Table 8: Easiest and hardest dialect per system and language (WER, full normalization) and the spread between them. Somali is the Northern Somali versus Benadiri Somali contrast; Whisper spreads are over collapsed output.

Language	System	Easiest dialect (WER)	Hardest dialect (WER)	Spread
Amharic	MMS-1B-all	Addis Ababa (0.529)	Shewa (0.673)	0.143
	omniASR-LLM-7B	Addis Ababa (0.485)	Wollo (0.630)	0.145
	omniASR-CTC-7B	Addis Ababa (0.474)	Shewa (0.614)	0.141
	Chirp 2	Addis Ababa (0.230)	Shewa (0.332)	0.102
	Chirp 3	Addis Ababa (0.277)	Gojam (0.359)	0.082
	Whisper-large-v3	Shewa (2.129)	L2 (2.689)	0.560
Oromo	MMS-1B-all	Southern (0.553)	West Central (0.595)	0.041
	omniASR-LLM-7B	West Central (0.499)	Eastern (0.575)	0.076
	omniASR-CTC-7B	West Central (0.674)	Eastern (0.738)	0.064
	Chirp 2	West Central (0.641)	Southern (0.794)	0.152
Somali	MMS-1B-all	Benadiri (0.656)	Northern (0.667)	0.011
	omniASR-LLM-7B	Northern (0.481)	Benadiri (0.506)	0.025
	omniASR-CTC-7B	Northern (0.741)	Benadiri (0.765)	0.024
	Chirp 2	Northern (0.660)	Benadiri (0.764)	0.103
	Whisper-large-v3	Benadiri (0.929)	Northern (1.054)	0.125
Tigrinya	MMS-1B-all	Z (0.759)	D (0.855)	0.097
	omniASR-LLM-7B	Z (0.747)	D (0.865)	0.117
	omniASR-CTC-7B	Z (0.569)	D (0.791)	0.223

Table 9: WER per dialect subgroup (full normalization) with interview-clustered 95% confidence intervals; column headers give utterances/interviews.

Amharic

System	Addis Ababa n=222/117	Shewa n=85/42	Gojam n=162/107	Gonder n=87/55	Wollo n=136/109	L2 n=196/91	Unknown n=112/91
MMS-1B-all	0.529 [0.502, 0.556]	0.673 [0.642, 0.704]	0.634 [0.599, 0.670]	0.599 [0.564, 0.633]	0.623 [0.596, 0.653]	0.556 [0.528, 0.582]	0.527 [0.491, 0.567]
omniASR-LLM-7B	0.485 [0.438, 0.536]	0.628 [0.548, 0.726]	0.625 [0.575, 0.675]	0.544 [0.465, 0.619]	0.630 [0.577, 0.686]	0.513 [0.466, 0.564]	0.529 [0.466, 0.591]
omniASR-CTC-7B	0.474 [0.430, 0.515]	0.614 [0.544, 0.672]	0.539 [0.491, 0.589]	0.476 [0.417, 0.534]	0.529 [0.483, 0.578]	0.485 [0.436, 0.541]	0.421 [0.376, 0.468]
Chirp 2	0.230 [0.213, 0.247]	0.332 [0.285, 0.390]	0.312 [0.276, 0.350]	0.275 [0.242, 0.307]	0.308 [0.282, 0.335]	0.269 [0.245, 0.293]	0.249 [0.226, 0.280]
Chirp 3	0.277 [0.253, 0.301]	0.337 [0.292, 0.396]	0.359 [0.319, 0.402]	0.311 [0.276, 0.348]	0.348 [0.319, 0.380]	0.286 [0.261, 0.313]	0.276 [0.242, 0.315]
Whisper-large-v3	2.175 [1.994, 2.375]	2.129 [1.796, 2.613]	2.419 [2.204, 2.677]	2.308 [2.006, 2.682]	2.555 [2.285, 2.876]	2.689 [2.445, 2.969]	2.604 [2.327, 2.925]

Oromo

System	West Central n=300/161	Eastern n=372/240	Southern n=328/219
MMS-1B-all	0.595 [0.571, 0.619]	0.577 [0.554, 0.599]	0.553 [0.532, 0.575]
omniASR-LLM-7B	0.499 [0.477, 0.521]	0.575 [0.552, 0.601]	0.567 [0.539, 0.594]
omniASR-CTC-7B	0.674 [0.638, 0.708]	0.738 [0.710, 0.765]	0.714 [0.677, 0.748]
Chirp 2	0.641 [0.612, 0.671]	0.746 [0.701, 0.803]	0.794 [0.729, 0.873]

Somali

System	Northern n=525/170	Benadiri n=291/106	Other n=184/71
MMS-1B-all	0.667 [0.651, 0.685]	0.656 [0.632, 0.682]	0.665 [0.630, 0.699]
omniASR-LLM-7B	0.481 [0.459, 0.506]	0.506 [0.477, 0.539]	0.509 [0.467, 0.554]
omniASR-CTC-7B	0.741 [0.712, 0.767]	0.765 [0.734, 0.795]	0.751 [0.718, 0.783]
Chirp 2	0.660 [0.601, 0.740]	0.764 [0.632, 0.931]	0.668 [0.581, 0.818]
Whisper-large-v3	1.054 [0.987, 1.139]	0.929 [0.909, 0.958]	0.982 [0.925, 1.076]

Tigrinya

System	Z n=542/206	L n=239/101	D n=219/89
MMS-1B-all	0.759 [0.745, 0.773]	0.802 [0.786, 0.817]	0.855 [0.836, 0.873]
omniASR-LLM-7B	0.747 [0.717, 0.783]	0.819 [0.776, 0.870]	0.865 [0.817, 0.924]
omniASR-CTC-7B	0.569 [0.541, 0.598]	0.696 [0.663, 0.731]	0.791 [0.753, 0.827]

Normalization tiers and self-reported numbers. Table 10 reports every pair under the verbatim, full, and phon schemes, alongside the per-language number the system's own paper publishes where one exists (MMS reports FLEURS WER scored verbatim; the Whisper figure is for large-v2). The remaining systems publish only CER (language-averaged for USM; per-language on its own held-out data for Omnilingual ASR).

End-of-sentence scheme. Table 11 keeps sentence-end markers as tokens and reports terminal-marker precision, recall, and F1 on HORN-ASR and FLEURS.

Table 10: WER and CER under the verbatim, full, and phon normalization schemes, and the self-reported per-language WER where published.

			WER			CER			
Benchmark	Language	System	verbatim	full	phon	verbatim	full	phon	self-reported WER
Horn-ASR	Amharic	MMS-1B-all	0.632	0.582	0.564	0.247	0.228	0.222	—
Horn-ASR	Amharic	omniASR-LLM-7B	0.635	0.554	0.515	0.180	0.159	0.147	—
Horn-ASR	Amharic	omniASR-CTC-7B	0.565	0.501	0.469	0.222	0.205	0.196	—
Horn-ASR	Amharic	Chirp 2	0.344	0.275	0.230	0.126	0.107	0.096	—
Horn-ASR	Amharic	Chirp 3	0.396	0.309	0.268	0.159	0.138	0.128	—
Horn-ASR	Amharic	Whisper-large-v3	2.401	2.401	2.401	1.434	1.461	1.460	—
Horn-ASR	Oromo	MMS-1B-all	0.637	0.575	0.575	0.189	0.179	0.179	—
Horn-ASR	Oromo	omniASR-LLM-7B	0.627	0.548	0.548	0.182	0.172	0.172	—
Horn-ASR	Oromo	omniASR-CTC-7B	0.758	0.710	0.710	0.383	0.375	0.375	—
Horn-ASR	Oromo	Chirp 2	0.784	0.728	0.728	0.322	0.312	0.312	—
Horn-ASR	Somali	MMS-1B-all	0.716	0.663	0.663	0.280	0.260	0.260	—
Horn-ASR	Somali	omniASR-LLM-7B	0.585	0.493	0.493	0.205	0.181	0.181	—
Horn-ASR	Somali	omniASR-CTC-7B	0.791	0.750	0.750	0.401	0.384	0.384	—
Horn-ASR	Somali	Chirp 2	0.760	0.693	0.693	0.346	0.328	0.328	—
Horn-ASR	Somali	Whisper-large-v3	1.011	1.004	1.004	0.454	0.430	0.430	—
Horn-ASR	Tigrinya	MMS-1B-all	0.814	0.786	0.773	0.369	0.351	0.342	—
Horn-ASR	Tigrinya	omniASR-LLM-7B	0.821	0.784	0.772	0.401	0.385	0.379	—
Horn-ASR	Tigrinya	omniASR-CTC-7B	0.684	0.637	0.635	0.349	0.342	0.340	—
FLEURS	Amharic	MMS-1B-all	0.383	0.282	0.257	0.105	0.079	0.072	53.3
FLEURS	Amharic	omniASR-LLM-7B	0.292	0.177	0.143	0.079	0.051	0.043	—
FLEURS	Amharic	omniASR-CTC-7B	0.516	0.415	0.387	0.106	0.079	0.071	—
FLEURS	Amharic	Chirp 2	0.233	0.178	0.148	0.076	0.059	0.052	—
FLEURS	Amharic	Chirp 3	0.220	0.155	0.123	0.064	0.049	0.042	—
FLEURS	Amharic	Whisper-large-v3	3.656	3.654	3.654	2.294	2.360	2.359	140.3
FLEURS	Oromo	MMS-1B-all	0.760	0.665	0.665	0.209	0.189	0.189	—
FLEURS	Oromo	omniASR-LLM-7B	0.691	0.580	0.580	0.185	0.164	0.164	—
FLEURS	Oromo	Chirp 2	0.611	0.516	0.516	0.171	0.154	0.154	—
FLEURS	Somali	MMS-1B-all	0.538	0.443	0.443	0.164	0.133	0.133	53.8
FLEURS	Somali	omniASR-LLM-7B	0.472	0.351	0.351	0.148	0.116	0.116	—
FLEURS	Somali	Chirp 2	0.455	0.406	0.406	0.157	0.143	0.143	—

LM-decoding control. Table 12 reports MMS-1B-all on HORN-ASR Somali with Meta's released Common-Crawl Somali language model across LM weights α and word-insertion bonuses β.

F INTER-ANNOTATOR AGREEMENT STUDY

Design. For each language, 100 test utterances were independently transcribed from the audio alone by three annotators (from scratch, no feedback, no visibility of each other's work), and the gender, dialect, and domain labels of the same utterances were independently re-annotated. The study was run after the benchmark was frozen and did not alter any released reference.

Transcription agreement. Pairwise, direction-averaged WER between annotators, scored under the same homophone-aware phon scheme applied to systems, with bootstrap 95% intervals: Amharic 0.255 [0.239, 0.271]; Somali 0.348 [0.333, 0.363]; Tigrinya 0.424 [0.406, 0.443]; Oromo 0.475 [0.455, 0.497]. CER agreement is two to three times better (0.103 to 0.193), and normalization-aware scoring closes 15 to 24% of nominal WER disagreement: these languages have no enforced orthographic standard, so a large share of apparent disagreement is homophone-grapheme and segmentation convention rather than content. These numbers set the resolution limit of single-reference WER on this task: differences between systems below this level cannot be attributed to recognition quality alone, which motivates the multi-scheme, CER, and chrF reporting of the evaluation protocol. They are an agreement level among independent first passes, not a human-accuracy ceiling.

Reference validation. Scoring each annotator directly against the released references on the same items (389 of 400 bridged), the references sit closer to every annotator than annotators sit to each other for Amharic (0.226 versus 0.254), Oromo (0.434 versus 0.474), and Tigrinya (0.383 versus 0.423), evidence that the multi-round transcription-review-editing pipeline (Section 4.5) produced references more central than any single pass. Somali is the exception (0.348 versus 0.345): its references behave like one more independent pass.

Table 11: End-of-sentence scheme: WER with markers scored as tokens, and terminal-marker precision / recall / F1, on HORN-ASR and FLEURS.

Benchmark	Language	System	WER (full)	WER (EOS)	Marker P	Marker R	Marker F1
Horn-ASR	Amharic	MMS-1B-all	0.582	0.614	0.00	0.00	0.00
Horn-ASR	Amharic	omniASR-LLM-7B	0.554	0.571	0.00	0.00	0.00
Horn-ASR	Amharic	omniASR-CTC-7B	0.501	0.541	0.00	0.00	0.00
Horn-ASR	Amharic	Chirp 2	0.275	0.316	0.85	0.38	0.53
Horn-ASR	Amharic	Chirp 3	0.309	0.363	0.34	0.10	0.15
Horn-ASR	Amharic	Whisper-large-v3	2.401	2.193	0.00	0.00	0.00
Horn-ASR	Oromo	MMS-1B-all	0.575	0.570	0.00	0.00	0.00
Horn-ASR	Oromo	omniASR-LLM-7B	0.548	0.545	0.00	0.00	0.00
Horn-ASR	Oromo	omniASR-CTC-7B	0.710	0.707	0.00	0.00	0.00
Horn-ASR	Oromo	Chirp 2	0.728	0.724	0.67	0.35	0.46
Horn-ASR	Somali	MMS-1B-all	0.663	0.675	0.00	0.00	0.00
Horn-ASR	Somali	omniASR-LLM-7B	0.493	0.511	0.00	0.00	0.00
Horn-ASR	Somali	omniASR-CTC-7B	0.750	0.759	0.00	0.00	0.00
Horn-ASR	Somali	Chirp 2	0.693	0.704	0.47	0.38	0.42
Horn-ASR	Somali	Whisper-large-v3	1.004	1.004	0.44	0.73	0.55
Horn-ASR	Tigrinya	MMS-1B-all	0.786	0.799	0.00	0.00	0.00
Horn-ASR	Tigrinya	omniASR-LLM-7B	0.784	0.785	0.00	0.00	0.00
Horn-ASR	Tigrinya	omniASR-CTC-7B	0.637	0.661	0.00	0.00	0.00
FLEURS	Amharic	MMS-1B-all	0.282	0.330	0.00	0.00	0.00
FLEURS	Amharic	omniASR-LLM-7B	0.177	0.233	0.00	0.00	0.00
FLEURS	Amharic	omniASR-CTC-7B	0.415	0.447	0.00	0.00	0.00
FLEURS	Amharic	Chirp 2	0.178	0.178	0.98	0.99	0.99
FLEURS	Amharic	Chirp 3	0.155	0.164	0.84	0.83	0.84
FLEURS	Amharic	Whisper-large-v3	3.654	3.419	0.00	0.00	0.00
FLEURS	Oromo	MMS-1B-all	0.665	0.693	0.00	0.00	0.00
FLEURS	Oromo	omniASR-LLM-7B	0.580	0.633	0.00	0.00	0.00
FLEURS	Oromo	Chirp 2	0.516	0.524	0.95	0.97	0.96
FLEURS	Somali	MMS-1B-all	0.443	0.467	0.00	0.00	0.00
FLEURS	Somali	omniASR-LLM-7B	0.351	0.379	1.00	0.00	0.00
FLEURS	Somali	Chirp 2	0.406	0.396	0.96	0.98	0.97

Table 12: LM-decoding control: MMS-1B-all with the released Common-Crawl Somali LM on HORN-ASR Somali.

MMS-1B-all, Somali, Horn-ASR	WER (raw)	WER (full)	CER (full)
greedy (no LM)	0.716	0.663	0.259
+CC LM (alpha=0.3, beta=0.0)	0.684	0.627	0.310
+CC LM (alpha=0.5, beta=1.5)	0.690	0.635	0.316
+CC LM (alpha=1.0, beta=0.0)	0.863	0.841	0.408
+CC LM (alpha=1.5, beta=0.0)	0.953	0.944	0.435
+CC LM (alpha=2.0, beta=0.0)	0.980	0.975	0.445

Label agreement. Fleiss' κ over three annotators: gender 0.80 to 0.99 (highly reliable), the axis the balance and gender analyses rest on; dialect 0.40 to 0.78 (moderate), adequate for the aggregate subgroups reported in Section 6.5 but not for fine-grained claims; domain 0.32 to 0.44 (low), so domain-disaggregated results are presented as exploratory only. Somali dialect is scored on the binary Northern Somali versus Benadiri Somali contrast ($\kappa = 0.78$, against about 0.21 for free-form fine-grained labels); one annotator whom the annotation team had flagged as uncertain is excluded from the Somali dialect κ, and the gender and domain conclusions are insensitive to their inclusion. Pairwise Cohen's κ on sentence-end markers, which justifies the end-of-sentence scoring scheme of Section 6.7, is 0.72 to 0.75 for Amharic, 0.65 to 0.69 for Oromo, 0.47 to 0.62 for Somali, and 0.11 to 0.56 for Tigrinya, whose end-of-sentence numbers we therefore treat as indicative only.

G RESPONSIBLE-RELEASE DETAIL

G.1 TAKEDOWN PROTOCOL WITH REQUESTER VERIFICATION

Speakers, broadcasters, or authenticated representatives may request removal of specific content from the dataset.

- **Eligible requesters.** The speaker themself, the broadcaster of record for the source recording, or a documented authorized representative (legal counsel, family member with documentation, organizational representative with a verifiable letter).
- **Verification.** Requesters confirm identity via a challenge, a fact about the recording session recoverable only from access to the original broadcaster's records. A request that cannot be verified is not honored but is logged.
- **Public removal log.** A public log of all removals (count, date, language, and reason category, with no PII) is maintained alongside each versioned release. This makes selective scrubbing visible without exposing speakers.
- **Cadence.** Verified removals are processed in the next versioned release.

The verification + public-log requirement is what protects the takedown protocol from being weaponized: an actor seeking to suppress an opposing speaker's voice cannot succeed without producing the requester credentials a verified speaker has, and even an attempted suppression appears in the public log.

G.2 Minors and post-hoc speaker withdrawal

Minors. Speakers identifiable as under 18 at recording time are excluded at segmentation; the per-language editorial team flags suspected minor-speaker segments. If a minor speaker is identified post-release, the takedown protocol applies and a representative (parent or legal guardian) acts as the eligible requester.

Post-hoc withdrawal of contributor labor. Native-speaker contributors (collectors, transcribers, editors) may withdraw their labor contribution from the released dataset at any time, without justification. The released benchmark preserves contribution provenance internally (Appendix B); a contributor withdrawal triggers an independent re-edit pass on the affected segments by another contributor before the next versioned release. We document this right because contributor compensation alone does not cover the power asymmetry between the dataset team and individual contributors.

G.3 Maintenance plan

The dataset is versioned on the Hugging Face Hub. Each new version increments a `vX.Y` tag and is accompanied by a changelog documenting corrections, additions, and deprecations. We commit to:

- **Errata cadence.** Confirmed errata are published with a monthly review cadence; users can submit reports through the Hub's discussion / issue interface plus a dedicated email documented on acceptance.
- **Refresh releases.** The pipeline can be re-run on more recent broadcasts. We commit to publishing at least one refresh release within 24 months of v1, which gives the field a moving target rather than a frozen evaluation dataset.
- **Named point of contact.** A maintenance POC is named on acceptance.

Releasing data without a maintenance plan is a known failure mode for academic datasets (Paullada et al., 2021); we treat ongoing maintenance as part of the release.

H Duration Statistics

Per-language segment-duration statistics of the released benchmark (moved from the main text for space).

I Source / Outlet Inventory

The source-pool inventory comprises 3,897 distinct interviews: 1,004 Amharic, 1,013 Oromo, 958 Somali, and 922 Tigrinya. The interviews were collected via paid native-speaker contributors per language; sources are publicly disseminated speech recordings (open broadcaster channels,

Table 13: Duration statistics (seconds) per language.

Language	Mean	Median	Std	Min	Max
Amharic	15.1	11.7	8.9	5	30
Oromo	14.8	11.3	8.8	5	30
Somali	13.0	9.7	8.1	5	30
Tigrinya	12.6	9.0	8.1	5	30

community-radio postings, podcasts, and interview-format videos on online media) from a mix of broadcaster, podcast, and online-media outlets. Per-broadcaster, per-channel, and per-year caps were applied at acquisition to limit any single source's effective share. Per-segment outlet identifiers are hashed in the released metadata to limit indirect re-identification of contributing channels. The full broadcaster–channel–year inventory (including raw and post-cap effective-share counts) is released with the dataset.

J Dialect Typology and References

Amharic. The Amharic subset is labelled following the dialect classification of Mengistu Tadese (Tadesse, 2018), spanning the prestige Addis Ababa variety, Shewa Amharic, and the principal northern regional dialects Gojam, Gonder, and Wollo. To this we add an *L2* category for non-native speakers (e.g. native Oromo or Tigrinya speakers interviewed in Amharic), recognizing the language's lingua-franca status in Ethiopia, and an *Unknown* residual bucket for segments the annotators could not reliably classify rather than fold into a wrong regional bucket.

Oromo. Following the standard three-way classification of Oromo dialects (Stroomer 1995; Owens 1985; Griefenow-Mewis 2001; Janko 2009), the Oromo subset covers Mecha-Tulama (Western / Central), Hararghe (Eastern), and Borana-Arsi-Guji (Southern), the three dialect zones with the largest speaker populations. Smaller varieties such as Wallo Oromo and Raya are not represented.

Somali. Following the dialect classification of Lamberti (1986) and Saeed (1999), the Somali subset uses five buckets: three sub-varieties of Northern Somali (Somaliland, Ogaadeen in eastern Ethiopia, and Puntland) alongside Benadiri Somali (Mogadishu-area mainstream coastal) and an *Other* residual for diaspora and minor regional varieties. The Maay group (Lamberti, 1986; Tosco, 2012) is not a separate bucket; Maay-area speech falls under *Other*.

Tigrinya. The Tigrinya subset is labelled following the dialect classification of Haileslasie et al. (2023), which synthesizes three Tigrinya dialect categories, denoted Z, L, and D, from prior dialectology work (Mehari 2021; Mengesha 2009).

K Topical Domain Coverage

Table 14 reports the per-segment counts for each of the fifteen topical-domain labels across the benchmark (Section 5.1). The counts are totals over the four languages.

L Cross-Interview Sampling Pseudocode

The two-phase greedy joint-balance sampler used for the cross-interview down-selection (Section 4.6) is sketched below. Inputs: candidate pool $\mathcal{C}$ of QC-passed single-speaker segments per language (size $\sim$3,000); target budget $T = 1{,}000$ per language; per-speaker hard cap $k = 6$; soft female-share floor $\tau = 0.5$.

Phase 1: per-speaker cap. For each language, restrict the candidate pool to at most k segments per speaker, sorted by descending QC rating; drop the remainder.

Table 14: Topical-domain bucket counts across the benchmark (totals over the four languages).

Domain	Segments
Politics	573
Social	527
Entertainment	493
Culture	400
Health	371
Business	349
Technology	237
History	210
Education	164
Agriculture	163
Other	140
Public Awareness	130
Religion	118
Sport	73
Travel	52

Phase 2: greedy diversity-aware fill. Initialise an empty selection $\mathcal{S}$; let $f(\mathcal{S})$ denote the running female share of $\mathcal{S}$. At each step, choose the candidate $c \in \mathcal{C} \setminus \mathcal{S}$ that maximises an inverse-frequency joint-axis diversity reward

$$r(c, \mathcal{S}) = \sum_{a \in \{\text{dialect,region,domain,speaker}\}} \frac{1}{1 + \text{count}_a(c, \mathcal{S})},$$

restricted to female speakers when $f(\mathcal{S}) < \tau$; $\text{count}_a(c, \mathcal{S})$ is the number of segments in $\mathcal{S}$ that share c's value on axis a. Continue until $|\mathcal{S}| = T$.

Properties. The joint-axis reward is monotone submodular, with the classical $(1 - 1/e)$ approximation guarantee for the diversity term; the female-share floor is applied as a per-step constraint, not folded into the reward. The reference Python implementation, including unit tests against the L_1-distance termination property, is included in the evaluation-protocol repository linked from Appendix D.

M Internal Ethics Protocol

The project was assessed under an internal protocol with three components.

- **Source admission.** Only publicly disseminated speech recordings were admitted, paired with a contributor-supplied reason for inclusion (dialect markers, topical-domain signals, or place keywords). The three-condition naturally-occurring test (Section 3.1) was applied at admission.
- **Content exclusion.** Content containing explicit violence, hate speech, or other regulator-flagged material was excluded at acquisition by contributors and confirmed by the per-language final editors. Speakers identifiable as under 18 at recording time were excluded at segmentation (Appendix G.2).
- **Contributor compensation.** All contributors (collectors, transcribers, and editors) were paid hourly at a rate benchmarked against the highest reported local monthly salary for skilled office work in each contributor's region, divided by a standard work-month, and increased to reflect specialised linguistic expertise.

The full text of the protocol document is released as a separate artifact with v1.

N SOURCE POOL

The unconditioned source pool before joint-balance sampling (moved from Section 4.7 for space). *Pool-level speaker identifiers are proxies assigned at collection, not verified identities; the per-language values and the total were computed at different stages of collection and are not additive.

Table 15: Source pool from which HORN-ASR was curated. Reports the unconditioned distribution before joint-balance sampling.

Language	Utts	Hours	Speaker IDs*	%M	%F	Vocab
Amharic	14,655	61.5	551	69	31	88,715
Oromo	16,211	66.5	617	79	21	96,811
Somali	14,773	53.3	391	72	28	77,825
Tigrinya	13,649	47.9	459	61	39	80,254
Total	59,288	229.1	634	71	29	335,578

O FLEURS AUDIT: ANNOTATION METHODOLOGY

Table 16: FLEURS dev/test audit, three covered Horn of Africa languages. "Labeled rows" counts FLEURS rows for which a speaker ID was assigned (i.e. rows that are duplicate readings of an already-present prompt); "Unique segments" counts the distinct text prompts those rows cover; "# Spkrs" is the count of distinct speaker IDs; "Speakers" lists the IDs.

Language	Split	Labeled rows	Unique segments	# Spkrs	Speakers	Gender
Amharic	dev	170	72	3	s1, s2, s3	FEMALE
Amharic	test	394	175	3	s1, s2, s3	FEMALE
Oromo	dev	19	16	4	s1, s2, s3, s4	MALE
Oromo	test	41	36	5	s1, s2, s3, s4, s5	MALE
Somali	dev	432	149	6	s1, s2, s3, s4, s5, s6	MALE
Somali	test	223	82	7	s1, s2, s3, s4, s5, s6, s7	MALE

We manually annotated speaker IDs (s1, s2, ...) on each labeled FLEURS dev/test row for the three covered Horn of Africa languages. All three languages were double-annotated. The labeled subset corresponds to FLEURS rows that are duplicate readings of the same prompt: for Amharic this is the entire pair-overlap subset distributed in the source agreement file; for Oromo and Somali the TSV mirrors the full FLEURS split, with the unlabeled remainder consisting of single-reading rows that did not require speaker attribution. Per-split labeled-row counts, unique-segment counts, distinct speaker counts, and split-level gender are reported in Table 16.