AfriSpeech-MultiBench: A Verticalized Multidomain Multicountry Benchmark Suite for African Accented English ASR

Anonymous ACL submission

Abstract

Recent advances in speech-enabled AI, including Google's NotebookLM and OpenAI's speech-to-speech API, are driving widespread interest in voice interfaces across sectors such as finance, health, agritech, legal services, and call-centers in the global north and south. Despite this momentum, there exists no publicly available application-specific model evaluation that caters to Africa's linguistic diversity. We present **Afrispeech-MultiBench**, the first domain-specific evaluation suite for over 100 African English accents across 10+ countries and six application domains: Finance, Legal, Medical, General dialogue, Call Center, and Named Entities. We benchmark a diverse range of open, closed, unimodal ASR and multimodal LLM-based speech recognition systems using both scripted and unscripted conversation drawn from various open African accented English speech datasets. Our empirical analysis reveals systematic variation: open-source ASR excels in scripted contexts but degrades on noisy, non-native dialogue; multimodal LLMs are more accent-robust yet struggle with domain-specific named entities; proprietary models deliver high accuracy on clean speech but vary significantly by country and domain. Smaller models fine-tuned on African English achieve competitive accuracy with lower latency, a practical advantage for deployment. By releasing this benchmark, we empower practitioners and researchers to select voice technologies suited to African use-cases, fostering inclusive voice applications for underserved communities.

1 Introduction

Automatic Speech Recognition (ASR) has become a foundational technology across numerous domains. In customer-support environments, ASR powers real-time call routing, intent detection, and agent assistance, substantially reducing response times and improving user satisfaction (Wang et al., 2023). In healthcare, voice-enabled digital scribes transcribe clinician–patient interactions on the fly, alleviating documentation burdens and cutting downstream transcription costs (van Buchem et al., 2021). Emerging applications in legal transcription (Saadany et al., 2023), financial trading desktops, and live subtitling further demonstrate the broad impact of ASR systems in both enterprise and consumer settings.

Selecting the optimal ASR model for a given task now often means choosing among powerful, pre-trained *foundation* systems rather than training bespoke models from scratch. Self-supervised representations such as wav2vec 2.0 (Baevski et al., 2020) learn rich audio features from large amounts of unlabeled speech and can be applied in a zero-shot or few-shot manner, achieving near-state-of-the-art word-error rates on standard benchmarks (Baevski et al., 2020). Large multitask models such as Whisper (Radford et al., 2023), trained on hundreds of thousands of hours of multilingual and multitask data, exhibit strong zero-shot transfer across domains and languages without additional fine-tuning (Radford et al., 2023). However, computational budgets, latency requirements, and domain mismatches mean that one foundation model may outperform another depending on the target task, be it medical dictation, legal proceedings, or informal conversational speech.

Accented speech, particularly non-Western and under-represented varieties, remains a persistent blind spot in mainstream evaluation suites. African accents exhibit rich phonetic and prosodic diversity, which can dramatically widen word-error-rate gaps when compared to North-American or British English (Dossou, 2025). Without a dedicated benchmark, practitioners cannot reliably predict which off-the-shelf ASR system will meet accuracy or latency targets on their specific African-accented corpus.

Accordingly, we present a unified evaluation suite that benchmarks leading ASR systems, **AfriSpeech-MultiBench** in zero-shot mode across medical, legal, conversational, and named-entity-rich African-accented English. The suite supplies standardized test sets, and transparent scoring protocols enabling practitioners to compare models and select the architecture most appropriate for their target application or for finetuning.

2 Related Work

IrokoBench introduced a comprehensive text-based evaluation across seventeen low-resource African languages, revealing significant performance gaps between large language models and human competence on tasks such as natural-language inference, reasoning and question answering (Adelani et al., 2025). The study underscores the necessity of domain-specific evaluation: without targeted test suites, systematic deficiencies remain undetected.

Within automatic speech recognition (ASR), progress is often measured through the community-maintained Open ASR Leaderboard, which continuously reports word-error rate (WER) and real-time factor on LibriSpeech (Panayotov et al., 2015), TED-LIUM 3 (Hernandez et al., 2018), GigaSpeech (Chen et al., 2021), VoxPopuli (Wang et al., 2021), AMI (Carletta et al., 2005), Earnings22 (Andrew et al., 2022), SPGISpeech (Guo et al., 2022), and Common Voice (Ardila et al., 2020). Although these datasets cover a range of domains, from read audiobooks to meeting-room recordings, they remain dominated by North-American and British English, providing limited insight into performance on African-accented English.

Empirical investigations confirm the practical consequences of this imbalance. Koenecke et al. documented a twofold increase in WER for African American Vernacular English relative to Standard American English across multiple commercial recognisers(Koenecke et al., 2020). A global audit involving speakers from 171 birth countries observed the largest error rates for sub-Saharan participants(DiChristofano et al., 2022). In the absence of African-accented evaluation sets, leaderboard rankings therefore offer an incomplete picture for stakeholders on the continent.

Modern recognisers are architecturally diverse. They include multilingual encoders such as Whisper (Radford et al., 2023) and XLS-R, proprietary cloud services (Microsoft Azure Speech-to-Text, Google Speech-to-Text), Conformer-based systems like Canary (Puvvada et al., 2024) and Parakeet (Rekesh et al., 2023), Speech-Augmented Language Models (SALMs) (Chen et al., 2023), and multimodal architectures such as SeamlessM4T (Schwenk et al., 2023). Their heterogeneous training regimes and objectives complicate any attempt to infer accent robustness from results on existing benchmarks alone.

Several African-accented corpora have been released to mitigate data scarcity. AfriSpeech-200 provides roughly 200 hours of read speech from more than 100 indigenous accents (Olatunji et al., 2023). AfriSpeech-Dialog adds spontaneous two-speaker conversations (Sanni et al., 2025); AfriSpeech-Parliament captures parliamentary debates (Intron Health, 2025a); Med-Convo-Nig focuses on Nigerian clinical tele-consultations (Intron Health, 2025c); Afri-Names targets named-entity-rich prompts (Intron Health, 2025b); and AfriSpeech-Countries assembles cross-regional accents under consistent recording conditions (Intron Health, 2025). Existing baseline evaluations do not cover modern speech recognition systems or lack broad application-specific results.

This study contributes three key advances. First, six publicly available African-accented corpora are harmonised into AfriSpeech-MultiBench, an evaluation suite spanning medical, legal, conversational and named-entity-rich speech. Second, sixteen contemporary recognisers covering multilingual, proprietary, Conformer-based, SALM and multimodal architectures—are evaluated in zero-shot mode, with both WER and real-time factor reported. Third, a fine-grained error analysis disaggregates results by accent cluster, phonetic context and domain, elucidating systematic failure modes and informing future data collection and model selection.

3 Benchmark Methodology

3.1 Source Datasets

We assemble six corpora to form AfriSpeech-MultiBench, covering diverse Anglophone African English accents. The distribution of sources is shown in Table 1.

- **AfriSpeech-200**: (Afri) a 200-hour, 67,577 clip dataset, 2,463 speakers across 120 indigenous accents from 13 African countries, span-

Domain	**Data Source**	**Samples**	**Hours**	**Countries**	**Accents**	**Speakers**
Medical	Afri (clinical), Dialog (medical), Med.Conv	3651	29.88	10	95	519
General	Afri (general), Dialog (general)	2741	13.06	9	84	455
Legal	Parl	8068	35.86	4	–	–
Named Entities	Names (names)	3121	2.18	3	6	–
Finance	Names (numbers), Names (commands)	3186	6.73	4	9	–
Call Center	Call (Private)	16	0.80	2	3	32
Total Unique		**18042**	**75.45**	**11**	**108**	**859**

Table 1: Domain-wise breakdown of the Afrispeech-Multibench benchmark. Parentheses denote domain-specific subsets. Full names of the datasets - Afri:Afrispeech, Dialog:Afrispeech-Dialog, Med.Conv:Med-Conv-Nig, Names: AfriNames. The Call Center source is private and not disclosed.

ning clinical and general domain read speech (Olatunji et al., 2023).

- **AfriSpeech-Dialog:** (Diag) about 50 long-form medical and non-medical conversational sessions with African-accented spontaneous English (about 7 hrs) (Sanni et al., 2025).

- **AfriSpeech-Parliamentary:** (Parl) A real-world noisy, multi-speaker dataset of transcribed parliamentary speech (about 35.86 hours, 8,068 clips) sampled from Nigeria, Ghana, South Africa, and Kenya. (Intron Health, 2025a).

- **Med-Conv-Nig:** (Med.Conv) about 25 long-form simulated doctor–patient conversations capturing multispecialty clinical interactions in Nigeria, featuring both male and female speakers and rich in medical vocabulary — tailored for evaluating domain-specific ASR in healthcare settings (Intron Health, 2025c).

- **AfriNames:** (Names) A read-speech corpus with subsets focused on African names (Name), numbers (Nums), and voice commands (Commands), e.g. "transfer $500 to my HSBC account"; comprising 6,307 single-speaker samples (about 8.92 hours), enriched with named entities and number utterances, spanning 12 distinct accents across four countries, particularly suited for evaluating ASR performance on entity-rich transcription tasks (Intron Health, 2025b)

- **AfriSpeech-Countries:** A mixture of Afrispeech-200, Afrispeech-Parliamentary, Afrinames and North African accented speech samples (Ctry-NA), totaling approximately 67 hours and 21,581 clips. The dataset spans seven African regions and includes both read and conversational speech. All samples are annotated by domain and country.

- **Afro-Call-Centers:** (Call) A private unreleased dataset capturing real-world agent–customer voice interactions rich in domain-specific vocabulary across finance, health, and customer support domains (Intron Health, 2025).

Dataset	**Hrs**	**Speakers**	**Accents**
Afrispeech	18.68	750	108
Afri-Diag	7.00	98	12
Parl	35.86	–	4
Med.Conv	4.20	11	1
Names	8.91	–	12
Countries (NA)*	4.61	–	7
Total Unique	**79.26**	**859**	**108**

Table 2: Corpus statistics (Test). Countries (NA) represents speech samples from Northern African countries not included in other test sets which already have other African countries. Dashes represent statistics not provided in the original release of the datasets.

3.2 Domains Studied

We define six domain categories for evaluation with dataset details described in Table 1:

- **Medical:** health-related medical speech and clinician–patient dialogues.

- **General:** read-speech sourced from Wikipedia and unscripted multispeaker dialogues.

- **Legal:** noisy parliamentary proceeding with overlapping speech.

- **Finance:** read speech enriched with numbers such as currencies, decimals, dates, measurements, locations, trading volumes, and financial institutions.

- **Call Center / Customer Support:** real-world agent–customer interactions

- **Named-Entities:** Named-Entity-Rich General clips with dense mentions of African person names, locations, organizations, and dates

3.3 Models

Architecture	Model	Size
Conformer	Nvidia Parakeet-tdt-0.6B-v2	0.6B
	Nvidia Parakeet-tdt-1.1B	1.1B
	Nvidia Parakeet-rnnt-1.1B	1.1B
	Nvidia Canary-1B-flash	1B
Whisper Variant	OpenAI Whisper-large-v3	1.54B
	Distil-Whisper-v3.5	756M
	Nyra Health CrisperWhisper	1.54B
SALM	IBM Granite-3.3-2B	2B
	Mistral Voxtral-Mini-3B	3B
	Nvidia Canary-Qwen-2.5B	2.5B
	Microsoft Phi-4 MM-Instruct	14B
Proprietary	Intron-Sahara	–
	OpenAI GPT-4o Transcribe	–
	Google Gemini-2.0 Fl	–
	AWS Transcribe	–
	Microsoft Azure Speech	–

Table 3: Descriptions of evaluated models, including model size, core architecture, and provider. Model sizes are in billions (B) of parameters when known.

We evaluate 16 modern ASR systems partly sourced from the top twenty entries on the Hugging Face Open ASR Leaderboard (snapshot: July 2025)[1] categorized into model families representing architectural breadth—Conformer, RNN-T, CTC, transducer hybrids, and speech-augmented language models (SALMs) and include both fully open-source checkpoints and proprietary services already deployed in commercial workflows.

- **NVIDIA's open models:** Open-source ASR models based on the FastConformer (Rekesh et al., 2023) such as the Parakeet variants: CTC, RNN-T and TDT (Galvez et al., 2024) in sizes of 0.6B and 1.1B, and the 1 billion parameter Canary-flash model pairing a FastConformer encoder with a transformer decoder (Puvvada et al., 2024).

- **Whisper Variants:** Transformer encoder decoder models based on Whisper (Radford et al., 2023). We consider the variants: Whisper-large-v3 (Radford et al., 2023), Distil-Whisper-v3.5[2], and CrisperWhisper (Zusag et al., 2024).

- **Open SALMs:** Multimodal LLMs and Speech-Augmented LLMs including IBM Granite-3.3-2B[3], Phi-4 Multimodal Instruct (Abdin et al., 2024), Nvidia Canary-Qwen[4], and Mistral's Voxtral Mini-3B (Liu et al., 2025).

- **Proprietary cloud ASR services:** OpenAI's GPT-4o transcribe[5], Google's Gemini-2.0-flash[6], AWS Transcribe[7], Azure Speech Recognition[8] and Intron[9]. Models are evaluated in zero-shot mode, with neither demonstrations (Min et al., 2022) nor domain-specific fine-tuning.

This broad selection of modern ASR systems facilitate an empirical comparison between commercially deployed services and publicly available checkpoints, capturing the architectural and commercial diversity of leading ASR systems, providing a realistic basis for accent-aware model selection.

3.4 Evaluation Protocol

- Primary metric: Word Error Rate (WER) measured per model, per domain, per country, and per dataset.

- Error analysis: Breakdown by domain, accent group (native vs non-native), named-entity errors, noise robustness; Open-source vs proprietary models, unimodal vs multimodal, large vs compact variants.

4 Experiments

- Dataset splits: We use held-out test sets per corpus, ensuring some accents appear only in testing to evaluate zero-shot generalization

[1] Leaderboard URL: https://huggingface.co/spaces/hf-audio/open_asr_leaderboard.

[2] https://huggingface.co/distil-whisper/distil-large-v3.5

[3] https://huggingface.co/ibm-granite/granite-speech-3.3-2b

[4] https://huggingface.co/nvidia/canary-qwen-2.5b

[5] https://platform.openai.com/docs/models/gpt-4o-transcribe

[6] https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-0-flash

[7] https://aws.amazon.com/transcribe/

[8] https://azure.microsoft.com/en-us/products/ai-services/ai-speech

[9] https://www.intron.io/

Model	**Open ASR Benchmarks**							**AfriSpeech-MultiBench**					
	Lib-S	TED-3	Giga	VoxP	AMI	Earn22	SPGI	Afrispeech	Afri-Diag	Parl	Med.Conv	Names	Call
Parakeet-tdt-0.6B-v2	1.69	3.38	9.74	5.95	11.16	11.15	2.17	30.20	**11.23**	18.45	29.41	41.88	20.96
Parakeet-tdt-1.1B	**1.40**	3.59	9.52	5.49	15.87	14.49	3.16	28.45	15.14	27.14	29.98	45.66	25.26
Parakeet-rnnt-1.1B	1.45	3.83	9.89	**5.44**	17.01	13.94	2.93	28.18	15.08	26.75	30.59	46.70	28.93
Canary-1B-flash	1.48	3.12	9.85	5.63	13.11	12.77	1.95	29.77	48.50	19.13	93.62	44.10	88.71
Whisper-large-v3	2.01	3.86	10.02	9.54	15.95	11.29	2.94	26.49	13.49	19.99	31.76	43.23	24.69
Distil-Whisper-v3.5	2.37	3.64	9.84	8.04	14.63	11.29	2.87	27.58	18.00	**11.50**	30.41	45.80	21.65
CrisperWhisper	1.82	3.2	10.24	9.82	**8.71**	12.89	2.7	63.80	72.72	79.35	83.12	70.14	35.52
IBM Granite-3.3-2B	1.64	4.12	11.05	6.55	10.22	13.86	3.96	34.38	99.59	20.67	96.30	49.51	–
Voxtral (Mistral)	1.86	–	10.04	6.78	–	12.18	2.04	20.17	68.42	21.10	78.73	49.36	–
Canary-Qwen-2.5B	1.61	**1.90**	**9.43**	5.66	10.19	**10.45**	**1.90**	29.87	96.64	18.18	97.89	42.91	–
Phi-4 MM-Instruct	1.68	2.89	9.77	5.93	11.45	10.50	3.11	26.48	88.91	36.73	130.17	44.28	–
Intron-Sahara	–	–	–	–	–	–	–	**16.35**	14.26	15.41	27.92	**8.17**	**20.08**
GPT-4o Transcribe	–	–	–	–	–	–	–	24.66	15.03	64.39	30.80	52.49	23.20
Google Gemini-2.0 Flash	–	–	–	–	–	–	–	27.80	12.02	20.51	27.59	50.12	22.39
AWS Transcribe	–	–	–	–	–	–	–	32.77	14.02	18.50	30.08	36.70	23.51
Azure Speech Recognition	–	–	–	–	–	–	–	28.41	13.29	18.75	**26.17**	35.69	–

Table 4: Word Error Rate (WER %) for each model on standard open ASR benchmarks and subsets of the Afrispeech-MultiBench dataset. Dashes represent results that were not available. Full names of datasets: Lib-S: LibriSpeech; TED-3: TED-LIUM 3; Giga: GigaSpeech; VoxP: VoxPopuli; AMI: AMI Meeting Corpus; Earn22: Earnings22; SPGI: SPGISpeech; Afrispeech: AfriSpeech-200; Afri-Diag: AfriSpeech-Dialogue; Parl: AfriSpeech-Parliamentary; Med.Conv: Med-Conv-Nig; Names: AfriNames; Call: Afro-Call-Centers.

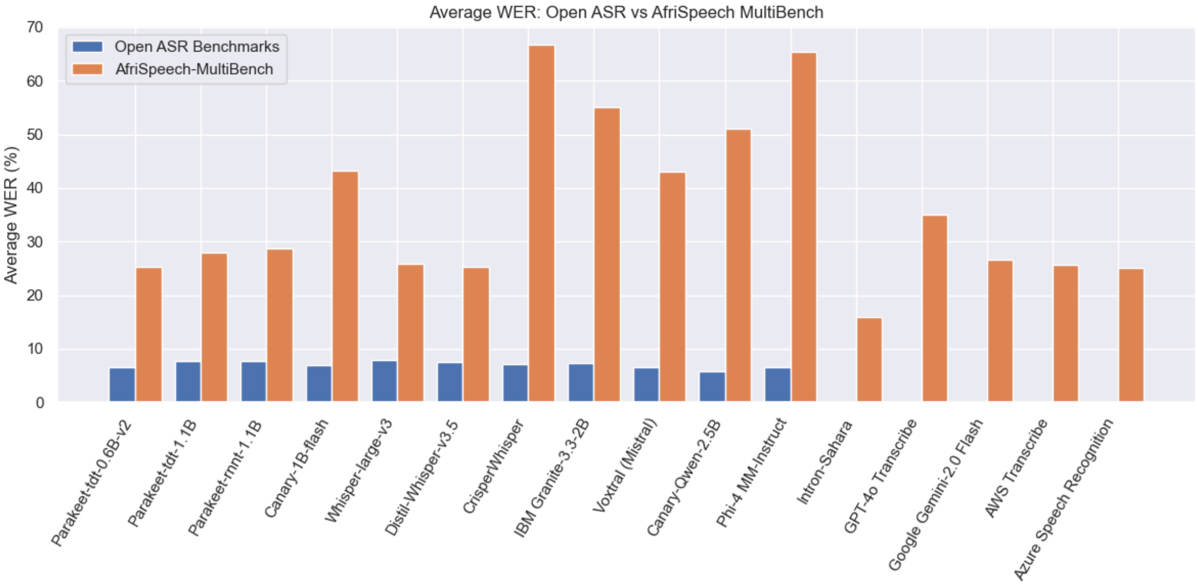

Figure 1: Average of Open ASR Leadearboard vs Afrispeech-Multibench

(e.g. 41 accents exclusively in test partition of AfriSpeech-200)

- Transcript Pre- and Post-processing: Model-specific transcript pre- and post-processing (described in Appendix section 7) normalized inputs, removed filler words, and mapped number words to their digit form, e.g. "twenty" to "20" and "first" to "1st".
- Inference setup: Uniform audio input preprocessing (16 kHz mono, no diarization) with default hyperparameters and decoding settings for ASR models and proprietary API calls. Local runs were on single T4 GPU (16GB).
- Prompting: We use consistent prompts for open and closed LLMs, e.g. "Transcribe this ENGLISH audio". Prompt details are provided in Appendix section 7.

We provide results for single runs.

5 Results

5.1 Overall Results

As show in Table 4, model performance on standard ASR benchmarks (e.g., LibriSpeech, TED-LIUM, AMI) fails to predict accuracy on African-accented, domain-specific speech. Leading open-source models like Parakeet-tdt-0.6B-v2 and Whisper-large-v3, which achieve WERs below 4% on LibriSpeech, degrade to 30–45% on general African speech and

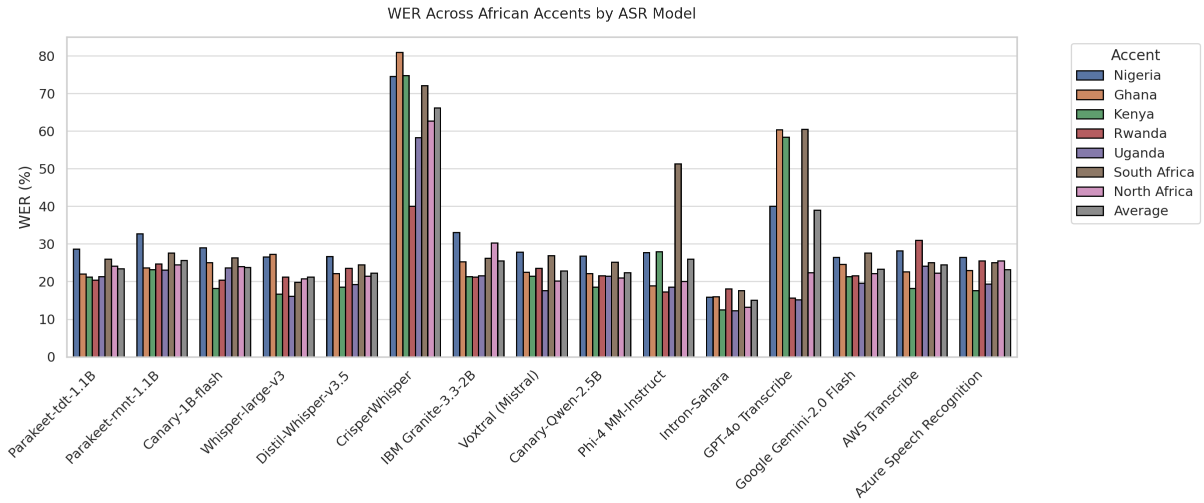

Figure 2: Word Error Rate (WER %) for each model across different African English accents in AfriSpeech-MultiBench. The average is computed across all listed accent categories.

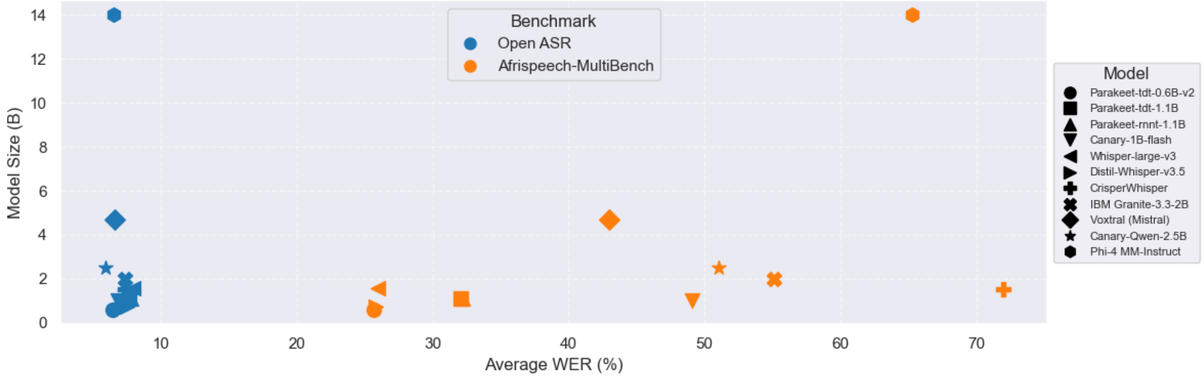

Figure 3: Model Sizes vs. Performance on Open ASR Benchmark (Blue) and Afrispeech-Multibench (Orange)

up to 70%+ on medical dialogue or named-entity-rich inputs in AfriSpeech-MultiBench. This pattern holds across architecture types, with all models showing 2–5× higher error rates on African data compared to leaderboard results. In contrast, Intron-Sahara, a regionally tuned model not featured on global leaderboards, consistently outperforms across domains—achieving 16.35% WER on general speech and just 8.17% on named entities.

5.2 Domain Performance

5.2.1 Medical

As show in Table 5, Intron-Sahara yields the lowest average WER (about 19.5%), significantly outperforming other models. Open models (Whisper-large-v3 and Parakeet-0.6Bv2) and proprietary (Gemini, GPT-4o, Azure) deliver average WERs of about 23–27% in these settings while Multimodal open LLMs like Phi-4 MM and IBM Granite perform poorly (>80% WER) in medical contexts despite taking top spots on the Open ASR Leaderboard.

5.2.2 Finance

As show in Table 6, Intron-Sahara excels here—achieving about 13.6% on numbers and about 1.8% on voice-commands, representing the finance domain. Proprietary models (Azure, AWS) perform moderately well (about 20–30% WER). Open-source and LLM models deliver higher error rates (about 40–55%).

5.2.3 Names

As show in Table 6, Intron-Sahara outperforms all others by a wide margin, reaching about 27% WER on African named entities. Open, larger, and proprietary models collapse with over 2x worse WERs (over 60%).

Model	Afri-Med	Diag	Med.Conv	Average
Parakeet-tdt-0.6B-v2	34.55	**11.23**	29.41	25.06
Parakeet-tdt-1.1B	33.79	15.14	29.98	29.98
Parakeet-rnnt-1.1B	33.45	15.08	30.59	30.59
Canary-1B-flash	34.77	72.23	78.92	78.92
Whisper-large-v3	32.59	17.22	31.76	27.19
Distil-Whisper-v3.5	32.18	16.77	30.63	26.53
CrisperWhisper	66.66	78.92	83.12	76.23
IBM Granite-3.3-2B	40.28	99.53	96.30	78.70
Voxtral (Mistral)	30.75	56.32	78.73	55.27
Canary-Qwen-2.5B	32.04	93.08	97.92	74.35
Phi-4 MM-Instruct	31.74	88.91	130.17	83.61
Intron-Sahara	**15.85**	13.44	29.10	**19.46**
GPT-4o Transcribe	28.54	15.03	30.80	24.79
Google Gemini-2.0 Flash	31.13	12.02	27.59	23.58
AWS Transcribe	42.22	14.02	30.08	28.77
Azure Speech Recognition	32.90	13.29	**26.17**	24.12
Average	34.59	39.51	53.83	42.89

Table 5: Word Error Rate (WER %) for each model on the medical domain subsets of AfriSpeech-MultiBench, including clinical notes, medical dialogues, and doctor–patient conversations. Dataset full name mappings: Afri-Med: Afrispeech Medical; Diag: AfriSpeech-Dialogue; Med.Conv: Med-Conv-Nig.

Model	Name	Commands	Nums
Parakeet-tdt-0.6B-v2	65.55	32.65	22.57
Parakeet-tdt-1.1B	76.44	33.67	26.47
Parakeet-rnnt-1.1B	75.78	35.36	26.66
Canary-1B-flash	75.69	30.05	20.15
Whisper-large-v3	73.1	31.58	18.11
Distil-Whisper-v3.5	68.15	37.28	15.28
CrisperWhisper	70.14	70.35	71.18
IBM Granite-3.3-2B	78.97	49.03	–
Voxtral (Mistral)	69.17	41.77	–
Canary-Qwen-2.5B	69.79	31.44	20.15
Phi-4 MM-Instruct	78.09	104.13	51.28
Intron-Sahara	**27.06**	**1.8**	**13.61**
GPT-4o Transcribe	67.43	46.67	17.45
Google Gemini-2.0-Fl	74.12	40.77	18.23
AWS Transcribe	60.07	27.60	20.21
Azure	67.15	23.42	22.43

Table 6: Word Error Rate (WER %) for each model on African named entites and Financial domain subsets of AfriSpeech-MultiBench. Dashes represent results that were not available.

5.2.4 Legal

Table 4 shows performance on the parliamentary dataset. Despite the high level of ambient noise and overlapping speakers in this dataset, Open Whisper variant (Distil-Whisper-v3.5, 11.5%) outperforms larger open and proprietary LLMs by a wide margin. Proprietary systems show slightly higher rates (about 21–27%) while Intron-Sahara (domain-tuned) obtains about 15.4% WER.

5.2.5 Call Center

Table 4 shows Intron marginally outperforming the conformer and whisper variants as well as proprietary ASR and LLMs providers on multispeaker call center conversations.

5.3 Accent and country variations

As shown in Table 7 and Figure 2, most models show pronounced degradation in Nigeria, South Africa, and Ghana (about 30%), relative to East and North Africa (about 24%). Most models perform comparably except GPT-4o and CrisperWhisper with WERs above 60%.

5.4 Model size vs performance

Figure 3 and Table 4 shows that, in a handfull of domains, larger SALMs (Granite, Phi-4, Voxtral, Canary-Qwen) only marginally outperform smaller architectures like conformer and Whisper variants half their size. In conversational speech, they are worse overall. Figure 3 indicates overall worse performance for open models with increasing size.

6 Discussion

This study yields a number of key insights that illuminate performance gaps and opportunities for advancing ASR systems in African settings:

6.1 Global benchmarks misrepresent African realities.

Leading models like Whisper and Parakeet achieve WERs below 10% on LibriSpeech and GigaSpeech, yet degrade to over 20–40% on African-accented data in AfriSpeech-MultiBench. This mismatch underscores the limits of current leaderboards in guiding ASR adoption across low-resource geographies.

6.2 Accent diversity drives large performance variance.

While models performed well on Kenyan and Ugandan English (average WERs as low as 12–18%), WERs doubled or tripled for West African and North African accents—exceeding 25% for many systems. This highlights the phonetic and prosodic diversity across the continent and the inadequacy of accent-agnostic training.

6.3 Conversational speech remains a major bottleneck.

Compared to read speech, performance worsened significantly on conversational corpora—AfriSpeech-Dialog Medical, Med Convo, and Parliamentary speech. These mirror Western benchmarks, where models also struggle on AMI and Earnings22 relative to LibriSpeech or SPGISpeech. However, the drop-off in African

Model	**Nigeria**	**Ghana**	**Kenya**	**Rwanda**	**Uganda**	**South Africa**	**North Africa**	**Average**
Parakeet-tdt-0.6B-v2	32.60	26.27	21.78	23.92	**9.38**	21.08	22.76	22.54
Parakeet-tdt-1.1B	28.65	22.08	21.21	20.39	21.35	25.96	24.13	23.40
Parakeet-rnnt-1.1B	32.76	23.69	23.19	24.71	23.08	27.59	24.42	25.63
Canary-1B-flash	29.00	25.06	18.25	20.39	23.65	26.30	24.04	23.81
Whisper-large-v3	26.53	27.22	16.70	21.18	16.16	19.85	20.72	21.19
Distil-Whisper-v3.5	26.69	22.16	18.51	23.53	19.22	24.45	21.43	22.28
CrisperWhisper	74.50	80.99	74.72	40.00	58.29	72.11	62.64	66.18
IBM Granite-3.3-2B	33.05	25.27	21.33	21.18	21.55	26.16	30.25	25.54
Voxtral (Mistral)	27.84	22.49	21.50	23.53	17.57	26.96	20.17	22.87
Canary-Qwen-2.5B	26.82	22.10	18.49	21.57	21.45	25.19	20.99	22.37
Phi-4 MM-Instruct	27.73	18.86	27.92	17.25	18.49	51.26	20.03	25.93
Intron-Sahara	**15.85**	**15.93**	**12.48**	18.04	12.26	**17.65**	**13.14**	**15.05**
GPT-4o Transcribe	40.03	60.41	58.38	**15.69**	15.22	60.43	22.40	38.94
Google Gemini-2.0 Flash	26.47	24.54	21.29	21.57	19.61	27.59	22.11	23.31
AWS Transcribe	28.16	22.59	18.18	30.98	24.11	25.05	22.23	24.47
Azure Speech Recognition	26.41	23.01	17.59	25.49	19.31	25.10	25.46	23.20
Average	31.44	28.92	25.72	23.09	21.29	31.42	24.81	26.67

Table 7: Word Error Rate (WER %) for each model across African accents in AfriSpeech-MultiBench, including the updated Parakeet-tdt-0.6B-v2 results.

conversational domains is more severe, revealing compound challenges likely due to accent, prosody, and domain shift.

6.4 Named entities and structured commands still confound models.

Most models scored above 40% WER on the AfriNames dataset, numbers, and financial voice commands, often failing to distinguish culturally unique or phonetically similar terms. This raises usability concerns in domains requiring accurate name capture or transactional integrity.

6.5 Model size and architecture don't predict reliability.

Smaller models like Parakeet-tdt-0.6B and Distil-Whisper sometimes matched larger peers on global benchmarks but showed inconsistent gains on African test sets. By contrast, Sahara—a regionally optimized model—consistently delivered best-in-class results across medical, legal, and conversational tasks.

6.6 Benchmarking must evolve beyond average-case accuracy.

AfriSpeech-MultiBench enables fine-grained, domain-aware evaluation that reflects real-world deployment conditions. It provides not only model ranking, but also insight into where and why systems fail—offering practical guidance for building domain- and region-specific ASR solutions in healthcare, law, finance, and public service delivery across Africa.

7 Conclusion

This study set out to address the gap between global ASR benchmarks and real-world performance on African-accented, domain-specific speech. Through AfriSpeech-MultiBench, we reveal that top-performing models on standard datasets like LibriSpeech and TED-3—achieving sub-5% WER—can exhibit 5–10× higher error rates on African speech, especially in medical, financial, and conversational domains. These disparities are consistent across open-source and proprietary systems, highlighting persistent geographic, linguistic, and domain biases in existing ASR development and evaluation pipelines.

Our findings underscore the need for regionally grounded benchmarks and models. Intron-Sahara, a model trained with African-specific data, consistently outperformed global leaders across domains and accents, particularly in name recognition, doctor–patient dialogue, and financial commands. By benchmarking 17 models across 8 African countries and 6 key domains, AfriSpeech-MultiBench provides actionable insights for building inclusive ASR systems. This work lays the foundation for future research and deployment efforts in healthcare, legal transcription, customer service, and multilingual voice applications across the African continent.

Limitations

While AfriSpeech-MultiBench offers a broad and diverse benchmark across African-accented English, several limitations warrant consideration. First, despite including over 10 countries and six domains, the benchmark does not yet cover all major linguistic regions in Africa or fully represent under-resourced countries with limited public data availability. Certain domains—such as manufacturing, education, and public safety—are not currently included, and even within included sectors like healthcare and finance, dataset sizes remain modest compared to global corpora, which may limit fine-grained error analysis and generalization of results.

Second, some datasets used are proxies rather than fully representative of their target verticals. For instance, parliamentary proceedings may not fully capture the legal domain's complexity, such as courtroom vernacular, legalese, or multilingual code-switching common in legal aid and judicial settings. Similarly, due to privacy constraints, customer support datasets from private call centers were not included, limiting direct benchmarking for commercial deployments. These gaps highlight both the urgent need and the opportunity for continued investment in domain-specific and geographically expansive data collection to build more comprehensive benchmarks for inclusive speech technologies.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024. Phi-4 technical report. *Preprint*, arXiv:2412.08905.

David Ifeoluwa Adelani, Jessica Ojo, Israel Abebe Azime, Jian Yun Zhuang, Jesujoba Oluwadara Alabi, Xuanli He, and 1 others. 2025. Irokobench: A benchmark for african languages in the age of large language models. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 2732–2757, Mexico City, Mexico. Association for Computational Linguistics.

Galen Andrew, Mingqing Chen, Jinyu Lu, and Kevin Sim. 2022. Earnings22: A 100-hour benchmark corpus for earnings-call ASR. In *Proceedings of Interspeech 2022*, pages 3158–3162.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, and 1 others. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, pages 4218–4222.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 12449–12460.

Jean Carletta, Simone Ashby, Séverine Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, and 1 others. 2005. The AMI meeting corpus: A pre-announcement. In *Proceedings of the Second International Conference on Machine Learning for Multimodal Interaction (MLMI)*, pages 30–44. Springer.

Chang Chen, Yiming Peng, Yuan Guo, Nanxin Yang, Shuai Zhang, Yongqiang Cui, and 1 others. 2021. Gigaspeech: An evolving, multi-domain ASR training corpus with 10,000 hours of audio. In *Proceedings of Interspeech 2021*, pages 3790–3794.

Zhehuai Chen, He Huang, Andrei Andrusenko, Oleksii Hrinchuk, Krishna Puvvada, Jason Li, and 1 others. 2023. SALM: Speech-augmented language model with in-context learning for speech recognition and translation. *arXiv preprint arXiv:2310.09424*.

Alex DiChristofano, Henry Shuster, Shefali Chandra, and Neal Patwari. 2022. Global performance disparities between english-language accents in automatic speech recognition. In *arXiv preprint arXiv:2208.01157*.

P. Dossou, Bonaventure F.2̇025. Advancing african-accented english speech recognition: Epistemic uncertainty-driven data selection for generalizable ASR models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, Bangkok, Thailand. Association for Computational Linguistics.

Daniel Galvez, Vladimir Bataev, Hainan Xu, and Tim Kaldewey. 2024. Speed of light exact greedy decoding for rnn-t speech recognition models on gpu. In *Interspeech 2024*, pages 277–281.

Cong Guo, Jing Zhang, Xiaohui Ma, Yongqiang Huang, Mike Lewis, Zhe Wei, and Shiliang Chen. 2022. SPGISpeech: 5,000 hours of transcribed financial audio for self-supervised speech representation learning. In *Proceedings of Interspeech 2022*, pages 3663–3667.

François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Estève. 2018. TED-LIUM 3: Twice as much data and corpus repartition for experiments on speaker adaptation. In *Speech and Computer (SPECOM)*, pages 198–208.

Intron Health. 2025. Afrispeech-countries: Cross-regional african-accented english speech benchmark. `https://huggingface.co/datasets/intronhealth/afrispeech-countries`.

Intron Health. 2025a. Afrispeech-parliament: Transcribed parliamentary sessions from four african nations. `https://huggingface.co/datasets/intronhealth/afrispeech-parliament`.

Intron Health. 2025b. Afri-names: African named-entity read-speech corpus. `https://huggingface.co/datasets/intronhealth/afri-names`.

Intron Health. 2025c. Med-convo-nig: Nigerian doctor–patient tele-consultation speech dataset. `https://huggingface.co/datasets/intronhealth/med-convo-nig`.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, and 1 others. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689.

Alexander H. Liu, Andy Ehrenberg, Andy Lo, Clément Denoix, Corentin Barreau, Guillaume Lample, Jean-Malo Delignon, Khyathi Raghavi Chandu, Patrick von Platen, Pavankumar Reddy Muddireddy, Sanchit Gandhi, Soham Ghosh, Srijan Mishra, Thomas Foubert, Abhinav Rastogi, Adam Yang, Albert Q. Jiang, Alexandre Sablayrolles, Amélie Héliou, and 87 others. 2025. Voxtral. *Preprint*, arXiv:2507.13264.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tobi Olatunji, Tejumade Afonja, Aditya Yadavalli, Chris Chinenye Emezue, Sahib Singh, Bonaventure F. P. Dossou, and 1 others. 2023. Afrispeech-200: Pan-african accented speech dataset for clinical and general domain asr. *Transactions of the Association for Computational Linguistics*, 11:1599–1617.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Krishna C. Puvvada, Piotr Żelasko, He Huang, Oleksii Hrinchuk, Nithin Rao Koluguri, Kunal Dhawan, Somshubra Majumdar, Elena Rastorgueva, Zhehuai Chen, Vitaly Lavrukhin, Jagadeesh Balam, and Boris Ginsburg. 2024. Less is more: Accurate speech recognition & translation without web-scale data. *arXiv preprint arXiv:2406.19674*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Dima Rekesh, Nithin Rao Koluguri, Samuel Kriman, Somshubra Majumdar, Vahid Noroozi, He Huang, Oleksii Hrinchuk, Krishna C. Puvvada, Ankur Kumar, Jagadeesh Balam, and Boris Ginsburg. 2023. Fast conformer with linearly scalable attention for efficient speech recognition. *arXiv preprint arXiv:2305.05084*.

Hadeel Saadany, Catherine Breslin, Constantin Orasan, and Sophie Walker. 2023. Better transcription of uk supreme court hearings. In *AI4AJ@ICAIL*.

Mardhiyah Sanni, Tassallah Abdullahi, Devendra Kayande, Emmanuel Ayodele, Naome Etori, Michael Mollel, and 1 others. 2025. Afrispeech-dialog: A benchmark dataset for spontaneous english conversations in healthcare and beyond. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Holger Schwenk, Loïc Barrault, Yu-An Chung, Francisco Guzmán, Juan Pino, and the Seamless Communication Team. 2023. Seamlessm4t: Massively multilingual and multimodal machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Marieke M. van Buchem, Hileen Boosman, Martijn P. Bauer, Ilse M. J. Kant, Simone A. Cammel, and

Ewout W. Steyerberg. 2021. The digital scribe in clinical practice: A scoping review and research agenda. *npj Digital Medicine*, 4:57.

Lingli Wang, Ni Huang, Yili Hong, Luning Liu, Xunhua Guo, and Guoqing Chen. 2023. Voice-based AI in call center customer service: A natural field experiment. *Production and Operations Management*, 32(4):1002–1018.

Weiyi Wang, Chau Tran, Fahim Azhar, Henrik Rottmann, Armand Joulin, and 1 others. 2021. Voxpopuli: A large-scale multilingual speech corpus for representation, semi-, and self-supervised learning. In *Proceedings of Interspeech 2021*, pages 993–997.

Mario Zusag, Laurin Wagner, and Bernhad Thallinger. 2024. Crisperwhisper: Accurate timestamps on verbatim speech transcriptions. In *Interspeech 2024*, pages 1265–1269.

Appendix

Pre- and Post-Processing

Audio pre-processing

Audio files are used exactly as distributed by the source datasets; no further segmentation or concatenation is performed. A single exception concerns the NVIDIA NeMo checkpoints (*parakeet-**, *canary-1B*), which require 16kHz mono input. When a file is multi-channel or sampled above 16kHz, it is down-mixed to mono and re-sampled with `sox` prior to inference. All other engines (Whisper variants, API endpoints) accept the original wave-forms without modification.

Transcript pre-processing

Reference and hypothesis strings undergo a three-stage normalisation pipeline, implemented exactly as in the public evaluation script:

1. `clean_text` — lower-cases, trims whitespace, removes punctuation, deletes 32 variants of *[inaudible]*, and removes frequent filler words (*uh*, *hmm*,*mmhmm*,...).
2. `text_to_numbers` — maps number words (*"twenty"* → 20) and ordinal words (*"first"* → 1st) to their digit form.
3. `EnglishTextNormalizer` — applies the Whisper normaliser for final case-folding and whitespace cleanup.

A sentinel token `abcxyz` replaces empty strings to avoid undefined denominators in word-error calculations.

Post-processing for Nemo models

NeMo/Parakeet outputs include automatically generated punctuation. Before the three-stage normaliser, inverse text normalisation is applied to restore standard spacing around commas and periods, ensuring a fair comparison with punctuation-free reference strings.

Metric

Word-error rate (WER) is computed with JIWER

$$\mathrm{WER}(r, h) = \frac{S + D + I}{|r|},$$

where S, D and I count substitutions, deletions and insertions needed to transform hypothesis h into reference r.

Prompting for Speech Augmented Language Models

Default prompts for open source speech augmented language models where used:

- Canary-Qwen-2.5B : "Transcribe the following: model.audio_locator_tag", "audio": ["speech.wav"]
- Mixtral (Voxtral-Mini-3B-2507): We used its apply_transcription_request function which takes an audio file and wraps it with inbuilt prompts for speech transcription.
- Google Gemini 2.0 Flash: The required prompt according to Google API documentation was used, prompt = """ Transcribe this ENGLISH audio. """
- Phi-4 Multimodal Instruct: <|user|><|audio_1|>Transcribe the audio to text<|end|><|assistant|>