# Afrispeech-Dialog: A Benchmark Dataset for Spontaneous English Conversations in Healthcare and Beyond

**Anonymous ACL submission**

## Abstract

Speech technologies are transforming interactions across various sectors, from healthcare to call centers and robots, yet their performance on African-accented conversations remains underexplored. We introduce Afrispeech-Dialog, a benchmark dataset of 50 simulated medical and non-medical African-accented English conversations, designed to evaluate automatic speech recognition (ASR) and related technologies. We assess state-of-the-art (SOTA) speaker diarization and ASR systems on long-form, accented speech, comparing their performance with native accents and discover a 10%+ performance degradation. Additionally, we explore medical conversation summarization capabilities of large language models (LLMs) to demonstrate the impact of ASR errors on downstream medical summaries, providing insights into the challenges and opportunities for speech technologies in the Global South. Our work highlights the need for more inclusive datasets to advance conversational AI in low-resource settings.

## 1 Introduction

Deep Learning approaches have revolutionized the field yielding significant performance gains across several natural language tasks (Bharadiya, 2023) especially for high-resource languages (Radford et al., 2023). While conversational speech recognition has continued to make significant strides in task automation in different domains such as medical (Biswas and Talukdar, 2024; Abacha et al., 2023; Yim et al., 2023), voice assistants (Pasandi and Pasandi, 2022; Mani et al., 2020), call center (Płaza et al., 2021), and robotics (Skantze, 2021), much of the research in Automatic Speech Recognition (ASR) has focused on monolingual speech with native accents (Aksënova et al., 2022), with considerable performance gaps in diverse linguistic, low resource, and accented contexts (Radford et al., 2023; Olatunji et al., 2023a).

In Anglophone African contexts, non-native accents further complicate the development of ASR systems resulting in poor generalization from state-of-the-art (SOTA) models (Hinsvark et al., 2021) in both general (Adelani et al., 2022; Olatunji et al., 2023a; Ogun et al., 2024; Tonja et al., 2024) and medical (Afonja et al., 2024; Olatunji et al., 2023b) contexts. Performant Speech recognition models are especially useful in overwhelmed health systems in developed countries where they can help to reduce documentation burden for overworked clinicians (Afonja et al., 2024; Olatunji et al., 2023b).

Recent advancements in medical conversation transcription and summarization using LLMs have led to wide adoption of these technologies in hospitals in developed countries (Michalopoulos et al., 2022; Yim et al., 2023) helping to improve clinical documentation quality and productivity (Galloway et al., 2024). Given the heavy patient loads (Baye et al., 2020) as well as the recent integration of AI/NLP/ASR systems into clinical workflows in African clinics [1,2], performant ambient transcription and summarization systems are highly desired. However, generalizability of such systems to African contexts remains underexplored (Afonja et al., 2024).

To tackle this problem, our contributions are as follows:

1. release AfriSpeech-Dialog a novel 7hr dataset of 50 long-form simulated English accented medical and non-medical conversations

2. benchmark the performance of SOTA speaker diarization, speech recognition, and LLM-based summarization on African accented conversations

[1] https://techcabal.com/2023/05/29/intron-health-brings-ai-to-african-healthcare/

[2] https://techtrendske.co.ke/2024/09/23/google-supports-jacaranda-healths-use-of-gen-ai-for-mothers-and-babies/

laying the groundwork for more inclusive ASR and NLP technologies in healthcare and beyond.

## 2 Related Works

### 2.1 ASR in Medical Conversations and Summarization

The role of ASR in medical documentation has grown significantly, particularly in telehealth and in-person patient-physician consultations (Korfiatis et al., 2022; Galloway et al., 2024; Michalopoulos et al., 2022; Yim et al., 2023). Accurate ASR in medical dialogue is critical, as transcription errors can lead to incorrect medical records. Several datasets have been developed to facilitate the study of ASR in medical contexts. PriMock57 (Korfiatis et al., 2022) provides primary care mock consultations in European English accents; Fareez et al. (2022) evaluates medical ASR on simulated patient-physician medical interviews with a focus on respiratory cases; Enarvi et al. (2020) experiments summarize ASR diarized doctor-patient ambulatory orthopedic surgery conversations, automatically generating medical reports. Le-Duc et al. (2024) proposed a real-time speech summarization system for medical conversations with the ability to generate summaries for every (local) and end of (global) utterances, eliminating the need for continuous update and revision of the summary state.

### 2.2 Non-medical conversational ASR

Pęzik et al. (2022) released DiaBiz, an Annotated Corpus of over 400hrs of Polish call center dialogs. Other conversational, parliamentary, or oratory datasets like AMI (Carletta et al., 2005), Earnings22 (Del Rio et al., 2022), Voxpopuli (Wang et al., 2021) have gained popularity on public ASR benchmarks. Conversational ASR has also been explored in other domains such as call centers (Płaza et al., 2021), and robotics (Skantze, 2021).

### 2.3 African Accented ASR

There has been growing interest in developing ASR systems that cater to African languages; for example, (Yılmaz et al., 2018) developed a multilingual ASR system for code-switched South African speech, East African ASR (Elamin et al., 2023). Multilingual ASR systems, such as EVI dataset (Spithourakis et al., 2022), offer a strong foundation for developing similar models in African contexts where data scarcity hinders progress. Olatunji et al. (2023b) released a pan-African accented English dataset for medical and general ASR.

### 2.4 Speaker Diarization in Multi-Speaker Conversations

To increase the efficiency of NLP/ASR systems, enormous contributions were made to researching the integration of SD models and methods into its pipeline. Serikov et al. (2024) provides a comparative analysis of SD models - Pyannote (Bredin et al., 2020), CLEVER[3], and NVIDIA NeMo (Harper et al., 2019) with their hyper-parameters choices and segmentation strategies on 20 different German dialects for diarization and identification (DI) task. However, with competitive performance, NeMo performs slightly better due to its multiscale segmentation concepts for identifying and removing shorter segments. On a similar DI task, (Chua et al., 2023) also benchmarked the performance of multilingual ASR models in open and closed tracks on the challenging MERLIon CCS English - Mandarin datasets - an extracted spontaneous and codeswitching parent-child conversation speeches.

## 3 Methodology

Figure 1 shows an overview of the dataset creation and benchmarking process.

### 3.1 Dataset

In this section, we outline our dataset creation process

#### 3.1.1 Collecting Conversations

We recorded simulated virtual and in-person medical and non-medical conversations from African medical and non-medical crowdworkers on the Intron Platform [4] similar to the process described in Olatunji et al. (2023b). Each conversation began with both speakers giving consent to record. Because real names were sometimes spoken in the consent segment, the first 10-15 seconds were removed to protect the identities of participants.

For medical conversations, following the process described in Fareez et al. (2022) and Korfiatis et al. (2022), a group of clinical experts prepared "patient cards" with disease conditions and demographics relevant to the African context. Doctor and Patient actors included medical professionals (e.g. doctors, nurses) familiar with Objective Structured Clinical Examinations (OSCE), a widely used assessment

[3] https://www.oxfordwaveresearch.com/products/cleaver/
[4] https://speech.intron.health

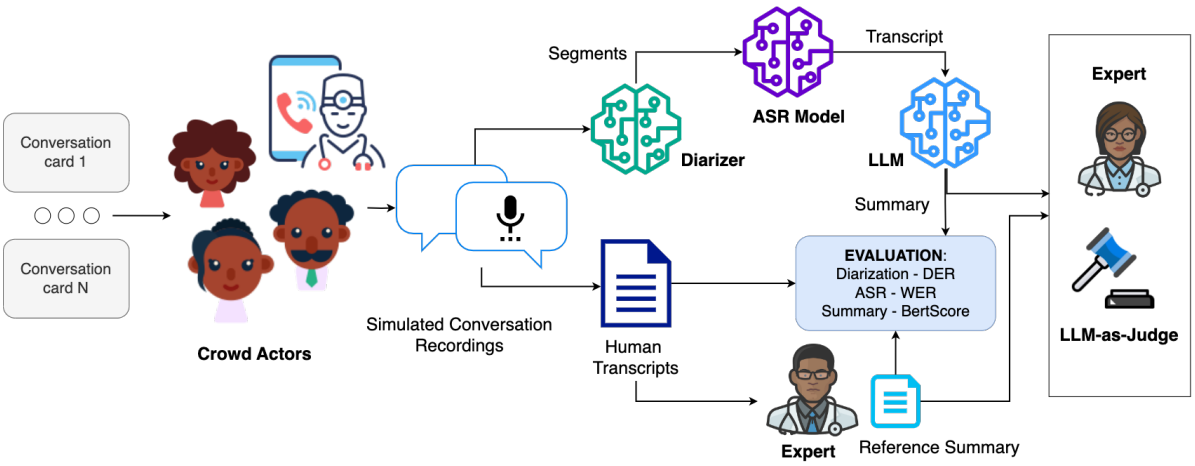


Figure 1: AfriSpeech Dialog: Dataset and Benchmarking Pipeline

in medical education that simulates doctor-patient interactions (Fareez et al., 2022). Each patient actor was provided with a detailed "patient card" that included information on their condition, demographics, and medical history, as shown in Table 1. Consistent with OSCE format, patient cards were hidden from doctor actors to faciliate a more natural consultation.

For general domain conversations, participants engaged in open discussions based on "topic cards" prepared by a team of reviewers. Each card contained a conversation topic, a brief description, and two discussion prompts to guide the conversation. The pair of participants (actors) had prior access to the cards and were advised to read through and understand them before starting the conversation. Table 2 shows a sample topic card.

The conversation recordings were stored as mono-channel, 16-bit wav files, with a 48 kHz sampling rate. A team of clinician reviewers reviewed the conversations and selected a high quality subset for this release. The dataset was collected across three African countries—Nigeria, Kenya, and South Africa. The speakers represent a diverse range of accents, spanning 11 accents: Hausa, Isoko, Idoma, Urhobo, Ijaw, Yoruba, Swahili, Sesotho, Igbo, Igala, and Ebira.

#### 3.1.2 Recording Characteristics

Dataset statistics are summarized in Table 3. In both medical and general domains, the dataset features two speakers in each conversation. Medical conversations were more structured, with doctors asking direct questions and patients providing responses, resulting in a more formal exchange. General conversations were more relaxed, with spontaneous discussions on various topics.

Overlapping speech occurs occasionally but is usually brief, involving short interjections like "yes" or "okay." Disfluencies and some code-switching reflect the natural flow of non-native English speakers. General conversations have fewer speaker turns but a higher average word count compared to medical ones, as speakers tend to talk longer to express their thoughts. These characteristics make the dataset valuable for testing speaker diarization and transcription models on African-accented speech.

#### 3.1.3 Transcription Process

All conversations were manually transcribed by professional transcribers. They were instructed to annotate speaker turns, and insert timestamps for each turn. Given the medical nature of some conversations, all transcribers were required to be familiar with medical terminology. Transcripts were then sent to a team of clinicians who created reference summaries for each transcript.

Contributors and annotators were paid $10 - $20 per hour per task depending on task complexity and clinical experience. Contributors considered payment satisfactory. The Dataset is released under a CC-BY-NC-SA 4.0 License.

### 3.2 Speaker Diarization

To benchmark diarization performance on this dataset, we selected three recent high-performing neural diarization models:

- **Pyannote** (Bredin, 2023): This model leverages a pre-trained neural network that com-

| | |
|---|---|
| **Condition** | Malaria |
| **Demographic (Age, Gender)** | 32-year-old Female |
| **Presenting Complaint** | Fever and chills (2 days) |
| **History of Presenting Complaint** | - Fever x 2 days (High grade, not relieved by medication)<br>- Chills x 2 days (Intermittent, severe)<br>- Headache x 2 days (Generalized, throbbing, 7/10 in severity)<br>- Fatigue and general body weakness<br>- No cough, diarrhea, vomiting, or urinary symptoms<br>- Patient lives in malaria-endemic area; no recent travel history |
| **Past Medical History (PMH)** | No chronic disease or surgery |
| **Family History** | No family history of similar illness |
| **Social History** | Does not drink alcohol or smoke |
| **Allergy History** | No known allergies |

Table 1: Example Patient Card for Medical Conversations

| **Topic: Cyberbullying** |
|---|
| **Overview:** Cyberbullying is a form of bullying that takes place online or through electronic communication. It involves using technology (social media, text messages, online forums) to intimidate or humiliate someone. Examples include spreading rumors, sending hurtful messages, and sharing embarrassing information without consent. |
| **Discussion Prompts:**<br>1. What steps do you take to protect yourself from cyberbullying?<br>2. Do you think social media has effective policies in place, or could they improve? |

Table 2: Sample Conversation Card for General Conversations

| | **Medical** | **General** |
|---|---|---|
| **Counts** | 20 | 29 |
| **Timestamped Counts** | 9 | 21 |
| **Avg. Num. of Turns** | 78.6 | 30.55 |
| **Total Duration (hours)** | 2.07 | 4.93 |
| **Avg. Word Count** | 725.3 | 1356.83 |
| **Num. of Countries** | 1 | 3 |
| **Num. of Accents** | 6 | 8 |
| **Genders (M,F)** | (14, 26) | (25, 33) |

Table 3: Statistics of the medical and non-medical datasets.

putes x-vector embeddings for speaker diarization. It uses an ECAPA-TDNN (Desplanques et al., 2020) architecture for speaker embeddings, shown to improve speaker separation in diarization tasks.

- **Reverb diarization v2** (Bhandari et al., 2024): This model is an extension of Pyannote 3.0, fine-tuned on 26,000 hours of annotated speech. However, this model uses WavLM (Chen et al., 2022) instead of the SincNet (Ravanelli and Bengio, 2018) features in the Pyannote 3.0 basic model.

- **Titanet**: This Nvidia diarization pipeline uses MarbleNet (Nguyen-Vo et al., 2022) for voice activity detection and Titanet-L (Koluguri et al., 2022) for embedding extraction. Titanet uses a 1D depth-wise separable convolutions with Squeeze-and-Excitation (SE) layers with global context, followed by a channel attention-based statistics pooling layer to map variable-length utterances to fixed-length embeddings (t-vectors) (Koluguri et al., 2022).

#### 3.2.1 Evaluation Metrics

We evaluated diarization performance using the **Diarization Error Rate (DER)** (Doddington et al., 2000). DER quantifies the percentage of time that speakers are misattributed or missed in the diarization output.

We used the *pyannote.metrics* library to calculate DER for each recording and computed the absolute DER for the entire dataset. The optimal mapping between the reference and hypothesis labels was obtained using the Hungarian algorithm (Kuhn, 1955),

ensuring an accurate alignment. We report DER on medical and general domain conversations.

### 3.3 Automatic Speech Recognition (ASR)

To benchmark ASR performance, we compare SOTA open source pretrained models, Whisper (Radford et al., 2023), Distil-Whisper (Gandhi et al., 2023), Nvidia Parakeet (Harper et al., 2019), Canary (Puvvada et al., 2024), MMS (Pratap et al., 2024) and Wav2vec2 (Baevski et al., 2020).

Open-source models offers transparency and reproducibility. They are often trained on diverse, real-world speech data, provide competitive baselines for dialog-specific tasks such as handling speaker variation, spontaneous speech, and overlapping speakers.

#### 3.3.1 Preprocessing

The original transcripts contained timestamps and speaker tags. We removed these items from the text as they are unnecessary for the ASR task. Long form audio recordings exceeded the context length of most ASR models. They were therefore chunked into 30 second segments for inference and transcript segments returned were concatenated.

#### 3.3.2 Evaluation Metrics

ASR performance was evaluated using **Word Error Rate (WER)**. WER measures the total number of insertions, deletions, and substitutions in the predicted text with respect to the total number of words in the reference text.

### 3.4 Medical Conversation Summarization

Nine LLMs (large, small, open, closed, general, and biomedical) were selected for summarizing doctor-patient dialogues. Each LLMs was presented a human conversation transcript and was prompted (Appendix section D) to generate a detailed summary.

**Closed-source general LLMs:** OpenAI (GPT-4o, GPT-3.5-turbo) (Achiam et al., 2023), and Anthropic Claude-3-Sonnet(Anthropic, 2023) represent leading general-purpose commercial LLMs.

**Open-source small general LLMs:** Meta-Llama-3.1-8B-Instruct, Meta-Llama-3.2-3B-Instruct (Dubey et al., 2024), microsoft-Phi-3-mini-4k-instruct (Abdin et al., 2024), and google-gemma-2-9b-it (Team et al., 2024) selected for their instruction-following abilities or multilingual support, which is essential for code-switching.

**Biomedical open-source LLMs:** m42-health-Llama3-Med42-8B(Christophe et al., 2024), OpenMeditron-Meditron3-8B(Chen et al., 2023) selected for their biomedical adaptation.

Examples of generated summaries are provided in Appendix section E.

#### 3.4.1 Evaluation

We evaluate summarization on the Afrispeech-Dialog medical subset using 3 complementary approaches.

#### 3.4.2 Quantitative Evaluation

We used the BERTScore (Zhang et al., 2019) to evaluate the quality of the LLM-generated summaries against the expert-generated reference summaries. Although BERTScore is widely used, studies like (Hanna and Bojar, 2021) have shown its limitations, particularly in capturing fine-grained semantic nuances and penalizing stylistic differences.

#### 3.4.3 Qualitative Evaluation

To address the limitations of BERTScore, we complement it with 2 qualitative evaluation approaches (Human and LLM-as-Judge) where summaries were evaluated on a 5-point scale (adapted from prior work Zheng et al., 2023; Liu et al., 2023) , 1 (worst) through 5 (best) on the following six criteria: recall of the diagnosis, accuracy of the treatment plan, avoidance of false or fabricated information, clarity and structure, and inclusion of important positive and negative clinical details. If any criterion (e.g. treatment plan) was absent from in the conversation (transcript), a score of 0 was to be assigned for that criterion. Detailed evaluation criteria can be found in the Appendix Section A.

**LLM-as-Judge** Consistent with the growing trend in recent studies (Zheng et al., 2023; Liu et al., 2023) we used generative models for automated summary evaluation. We used the OpenAI's "o1" model (Temsah et al., 2024) prompting based on the criteria mentioned above. Verbatim prompts can be found in the Appendix Section B.

**Human Evaluation** In a blind study, we randomly present pairs of human transcripts and LLM- or human-generated summaries to a team of 4 clinical experts. The experts compared the information available from summaries to conversation transcripts using the 6 criteria listed above (Kanithi et al., 2024; Singhal et al., 2023; Wang et al., 2023).

Each summary was independently rated by 2 experts.

#### 3.4.4 Error Propagation from Cascading Models

Since real-world conversation summarization systems rely on imperfect ASR transcripts and accented medical ASR transcription is challenging for several ASR systems (Afonja et al., 2024), we further evaluated summaries generated based on predicted (machine) transcripts to determine if there was a drop in quality when compared with summaries generated based on human transcripts (Giuseppe et al., 2021). We measured summary Quality using LLM-as-Judge.

## 4 Experiments

### 4.1 Diarization

We download and run inference using publicly released checkpoints from HuggingFace (Wolf et al., 2020), with default hyperparameters. We set the collar to 0.0, meaning no margin was allowed around speaker transitions, ensuring that even short overlaps (e.g., "yes" or "okay") were evaluated directly without any tolerance. Overlapping speech was also not excluded from the evaluation.

We ran inference on a single Nvidia T4 GPU. Inference for Pyannote and Reverb took approximately 2 hours while the Titanet took about 30 minutes. Results represent single runs.

### 4.2 ASR

Models were downloaded from publicly available huggingface (Wolf et al., 2020) checkpoints with default hyperparameters and the default generation configuration was used. We ran inference on Nvidia T4 GPUs. The inference required an average of around 30 minutes for the whole dataset for the open-source models. Results represent single runs.

### 4.3 Summarization

For open-source LLMs, we used publicly available checkpoints from HuggingFace (Wolf et al., 2020) without altering their default hyperparameters, except for setting max_new_tokens to 1024. Closed-source models were accessed via their respective APIs, also using default hyperparameters. The prompt template was adapted from prior work (Zheng et al., 2023; Liu et al., 2023), and to ensure consistency, the same prompt was used across all models (details can be found in the Appendix).

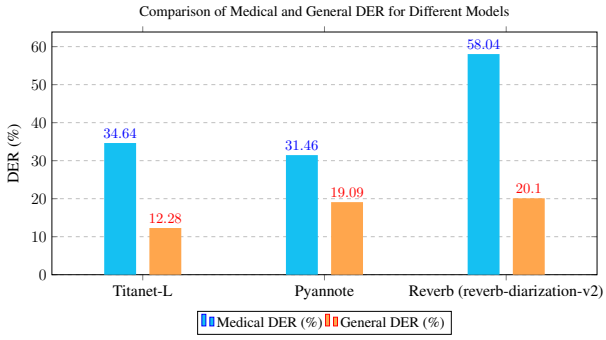


Figure 2: Comparison of Medical and General DER for Different Models

We conducted the summarization experiments under two scenarios: (1) generating summaries from human-produced transcripts and (2) generating summaries from transcripts created by our best-performing ASR model (Whisper-large-v3).

## 5 Results and Discussion

### 5.1 Diarization

We compute DER separately for a subset of conversations with accurate timestamps– medical (9 samples, Med DER) and general (21 samples, Gen DER) domain conversations. The results are shown in Table 4 and Figure 2. We also show the performance of these models on conversational datasets on other accents, using the values reported in (Bredin, 2023; Koluguri et al., 2022; Landini et al., 2022).

The models consistently performed better on general domain conversations compared to medical conversations, likely due to their relaxed structure and fewer interruptions. Results show diarization results on Afrispeech-Dialog are better than AMI and DIHARD likely because of the simulated and structured nature of conversations.

### 5.2 ASR

We report WER for all conversations, as well as separately for the medical and non-medical portions of the data (Table 5). The **openai/whisper-large** models achieved the best performance, followed by **nvidia/canary**, **nvidia/parakeet** and models and lastly **wav2vec2** based models.

The results, as presented in Table 5 and Appendix Figure 3, demonstrate a clear trend: models exhibit superior performance on non-medical (general domain) audio compared to medical domain audio, roughly 5% better WER, suggesting medical conversations are more challenging likely

| Model | DER (%) | Med DER (%) | Gen DER (%) | AMI DER (%) | DIHARD DER(%) |
|---|---|---|---|---|---|
| Titanet-L | **16.27** | 34.64 | **12.28** | (1.89) | - |
| Pyannote | 21.30 | **31.46** | 19.09 | 24.8 (4.6) | 34.4 |
| Reverb | 26.87 | 58.04 | 20.10 | - | - |

Table 4: Diarization Error Rate (DER) for all 30 audios, with detailed results for the Medical (Med. DER) and General (Gen. DER) subsets. The AMI DER and DIHARD DER columns show performance on the AMI MixHeadset (Carletta et al., 2005) and DIHARD II (Ryant et al., 2019) datasets, respectively. Lower DER is better, and (*) indicates results where overlapping speech regions were ignored.

| Model | WER | Med WER | Gen WER | AMI | Earnings22 | VoxPop |
|---|---|---|---|---|---|---|
| whisper-medium | 21.27 | 26.49 | **19.47** | 16.6 | - | 7.4 |
| whisper-large-v2 | **20.82** | **23.74** | 19.81 | 16.4 | 12.05 | 7.3 |
| whisper-large-v3 | **20.38** | **23.81** | **19.19** | 16.01 | **11.3** | - |
| whisper-large-v3-turbo | 21.93 | 25.58 | 20.67 | - | - | - |
| distil-whisper-large-v2 | 25.38 | 30.43 | 23.63 | **14.67** | 12.19 | 8.24 |
| distil-whisper-large-v3 | 21.20 | 25.67 | 19.58 | 15.95 | **11.29** | 8.25 |
| parakeet-rnnt-1.1b | 28.16 | 34.03 | 26.13 | 17.1 | 15.15 | **5.44** |
| parakeet-ctct-1.1b | 28.97 | 34.16 | 27.19 | 15.67 | 13.75 | 6.56 |
| parakeet-tdt-1.1b | 28.69 | 33.57 | 27.01 | 15.9 | 14.65 | **5.49** |
| canary-1b | 22.82 | 27.40 | 21.25 | **13.53** | 12.05 | 5.79 |
| wav2vec2-large-960h | 86.34 | 88.35 | 81.17 | 37.0 | - | 17.9 |
| mms-1b-all | 61.75 | 69.04 | 59.22 | 42.02 | 31.17 | 17.63 |

Table 5: WER for various pre-trained models on whole dataset, Medical and General (non-medical) portions of the dataset, along with AMI (Carletta et al., 2005), Earnings22 (Del Rio et al., 2022), and Voxpopuli (Wang et al., 2021) datasets. Lower is better. Top 2 models in Bold

due to accented medical jargon such as medication names, diagnoses, and density of numerical entities (Afonja et al., 2024).

Following trends in Olatunji et al. (2023b), Wav2vec2-large-960h finetuned exclusively on read librispeech audiobooks (Panayotov et al., 2015) performs worst, confirming that model size (number of parameters), training set size, training data domain (in-the-wild, conversational) and multilingual pretraining play a role in generalizability to accented speech.

Additionally, we compare performance on other conversational datasets reported in Table 5 and show a 5 to 20 point (absolute) performance drop on average suggesting poor generalizability of SOTA models to African-accented conversation English speech. This degradation underscores the challenges ASR models face with African accents, highlighting the pressing need for high-quality, domain-specific datasets to improve ASR performance for low-resource accents and languages.

### 5.3 Summarization

Table 6 shows our summarization results.

In BERTScore-H summaries are generated based on human transcripts and evaluated against human reference summaries. In BERTScore-M summaries are generated based on machine-predicted transcripts (Whisper-Large-v3) and evaluated against human reference summaries.

LLM-Eval-H represents the LLM-as-Judge results when key details in the human transcripts are sought from summaries generated from human transcripts. LLM-Eval-M represents the LLM-as-Judge results when key details in the human transcripts are sought summaries generated from machine-predicted transcripts

Table 6 shows that general domain models like GTP4o and Claude consistently achieve high scores across all evaluation metrics followed by Biomedical LLMs where generated summaries are semantically more similar to reference summaries.

Open general smaller LLMs like Gemma and the Llama models demonstrate lower semantic similarity with reference summaries (BERTScore) and their performance on LLM-Eval metrics demonstrate an inferior ability to retrieve critical information from the conversation. The weakest performer

| Model | BertScore-H(%) | LLM-Eval-H(%) | BertScore-M(%) | LLM-Eval-M(%) |
|---|---|---|---|---|
| **GTP4o** | **91.34** | **72.94** | **90.94** | **70.98** |
| Gpt3.5-turbo | 89.87 | 69.41 | 90.81 | 64.51 |
| Claude_Sonnet | 88.39 | **73.53** | 88.50 | 69.42 |
| gemma-2-9b | 86.00 | 71.37 | 85.48 | **71.57** |
| Meditron3-8B | 89.83 | 60.20 | 90.40 | 59.61 |
| Llama3-Med42-8B | **90.56** | 67.45 | **91.34** | 65.29 |
| Llama-3.1-8B-Instruct | 86.44 | 69.41 | 86.44 | 62.35 |
| Llama-3.2-3B-Instruct | 85.60 | 65.10 | 87.09 | 63.73 |
| Phi3-mini-instruct | 88.05 | 57.45 | 89.66 | 60.39 |

Table 6: Summarization performance of various models, evaluated using BERTScore (F1), LLM-Eval, and Human Evaluation. BERTScore-H and LLM-Eval-H represent the metrics when summaries are generated and evaluated using human-generated transcripts, while BERTScore-M and LLM-Eval-M represent the metrics for machine-generated transcripts. Higher is better. Top 2 models in Bold

overall was Phi3-mini-4k-instruct, particularly in the LLM-Eval scores, signaling significant challenges in medical summarization.

### 5.4 Error Propagation from cascading models

While summarizing ASR transcripts reveal no significant difference in semantic similarity to reference summaries (BERTScore-H vs BERTScore-M), a more nuanced pattern emerges with LLM-Eval-H and LLM-Eval-M metrics where many models show a notable drop in performance. LLM-Eval-M scores are on average 2 to 5 points lower than LLM-Eval-H. This indicates that ASR errors (noisy transcription outputs) may limit the summarization ability of LLMs leading to loss of critical healthcare information.

Table 7 presents a comparison between a human-written transcript and the corresponding summary generated by GPT4o. While GPT4o captures key details such as pain severity, lifestyle factors, and medical history, its summary omits a vital detail: the patient's experience of rectal bleeding during urination, which is crucial to the case. For a complete presentation of the case and summaries generated by other models, please refer to the Appendix section E.

## 6 Conclusion

This study highlights several challenges with accented conversational ASR, especially its impact on downstream medical summarization. We benchmark speaker diarization, ASR, and medical summarization, drawing attention to gaps and opportunities to improve accented conversational ASR in the African context.

## 7 Limitations

While simulated conversations provide the opportunity to benchmark ASR systems, they may not fully reflect the complexity of real-world natural dialogues where several interruptions, distractions and background noise are more typical. Additionally, the age range of contributors does not accurately reflect the diversity of real-world speaker populations, potentially impacting the generalizability of the results. The use of general large language models (LLMs) as evaluators may also introduce biases. At 5hrs, the dataset does not reflect the full breadth of medical or general conversations as several domains, subdomains, topics, and medical specialties were not covered in this release. Nonetheless, this work significantly contributes to addressing the gap in ASR systems' adaptability to African accents. Future research should aim to incorporate real-world data, broader age representation, and further optimize models to enhance the robustness and applicability of ASR technologies in diverse African linguistic environments.

## 8 Ethical Considerations

In a world where voice cloning and identity theft are major cybersecurity threats, releasing a voice-based dataset is not without risk. Each recording began with contributors expressing consent to the recording. But since real names were sometimes used during consent, these segments were removed to protect the identities of participants. During doctor-patient conversations, pseudonyms

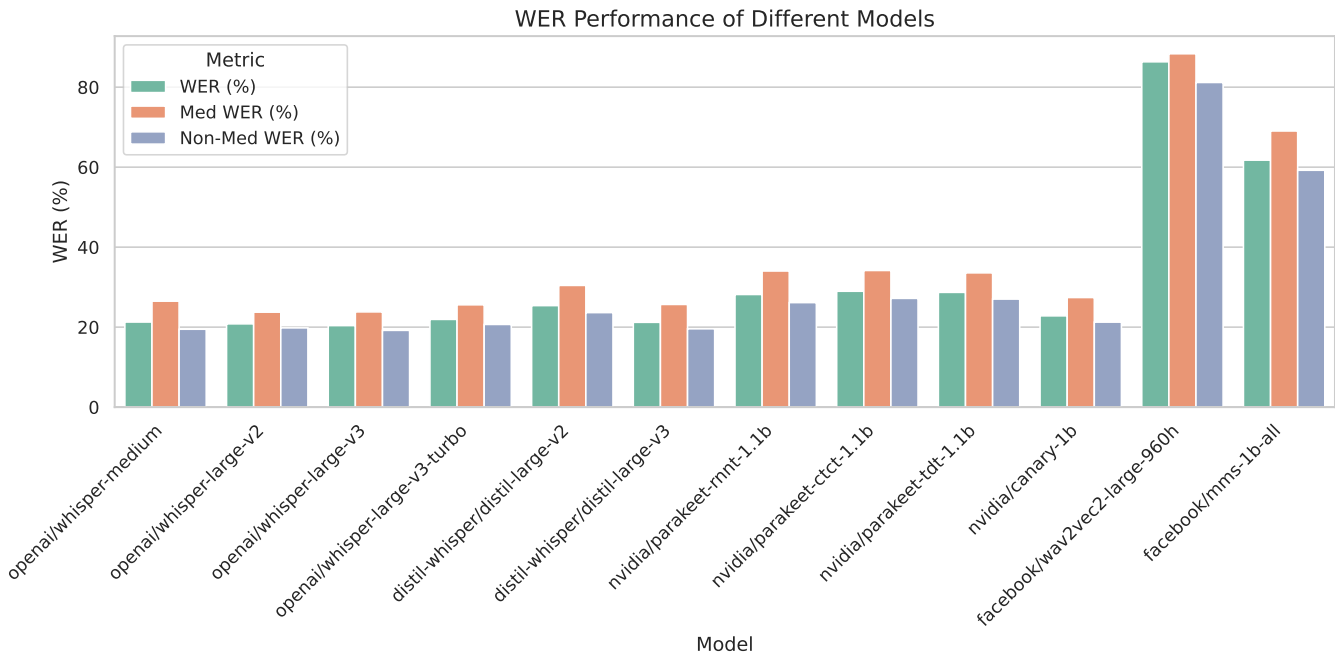


Figure 3: Comparison of Medical and General WER for Different Models

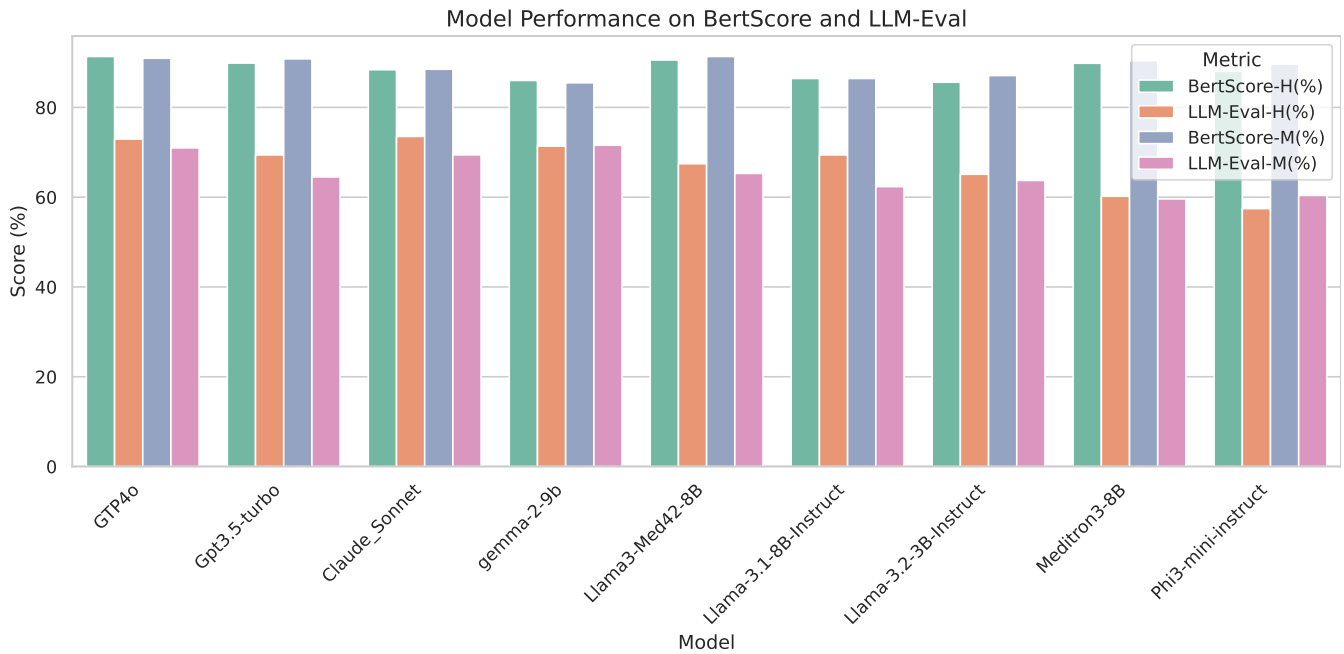


Figure 4: Summarization resuls for several LLMs

| Transcript |
|---|
| **Doctor:** Okay. You're welcome, my dear. What brought you here? |
| **Patient:** I have rectal bleeding and abdominal pain. The pain started two weeks ago, and the rectal bleeding happens when I urinate. The blood is not mixed with stool. |
| **Doctor:** How severe is the abdominal pain, from a scale of 1 to 10? |
| **Patient:** 6. |
| **Doctor:** Which side of the abdominal region do you feel the pain? |
| **Patient:** I feel pain in my lower abdominal region, and it worsens with pressure. |
| **Doctor** Have you undergone any surgery before? Do you smoke or drink alcohol? |
| **Patient:** No. |
| **Doctor:** Does anyone in your family have similar issues? |
| **Patient:** Yes, my mother had colorectal cancer. |
| **Doctor:** What foods do you normally eat? |
| **Patient:** I like taking swallow. |
| **Doctor:** Are you allergic to any food? |
| **Patient:** No. |

**GPT4o's Summary:** Dr. Philip conducted a consultation with Mrs. Ayike, a 45-year-old woman. She reported experiencing rectal bleeding and abdominal pain for the past two weeks. The pain, rated as a 6 out of 10 on the pain scale, is localized in the lower abdominal region and worsens with pressure. Mrs. Ayike also noted a loss of appetite and fatigue since the onset of her symptoms. She has not sought medical attention prior to this visit and has no history of surgery. She neither smokes nor drinks alcohol. Although Mrs. Ayike does not travel, she revealed a family history of colorectal cancer through her mother. Her usual diet includes "swallow," and she has no food allergies. Mrs. Ayike is married with four children, but currently lives separately from her husband. Dr. Philip concluded the session by acknowledging Mrs. Ayike's details.

Table 7: Snippet of a doctor-patient conversation (transcript) written by a clinician and the corresponding summary by GPT4o.

were used maintain anonymity. Furthermore, releasing real world doctor-patient conversations risk breaching privacy and confidentiality requirements hence the preference for simulated conversations.

### 8.1 References

## Acknowledgments

## References


Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. 2023. An empirical study of clinical note generation from doctor-patient encounters. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2291–2302.

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

David Ifeoluwa Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba O Alabi, Shamsuddeen H Muhammad, Peter Nabende, et al. 2022. Masakhaner 2.0: Africa-centric transfer learning for named entity recognition. *arXiv preprint arXiv:2210.12391*.

Tejumade Afonja, Tobi Olatunji, Sewade Ogun, Naome A Etori, Abraham Owodunni, and Moshood Yekini. 2024. Performant asr models for medical entities in accented speech. *arXiv preprint arXiv:2406.12387*.

Alëna Aksënova, Zhehuai Chen, Chung-Cheng Chiu, Daan van Esch, Pavel Golik, Wei Han, Levi King, Bhuvana Ramabhadran, Andrew Rosenberg, Suzan Schwartz, et al. 2022. Accented speech recognition: Benchmarking, pre-training, and diverse data. *arXiv preprint arXiv:2205.08014*.

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks

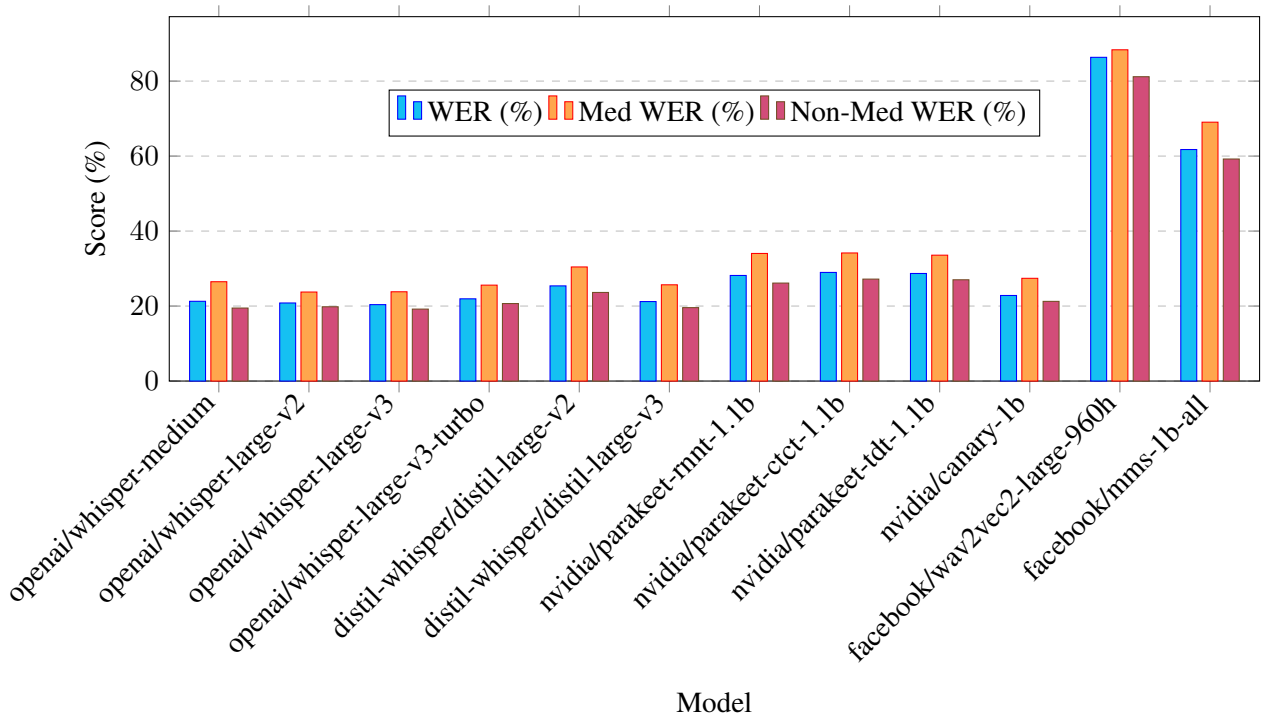


Figure 5: WER, Med WER, and Non-Med WER for Various Models


and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Anthropic. 2023. The claude 3 model family: Opus, sonnet, haiku. Accessed: 2024-10-15.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Yohannes Baye, Tesfaye Demeke, Nigusie Birhan, Agumasie Semahegn, and Simon Birhanu. 2020. Nurses' work-related stress and associated factors in governmental hospitals in harar, eastern ethiopia: A cross-sectional study. *PloS one*, 15(8):e0236782.

Nishchal Bhandari, Danny Chen, Miguel Ángel del Río Fernández, Natalie Delworth, Jennifer Drexler Fox, Migüel Jetté, Quinten McNamara, Corey Miller, Ondřej Novotný, Ján Profant, Nan Qin, Martin Ratajczak, and Jean-Philippe Robichaud. 2024. Reverb: Open-source asr and diarization from rev. *Preprint*, arXiv:2410.03930.

Jasmin Bharadiya. 2023. A comprehensive survey of deep learning techniques natural language processing. *European Journal of Technology*, 7(1):58–66.

Anjanava Biswas and Wrick Talukdar. 2024. Intelligent clinical documentation: Harnessing generative ai for patient-centric clinical note generation. *arXiv preprint arXiv:2405.18346*.

Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. 2020. Pyannote. audio: neural building blocks for speaker diarization. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7124–7128. IEEE.

Hervé Bredin. 2023. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *Proc. INTERSPEECH 2023*.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, et al. 2005. The ami meeting corpus: A pre-announcement. In *International workshop on machine learning for multimodal interaction*, pages 28–39. Springer.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf,

Amirkeivan Mohtashami, et al. 2023. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

Clément Christophe, Praveen K Kanithi, Tathagata Raha, Shadab Khan, and Marco AF Pimentel. 2024. Med42-v2: A suite of clinical llms.

Victoria YH Chua, Hexin Liu, Leibny Paola Garcia Perera, Fei Ting Woon, Jinyi Wong, Xiangyu Zhang, Sanjeev Khudanpur, Andy WH Khong, Justin Dauwels, and Suzy J Styles. 2023. Merlion ccs challenge: A english-mandarin code-switching child-directed speech corpus for language identification and diarization. *arXiv preprint arXiv:2305.18881*.

Miguel Del Rio, Peter Ha, Quinten McNamara, Corey Miller, and Shipra Chandra. 2022. Earnings-22: A practical benchmark for accents in the wild. *arXiv preprint arXiv:2203.15591*.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *arXiv preprint arXiv:2005.07143*.

George R Doddington, Mark A Przybocki, Alvin F Martin, and Douglas A Reynolds. 2000. The nist speaker recognition evaluation–overview, methodology, systems, results, perspective. *Speech communication*, 31(2-3):225–254.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Moayad Elamin, Yonas Chanie, Paul Ewuzie, and Samuel Rutunda. 2023. Multilingual automatic speech recognition for kinyarwanda, swahili, and luganda: Advancing asr in select east african languages. In *4th Workshop on African Natural Language Processing*.

Seppo Enarvi, Marilisa Amoia, Miguel Del-Agua Teba, Brian Delaney, Frank Diehl, Stefan Hahn, Kristina Harris, Liam McGrath, Yue Pan, Joel Pinto, et al. 2020. Generating medical reports from patient-doctor conversations using sequence-to-sequence models. In *Proceedings of the first workshop on natural language processing for medical conversations*, pages 22–30.

Faiha Fareez, Tishya Parikh, Christopher Wavell, Saba Shahab, Meghan Chevalier, Scott Good, Isabella De Blasi, Rafik Rhouma, Christopher McMahon, Jean-Paul Lam, et al. 2022. A dataset of simulated patient-physician medical interviews with a focus on respiratory cases. *Scientific Data*, 9(1):313.

J Luke Galloway, Dominique Munroe, Pamela D Vohra-Khullar, Christopher Holland, Mary A Solis, Miranda A Moore, and Reema H Dbouk. 2024. Impact of an artificial intelligence-based solution on clinicians' clinical documentation experience: Initial findings using ambient listening technology. *Journal of General Internal Medicine*, pages 1–3.

Sanchit Gandhi, Patrick von Platen, and Alexander M Rush. 2023. Distil-whisper: Robust knowledge distillation via large-scale pseudo labelling. *arXiv preprint arXiv:2311.00430*.

Martucci Giuseppe, Mauro Cettolo, Matteo Negri, and Marco Turchi. 2021. Lexical modeling of asr errors for robust speech translation. In *Proceedings of Interspeech 2021*, pages 2282–2286.

Michael Hanna and Ondřej Bojar. 2021. A fine-grained analysis of bertscore. In *Proceedings of the Sixth Conference on Machine Translation*, pages 507–517.

Eric Harper, Somshubra Majumdar, Oleksii Kuchaiev, Li Jason, Yang Zhang, E Bakhturina, V Noroozi, S Subramanian, K Nithin, H Jocelyn, et al. 2019. Nemo: A toolkit for conversational ai and large language models. *Computer software], URL: https://github. com/NVIDIA/NeMo*.

Arthur Hinsvark, Natalie Delworth, Miguel Del Rio, Quinten McNamara, Joshua Dong, Ryan Westerman, Michelle Huang, Joseph Palakapilly, Jennifer Drexler, Ilya Pirkin, et al. 2021. Accented speech recognition: A survey. *arXiv preprint arXiv:2104.10747*.

Praveen K Kanithi, Clément Christophe, Marco AF Pimentel, Tathagata Raha, Nada Saadi, Hamza Javed, Svetlana Maslenkova, Nasir Hayat, Ronnie Rajan, and Shadab Khan. 2024. Medic: Towards a comprehensive framework for evaluating llms in clinical applications. *arXiv preprint arXiv:2409.07314*.

Nithin Rao Koluguri, Taejin Park, and Boris Ginsburg. 2022. Titanet: Neural model for speaker representation with 1d depth-wise separable convolutions and global context. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8102–8106.

Alex Papadopoulos Korfiatis, Francesco Moramarco, Radmila Sarac, and Aleksandar Savkov. 2022. Primock57: A dataset of primary care mock consultations. *arXiv preprint arXiv:2204.00333*.

Harold W Kuhn. 1955. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97.

Federico Landini, Ján Profant, Mireia Diez, and Lukáš Burget. 2022. Bayesian hmm clustering of x-vector sequences (vbx) in speaker diarization: Theory, implementation and analysis on standard tasks. *Computer Speech Language*, 71:101254.

Khai Le-Duc, Khai-Nguyen Nguyen, Long Vo-Dang, and Truong-Son Hy. 2024. Real-time speech summarization for medical conversations. *arXiv preprint arXiv:2406.15888*.

Yixin Liu, Alexander R Fabbri, Jiawen Chen, Yilun Zhao, Simeng Han, Shafiq Joty, Pengfei Liu, Dragomir Radev, Chien-Sheng Wu, and Arman Cohan. 2023. Benchmarking generation and evaluation capabilities of large language models for instruction controllable summarization. *arXiv preprint arXiv:2311.09184*.

Anirudh Mani, Shruti Palaskar, and Sandeep Konam. 2020. Towards understanding asr error correction for medical conversations. In *Proceedings of the first workshop on natural language processing for medical conversations*, pages 7–11.

George Michalopoulos, Kyle Williams, Gagandeep Singh, and Thomas Lin. 2022. Medicalsum: A guided clinical abstractive summarization model for generating medical reports from patient-doctor conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4741–4749.

Hoang-Thi Nguyen-Vo, Huy Nguycn-Gia, Hoan-Duy Nguyen-Tran, Hoang Pham-Minh, Hung Vo-Thanh, and Hao Do-Due. 2022. Marblenet: A deep neural network solution for vietnamese voice activity detection. In *2022 9th NAFOSTED Conference on Information and Computer Science (NICS)*, pages 154–158.

Sewade Ogun, Abraham T Owodunni, Tobi Olatunji, Eniola Alese, Babatunde Oladimeji, Tejumade Afonja, Kayode Olaleye, Naome A Etori, and Tosin Adewumi. 2024. 1000 african voices: Advancing inclusive multi-speaker multi-accent speech synthesis. *arXiv preprint arXiv:2406.11727*.

Tobi Olatunji, Tejumade Afonja, Bonaventure FP Dossou, Atnafu Lambebo Tonja, Chris Chinenye Emezue, Amina Mardiyyah Rufai, and Sahib Singh. 2023a. Afrinames: Most asr models" butcher" african names. *arXiv preprint arXiv:2306.00253*.

Tobi Olatunji, Tejumade Afonja, Aditya Yadavalli, Chris Chinenye Emezue, Sahib Singh, Bonaventure FP Dossou, Joanne Osuchukwu, Salomey Osei, Atnafu Lambebo Tonja, Naome Etori, et al. 2023b. Afrispeech-200: Pan-african accented speech dataset for clinical and general domain asr. *Transactions of the Association for Computational Linguistics*, 11:1669–1685.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Hannaneh B Pasandi and Haniyeh B Pasandi. 2022. Evaluation of asr systems for conversational speech: A linguistic perspective. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*, pages 962–965.

Piotr Pęzik, Gosia Krawentek, Sylwia Karasińska, Paweł Wilk, Paulina Rybińska, Anna Cichosz, Angelika Peljak-Łapińska, Mikołaj Deckert, and Michał Adamczyk. 2022. Diabiz–an annotated corpus of polish call center dialogs. In *Proceedings of the thirteenth language resources and evaluation conference*, pages 723–726.

Mirosław Płaza, Łukasz Pawlik, and Stanisław Deniziak. 2021. Call transcription methodology for contact center systems. *IEEE Access*, 9:110975–110988.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, et al. 2024. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52.

Krishna C Puvvada, Piotr Żelasko, He Huang, Oleksii Hrinchuk, Nithin Rao Koluguri, Kunal Dhawan, Somshubra Majumdar, Elena Rastorgueva, Zhehuai Chen, Vitaly Lavrukhin, et al. 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 *International conference on machine learning*, pages 28492–28518. PMLR.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

Mirco Ravanelli and Yoshua Bengio. 2018. Speaker recognition from raw waveform with sincnet. *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 1021–1028.

Neville Ryant, Kenneth Church, Christopher Cieri, Alejandrina Cristia, Jun Du, Sriram Ganapathy, and Mark Liberman. 2019. The second dihard diarization challenge: Dataset, task, and baselines. pages 978–982.

Oleg Serikov, Ekaterina Voloshina, Anna Postnikova, Saliha Muradoğlu, Éric Le Ferrand, Elena Klyachko, Ekaterina Vylomova, Tatiana Shavrina, and Francis Tyers. 2024. Proceedings of the 3rd workshop on nlp applications to field linguistics (field matters 2024). In *Proceedings of the 3rd Workshop on NLP Applications to Field Linguistics (Field Matters 2024)*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Gabriel Skantze. 2021. Turn-taking in conversational systems and human-robot interaction: a review. *Computer Speech & Language*, 67:101178.

Georgios P Spithourakis, Ivan Vulić, Michał Lis, Iñigo Casanueva, and Paweł Budzianowski. 2022. Evi: Multilingual spoken dialogue tasks and dataset for knowledge-based enrolment, verification, and identification. *arXiv preprint arXiv:2204.13496*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Mohamad-Hani Temsah, Amr Jamal, Khalid Alhasan, Abdulkarim A Temsah, and Khalid H Malki. 2024. Openai o1-preview vs. chatgpt in healthcare: A new frontier in medical ai reasoning. *Cureus*, 16(10).

Atnafu Lambebo Tonja, Israel Abebe Azime, Tadesse Destaw Belay, Mesay Gemeda Yigezu, Moges Ahmed Mehamed, Abinew Ali Ayele, Ebrahim Chekol Jibril, Michael Melese Woldeyohannis, Olga Kolesnikova, Philipp Slusallek, et al. 2024. Ethiollm: Multilingual large language models for ethiopian languages with task evaluation. *arXiv preprint arXiv:2403.13737*.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 993–1003, Online. Association for Computational Linguistics.

Lucy Lu Wang, Yulia Otmakhova, Jay DeYoung, Thinh Hung Truong, Bailey E Kuehl, Erin Bransom, and Byron C Wallace. 2023. Automated metrics for medical multi-document summarization disagree with human evaluations. *arXiv preprint arXiv:2305.13693*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Emre Yılmaz, Astik Biswas, Ewald van der Westhuizen, Febe de Wet, and Thomas Niesler. 2018. Building a unified code-switching asr system for south african languages. *arXiv preprint arXiv:1807.10949*.

Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. Aci-bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *Scientific Data*, 10(1):586.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

## A Human Evaluation Criteria

This section provides details on the criteria used to evaluate the AI-generated summaries of doctor-patient conversations. Each summary was evaluated based on the following aspects: 1. How well does the summary include all positive clinical signs, symptoms, or details in the transcript? 2. How well does the summary include all negative clinical signs, symptoms, or details in the transcript? 3. How well does the summary recall the diagnosis from the transcript? 4. How accurately does the summary reflect the treatment plan in the transcript? 5. To what degree does the summary avoid false, incorrect, or fabricated information? 6. How concise, clear, and well-structured is the summary?

## B Prompt Template for the LLM Evaluator

This figure outlines the prompt template utilized for the evaluation conducted by OpenAI's o1 model, serving as the LLM evaluator.

You are a medical expert. Your task is to evaluate the accuracy and quality of an AI-generated summary of a doctor-patient conversation. Please provide an impartial assessment based on the following criteria, rating each from 1 to 5 (1 = lowest, 5 = highest). If a specific criterion is not addressed in the transcript, assign a score of 0.

Carefully assess the summary using these criteria:

**Transcript:** {transcript}

**Summary:** {summary}

**Evaluation Criteria:**
1. Completeness of positive clinical signs, symptoms, or relevant details: Does the summary capture all important positive findings from the transcript?
2. Completeness of negative clinical signs, symptoms, or relevant details: Does the summary capture all important negative findings from the transcript?
3. Accuracy of diagnosis: Does the summary correctly reflect the diagnosis mentioned in the transcript?
4. Accuracy of treatment plan: Does the summary correctly convey the treatment plan discussed in the transcript?
5. Factual accuracy: Does the summary avoid any false, incorrect, or fabricated information?
6. Clarity and structure: Is the summary concise, clear, and well-structured?

Please return the evaluation scores in the following JSON format:
{
"Criterion 1": score,
"Criterion 2": score,
"Criterion 3": score,
"Criterion 4": score,
"Criterion 5": score,
"Criterion 6": score
}

## C Evaluation Metric for the LLM Evaluator

Each criterion score has a maximum of 5 points. Since there are 6 criteria for each row of data, the perfect score for one row is 5×6=30 points. If we have 10 rows, and each row has a perfect score, the total possible score for the dataset is 5×6×10=300 points. This is considered the "perfect accuracy" score. To compute the percentage score for a model: For each row, sum up the points awarded for all 6 criteria. Then, sum up the scores for all the rows to get the total score. Finally, I calculate the percentage by dividing the total score by the perfect accuracy and multiplying by 100. $\text{Percentage}_{Score} = (TotalScoreAcrossAllRows/56NumberofRows)100$

## D Prompt Template for Summarization

This figure outlines the prompt template utilized for the summarization task.

You are a medical expert; you will be provided a transcript that contains a conversation between a doctor and a patient. The task is for you to summarize the key points in the conversation.
**Transcript:** {transcript}

## E Sample of LLM Generated Summaries for The Transcript in Table 8

This section gives an example of the LLM-generated summaries for the transcript in Table 8. The summaries are:

**Claude's Summary:** Based on the conversation, the key points are: 1. The patient, Mrs. Ayike, is a 45-year-old woman presenting with rectal bleeding and abdominal pain for the past two weeks. 2. The rectal bleeding occurs when she urinates, and it is not mixed with stool.3. The abdominal pain is rated as 6 out of 10 in severity, and it is localized in the lower abdominal region.4. She has lost her appetite and feels tired since the symptoms started.5. She has not traveled recently and is not a frequent traveler.6. She is married but currently separated from her husband, living in a different state.n7. She has four children.8. She has not undergone any surgery before. 9. She does not smoke or consume alcohol.10. Her mother had colorectal cancer in the past, indicating a family history of the condition.11. Her typical diet includes swallows (a type of food), and she is not allergic to any food. The doctor has gathered essential information about the patient's symptoms,

medical history, and lifestyle to aid in the diagnosis and treatment process.

**Gemma's Summary:** Mrs. Ayike, a 45-year-old woman, presented to Dr. Philip complaining of two weeks of abdominal pain and rectal bleeding. Key points:Symptoms:Lower abdominal pain rated 6/10, worsening with pressure. Rectal bleeding during urination, not mixed with stool. Loss of appetite and fatigue. Medical History: No previous surgeries. No smoking or alcohol consumption. Family history of colorectal cancer (mother). Lifestyle: Prefers a diet high in swallow. No known food allergies. Not a traveler. Separated from her husband, living with four children. Next steps: Based on the symptoms and family history, Dr. Philip will likely order further investigations to determine the cause of Mrs. Ayike's abdominal pain and rectal bleeding. This may include blood tests, stool tests, imaging studies (such as a colonoscopy or CT scan), and possibly a biopsy.

**Med42-8B's summary:** Mrs. Ayike, a 45-year-old female, presents with rectal bleeding and lower abdominal pain, rated 6/10 in severity, which started two weeks ago. She reports seeing red blood in her urine but not mixed with stool. She has not sought medical attention previously for these symptoms. She describes her appetite as poor and feels tired. She denies any recent travel, smoking, or alcohol use. She is married with four children but is currently separated from her husband. Her mother had colorectal cancer. She consumes a diet primarily consisting of fufu (swallow). There are no known food allergies.

**GPT3.5's summary:** Mrs. Aike, a 45-year-old patient, visited Dr. Philip due to rectal bleeding and abdominal pain that started two weeks ago. The abdominal pain is severe, rated at a 6 on a scale of 1 to 3, and is localized in the lower abdominal region. She has also experienced a loss of appetite, fatigue, and difficulty eating since the pain started. Mrs. Aike is married with four children, not living with her husband, and has no history of surgeries, smoking, or alcohol consumption. Her mother had colorectal cancer before, indicating a family history of similar issues. She prefers eating swallow and is not allergic to any food. This is her first visit to seek medical help for these symptoms.

**Meditron3-8B's summary:** The patient, Mrs. Ayike, is a 45-year-old woman who has been experiencing rectal bleeding and abdominal pain for two weeks. She has not undergone any surgery before and does not smoke or take alcohol. Her mother had colorectal cancer before. She has lost her appetite and feels tired. The doctor asks her about her family history, food allergies, and the severity of her symptoms. The doctor also asks about her marital status and children. The patient's symptoms are similar to those of colorectal cancer, which her mother had.

**Llama-3.1-8B-Instruct's summary:** Based on the provided transcript, here are the key points from the conversation between Dr. Philip and Mrs. Ayike: 1. Patient's Introduction: Mrs. Ayike, a 45-year-old woman, presents to the doctor with rectal bleeding and abdominal pain. 2.Symptoms: The abdominal pain started two weeks ago, and she experiences rectal bleeding when urinating, but not mixed with stool. The pain is located in the lower abdominal region and worsens with pressure. 3. Medical History: Mrs. Ayike has not undergone any surgery before and does not have a history of smoking or alcohol consumption. 4. Family History: Her mother had colorectal cancer before.5. Diet and Allergies: Mrs. Ayike likes eating swallow (a type of food) and is not allergic to any food. 6.Lifestyle : She is married with four children but is currently separated from her husband. She does not travel and has not been able to eat well or feel energetic due to her symptoms. 7. Current Condition: Mrs. Ayike rates her abdominal pain as a 6 out of 10 and has lost appetite and feels tired since the onset of symptoms. Based on the information provided, Dr. Philip will likely perform a thorough examination and possibly order diagnostic tests to determine the cause of Mrs. Ayike's symptoms. Given her family history of colorectal cancer, a colonoscopy may be recommended to rule out any potential malignancies.

**Llama-3.2-3B-Instruct's summary:** Here's a summary of the key points in the conversation between Dr. Philip and Mrs. Ayike: Patient's Symptoms: Rectal bleeding (not mixed with stool), Abdominal pain (rated 6 out of 10) in the lower abdominal region, Loss of appetite and tiredness for the past two weeks. Medical History: No previous surgeries, No smoking or

alcohol consumption, No known allergies to food Family History: Mother had colorectal cancer before. Lifestyle and Habits: Married with four children, but not living with her husband. No history of traveling. Prefers swallowing food. Other Information: Mrs. Ayike is 45 years old. She has not been able to eat well due to the abdominal pain. She is seeking medical attention for the first time, as the symptoms started two weeks ago.

**Phi3-mini-instruct's summary:** Dr. Philip, a medical expert, is speaking with Mrs. Ayike, a 45-year-old woman who has been experiencing rectal bleeding and abdominal pain for two weeks. Mrs. Ayike has not had any check-ups for her symptoms and has lost her appetite, feeling tired. She has not traveled recently and is not a smoker. She does not consume alcohol and has no known food allergies. Mrs. Ayike's mother had colorectal cancer. Dr. Philip asks about her diet, which mainly consists of swallow, and inquires about any family history of similar issues.

| **Transcript** |
| --- |
| **Speaker 1:** Good afternoon. My name is Dr. Philip, and what's your name? |
| **Speaker 2:** My name is Mrs. Ayike. |
| **Speaker 1:** Okay, how old are you? |
| **Speaker 2:** I am 45 years old. |
| **Speaker 1:** Okay. You're welcome, my dear. What brought you here? |
| **Speaker 2:** I have rectal bleeding and abdominal pain. |
| **Speaker 1:** Can you tell me more about it? |
| **Speaker 2:** The pain started two weeks ago, and the rectal bleeding happens when I urinate. The blood is not mixed with stool. |
| **Speaker 1:** Have you gone for any check-ups? |
| **Speaker 2:** No, this is my first time coming here. |
| **Speaker 1:** How severe is the abdominal pain, from a scale of 1 to 10? |
| **Speaker 2:** 6. |
| **Speaker 1:** Which side of the abdominal region do you feel the pain? |
| **Speaker 2:** I feel pain in my lower abdominal region, and it worsens with pressure. And since the last two weeks, as I've started the abdominal pain, I have not been able to eat well. I lost appetite. I feel tired. |
| **Speaker 1:** Have you traveled recently? |
| **Speaker 2:** No, I don't travel. |
| **Speaker 1:** Are you married? |
| **Speaker 2:** Yes, I am married, with four children, but my husband lives in another state. |
| **Speaker 1:** Have you undergone any surgery before? |
| **Speaker 2:** No. |
| **Speaker 1:** Do you smoke or drink alcohol? |
| **Speaker 2:** No. |
| **Speaker 1:** Does anyone in your family have similar issues? |
| **Speaker 2:** Yes, my mother had colorectal cancer. |
| **Speaker 1:** What foods do you normally eat? |
| **Speaker 2:** I like taking swallow. I like swallow a lot. |
| **Speaker 1:** Are you allergic to any food? |
| **Speaker 2:** No. |

Table 8: Snippet of a doctor-patient conversation (transcript) written by a clinician