Robust Semantic Reasoning in Audio Language Models via In-Context Learning

FirstNameA LastNameA [1,*], *FirstNameB InitialB LastNameB* [2,3,*,**], *FirstNameC LastNameC* [1,3]

[1] Address Affiliation 1, Country Affiliation 1
[2] Address Affiliation 2, Country Affiliation 2
[3] Address Affiliation 3, Country Affiliation 3

first@university.edu, second@companyA.com, third@companyB.ai

Abstract

Audio–language models (ALMs) have recently shown strong zero-shot performance on speech understanding tasks, yet their robustness to accented speech and semantic reasoning remains underexplored. In this work, we investigate whether reasoning failures in ALMs stem primarily from acoustic mismatch or from linguistic decision bias. We evaluate multiple generative ALMs on audio entailment across accented and domain-shifted datasets, observing pronounced class imbalance and entailment dominance despite competitive overall accuracy. We then introduce an in-context learning (ICL) framework that conditions next-token prediction models with balanced semantic exemplars to recalibrate reasoning boundaries without parameter updates. Results show that ICL improves class balance and macro-F1 on accented data than on domain-matched speech, suggesting that many observed failures arise from linguistic inference bias rather than purely acoustic degradation. Our findings provide new evidence that contextual semantic calibration is an effective, lightweight strategy for improving reasoning reliability in audio–language models under accent variability.

Index Terms: audio language models, speech recognition, accented speech, semantic reasoning, in-context learning

1. Introduction

Audio–language models (ALMs) integrate speech encoders with large language model (LLM) decoders, enabling direct reasoning over spoken input without explicit transcription [1, 2]. Recent systems demonstrate strong performance on audio captioning, question answering, and speech-based natural language inference [3, 4]. However, competitive aggregate accuracy does not necessarily imply stable semantic reasoning [5, 6]. In particular, performance under accented or non-standard speech often reveals class imbalance, prediction collapse, and systematic over-reliance on specific labels [7, 8]. In real-world applications such as healthcare [9], education, and multilingual communication, accent variability is pervasive, raising an important question: do ALMs fail because they cannot represent the acoustics reliably, or because their reasoning boundaries are poorly calibrated?

Most prior work has framed robustness primarily as an acoustic problem, focusing on speech recognition quality or representation invariance [10, 11]. Under this view, downstream reasoning errors are assumed to stem from degraded audio embeddings [12]. However, emerging observations suggest that even when transcripts are intelligible, ALMs exhibit strong decision biases, such as over-predicting entailment or neutral labels [13, 2]. This pattern indicates that reasoning instability may arise from linguistic calibration issues rather than purely acoustic mismatch [2]. Disentangling these two failure modes acoustic degradation versus semantic decision-boundary bias is critical for designing principled mitigation strategies.

In this work, we investigate this distinction through the lens of in-context learning (ICL) applied to generative audio–language models. ICL provides a lightweight mechanism for semantic recalibration by conditioning predictions on labeled exemplars without updating model parameters [14, 15, 16, 17]. By comparing zero-shot performance, ASR–LLM cascades, and three ICL variants (audio-only, audio-plus-transcript, and audio-plus-transcript-plus hypothesis), we directly test whether contextual semantic conditioning can correct reasoning instability and whether improvements depend on textual reinforcement. Our contributions are fourfold:

- **Controlled Analysis of Audio Entailment Reasoning Stability.** We provide a comprehensive evaluation of seven generative ALMs, three contrastive ALMs, and ASR–LLM cascades on two speech-based entailment datasets. We show that zero-shot performance can mask severe class imbalance and entailment dominance, highlighting the need for fine-grained reasoning diagnostics beyond aggregate accuracy.
- **In-Context Semantic Calibration Framework.** We introduce an in-context learning overlay for audio entailment that systematically scales from 1 to 10 shots using balanced audio–hypothesis exemplars. This framework enables controlled analysis of reasoning calibration without fine-tuning and isolates the role of contextual semantic guidance in multimodal inference.
- **Attribution of Failure Modes: Linguistic vs. Acoustic.** By comparing three ICL variants (audio-only, audio-plus-transcript, and audio-plus-transcript-plus hypothesis) conditioning against ASR–LLM baselines, we provide empirical evidence that a substantial portion of reasoning instability under accent variability originates from linguistic decision-boundary bias rather than solely from acoustic representation degradation.
- **Accent-diverse audio entailment benchmark for robust multimodal reasoning.** We curate and formalize an accent-diverse audio entailment benchmark that isolates semantic reasoning under realistic speech variability. The dataset includes logically challenging hypothesis constructions and balanced label distributions and is split to ensure that no test speaker or hypothesis overlaps with the exemplar pool. This design supports reproducible evaluation of accent robustness across languages and domains.

[*] These authors contributed equally.
[**] indicates the corresponding author.

Together, our findings suggest that accent robustness in ALMs is not merely an acoustic challenge but also a problem of semantic calibration. We demonstrate that lightweight contextual conditioning can reshape reasoning behavior, offering a practical direction for improving reliability in speech multimodal systems.

2. Related Work

2.1. In-Context Learning in Large Language Models

In-context learning (ICL) emerged as a defining capability of large language models (LLMs) with the introduction of GPT-3 [18], demonstrating that models can perform few-shot adaptation without parameter updates [18]. Subsequent work showed that ICL enables competitive performance across reasoning, classification, and structured inference tasks when prompted with exemplar demonstrations [19, 16, 20].

Several studies have analyzed the mechanisms underlying ICL. Min et al. [21] argue that label space specification plays a central role, while Von et al. [22, 23] propose that transformers perform implicit gradient descent within context. Others have observed diminishing returns with increasing shot counts [18], highlighting saturation effects similar to those observed in our audio entailment setting.

However, most ICL research has focused exclusively on text-based models. Far less work has examined whether similar contextual scaling behavior transfers to audio-language models operating directly on speech inputs [24].

2.2. In-Context Learning in Speech and Audio-Language Models

Recent multimodal models extend ICL capabilities to speech and audio inputs. Models such as AudioPaLM [25], SALMONN [26], and Qwen-Audio [10] demonstrate speech understanding and generation by integrating acoustic encoders with large language models. These architectures enable zero-shot and few-shot transfer across speech tasks including question answering, summarization, and dialogue [27, 16].

Despite these advances, systematic evaluation of ICL scaling behavior in audio-language models remains limited [27]. Prior work typically reports aggregate zero-shot or few-shot results without analyzing shot efficiency, saturation dynamics, or conditioning strategies under accent variability [15]. Moreover, most benchmarks rely on English speech from high-resource domains, leaving robustness under accent diversity largely unexplored [27]. A high-level overview of our multimodal architecture and ICL pipeline is shown in Figure 1.

Our work contributes by explicitly evaluating ICL across multiple shot counts and conditioning variants (audio-only, audio + transcript, audio + transcript + hypothesis), revealing non-monotonic scaling and diminishing returns beyond moderate context lengths.

2.3. ASR and Low-Resource / Accented Speech

Automatic speech recognition (ASR) performance degrades significantly under accent variation and low-resource conditions [28]. Studies have documented disparities in word error rates across dialects and accents, particularly for African American English and other underrepresented speech varieties [28].

Recent multilingual ASR systems such as Whisper [29] and XLS-R [30] have improved cross-lingual and cross-accent robustness through large-scale self-supervised pretraining. Nevertheless, recognition errors under accent shift can propagate into downstream tasks, affecting semantic inference reliability.

Cascaded ASR–LLM pipelines attempt to mitigate this by decoupling acoustic modeling from reasoning [1]. However, such systems depend critically on transcription fidelity. Our results show that cascaded systems remain competitive, particularly in domain-specific settings, but end-to-end audio-language models can match their performance when sufficiently large and well-trained.

2.4. Natural Language Inference and Audio Entailment

Natural language inference (NLI) has long served as a benchmark for structured semantic reasoning [13]. Extending entailment tasks to multimodal settings has led to visual entailment benchmarks [31] and speech-based semantic tasks. However, audio entailment remains underexplored, particularly under accent variation.

Most existing speech understanding benchmarks evaluate transcription accuracy or intent classification rather than structured logical inference. By framing audio reasoning as a three-way entailment task (entailment, neutral, contradiction), our work bridges speech recognition and semantic reasoning evaluation.

3. Methods

3.1. Task Formulation

We formulate audio entailment as a three-way multimodal inference task. Each instance consists of an audio premise $a \in \mathcal{A}$, a textual hypothesis $h \in \mathcal{H}$, and a label $y \in \mathcal{Y}$, where

$$\mathcal{Y} = \{\text{ENTAILMENT}, \text{CONTRADICTION}, \text{NEUTRAL}\}.$$

Given a model f_θ, the objective is to predict

$$\hat{y} = f_\theta(a, h).$$

For generative audio–language models (ALMs), prediction is implemented as next-token classification:

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} P_\theta(y \mid a, h, \text{prompt}),$$

where the prompt may optionally contain in-context exemplars. For contrastive ALMs, audio and hypothesis are embedded into a shared latent space:

$$z_a = g_\theta(a), \quad z_h = t_\theta(h),$$

and classification is performed via similarity scoring against class-conditioned templates.

3.2. Dataset Construction

We construct two accented speech entailment datasets from *Afrispeech-200* and *Afrispeech-Medical*. Both datasets are designed to probe semantic reasoning under accent and domain variability rather than transcription fidelity.

3.2.1. AfriSpeech-200 (General Domain)

The **AfriSpeech-200 [32]** dataset is a large-scale Pan-African English accented corpus originally designed for general domain ASR. It comprises approximately 200 hours of audio from 2,463 unique speakers representing 120 indigenous accents across 13 countries. The dataset is linguistically diverse,

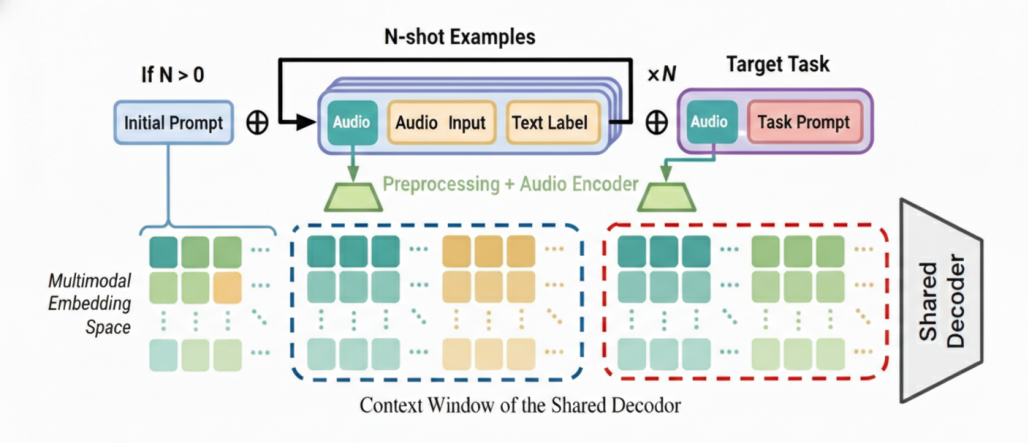

Figure 1: *Multimodal audio–language model architecture for N-shot learning. The framework illustrates the integration of an initial prompt, N audio–text example pairs, and a target task. Audio inputs are processed via a dedicated encoder and mapped into a shared multimodal embedding space alongside text tokens.*

featuring accents from five major language families, including Niger-Congo and Afro-Asiatic, with a significant portion of speakers originating from Nigeria (approx. 67%).

3.2.2. AfriSpeech-Medical (Clinical Domain)

The **AfriSpeech-Medical [30]** dataset specifically targets the challenges of clinical speech recognition in the Global South. It incorporates domain-specific terminology such as complex drug names, medical subspecialties, and physician-patient interaction jargon.

3.2.3. Multi-LLM Hypothesis Generation

For each transcript, candidate hypotheses were generated using three distinct large language models: Llama 3.1 8B, Mistral Large 3, and Qwen 2 7B. Using multiple generators mitigates stylistic bias and prevents coupling the benchmark to a single decoder family.

Each model was prompted to generate hypotheses corresponding to entailment, contradiction, and neutral relationships. Generation prompts explicitly constrained outputs to:

- remain grounded in the transcript,
- avoid introducing unstated world knowledge,
- vary logical structure (negation, modality, quantifier shifts, temporal contrasts),
- preserve lexical diversity across models.

This multi-source generation strategy increases semantic diversity and reduces decoder-specific artifacts in the benchmark.

3.2.4. Human Verification Protocol

All generated hypotheses were audited by a team of three trained annotators with backgrounds in linguistics and speech technology. Annotators followed a structured protocol requiring them to: (1) evaluate the entire audio premise before reviewing hypotheses, (2) assign labels based strictly on acoustic evidence, (3) refine hallucinated or ambiguous text while preserving intended semantic relations, and (4) exclude low-quality audio segments.

To ensure labeling reliability, each hypothesis was independently reviewed by two annotators, yielding a high inter-annotator agreement with a correlation coefficient of $\kappa = 0.91$. Disagreements were resolved through adjudication by a third annotator, and a senior author performed periodic spot checks to ensure corpus-wide consistency.ensure corpus-wide consistency.

3.2.4.1. Exemplar selection and balancing.

To construct the exemplar pools used in our few-shot prompts (§4.3), we generated a pool of 50 candidate audio–hypothesis pairs per dataset covering a wide variety of linguistic phenomena (e.g., lexical frequency, sentence length, negation) and relation types (entailment, contradiction, neutral). From this pool we selected 10 exemplars for each dataset such that each relation class and accent group is equally represented. All exemplars are disjoint from the evaluation sets there is no speaker overlap across splits, and prompts are fixed across runs to avoid data leakage and prompt induced bias. Balanced exemplar selection improves the reproducibility of our ICL experiments by controlling for prompt structure.

3.2.5. Agreement Analysis

Agreement between automated label mapping and human judgments was evaluated on a 300-sample subset using Cohen's κ:

$$\kappa = \frac{p_o - p_e}{1 - p_e},$$

where p_o is observed agreement and p_e is expected agreement by chance.

Automated mapping achieved 93.4% raw agreement with $\kappa = 0.89$, indicating near-perfect agreement and validating the reliability of the evaluation pipeline.

3.3. Evaluated Models

We evaluate both generative and contrastive audio–language models. Generative ALMs combine a speech encoder with an autoregressive large language model (LLM) decoder. The decoder plays a central role in semantic reasoning, as next-token prediction determines the final entailment label. Differences in decoder architecture, parameter scale, and instruction tuning therefore directly influence reasoning calibration and class bias.

Table 1: *Architectural specifications of evaluated audio–language models. Generative models rely on autoregressive LLM decoders for label prediction, whereas contrastive models operate via aligned embedding similarity.*

Model	Params	Text Decoder
Qwen2.5-Omni	7B	Qwen2.5-7B-Instruct
Qwen2-Audio	7B	Qwen2-7B-Instruct
SALMONN	13B	Vicuna-13B
Kimi-Audio	12.5B	Kimi-LLM-v1
GAMA	7B	Vicuna-7B (v1.5)
AudioFlamingo2	3B	Qwen2.5-3B
AudioFlamingo3	7B	Qwen2.5-7B
MSCLAP (2022)	–	HTSAT / RoBERTa
MSCLAP (2023)	–	HTSAT / RoBERTa
LAION-CLAP	–	Swin / RoBERTa

In contrast, contrastive ALMs rely on joint audio–text embedding alignment without explicit autoregressive reasoning [33]. This architectural distinction allows us to isolate decoder-driven semantic inference from purely representation-based matching.

Table 1 summarizes the architectural backbone and text decoder components of all evaluated systems.

3.4. Zero-Shot Baselines

We establish zero-shot baselines prior to introducing in-context learning.

- Generative ALMs: Seven generative ALMs are evaluated in a direct audio-to-label configuration without fine-tuning or exemplars.
- Contrastive ALMs: Three contrastive ALMs compute similarity between audio embeddings and textual hypotheses to derive class predictions.
- ASR–LLM Cascade: To isolate linguistic reasoning from acoustic encoding, we construct an ASR–LLM pipeline. Audio is first transcribed using Whisper Large-v3 and IBM Granite ASR systems, producing transcript $\tilde{a}$. The transcript is then passed to text-only LLMs (Qwen2, LLaMA 3.2, and Mistral):

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} P(y \mid \tilde{a}, h).$$

This cascade provides a reference condition in which reasoning operates over explicit textual premises.

3.5. In-Context Learning Overlay

To investigate whether reasoning failures arise from acoustic degradation or decision-boundary miscalibration, we introduce an in-context learning (ICL) overlay applied to generative ALMs.

Let

$$E_k = \{(a_j, h_j, y_j)\}_{j=1}^{k}$$

denote a set of k labeled exemplars prepended to each test instance. Prediction becomes:

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} P_\theta(y \mid E_k, a, h).$$

We evaluate shot levels $k \in \{1, 3, 5, 7, 10\}$ using a fixed pool of ten balanced exemplars covering diverse logical phenomena. Three conditioning variants are evaluated:

ICL Prompt Template

Task: Determine whether the hypothesis is entailed by the spoken premise. You must respond with exactly one label: `entailment`, `contradiction`, or `neutral`.

Few-shot Examples:
Example 1: Premise (audio): **[audio clip A]** Hypothesis: **[hypothesis A]** Answer: entailment

Example 2: Premise (audio): **[audio clip B]** Hypothesis: **[hypothesis B]** Answer: contradiction

Example 3: Premise (audio): **[audio clip C]** Hypothesis: **[hypothesis C]** Answer: neutral

Target Instance: Premise (audio): **[target audio]** Hypothesis: **[target hypothesis]**
Answer:

Figure 2: *In-context learning prompt template for audio entailment. Brackets indicate placeholders for audio and textual inputs.*

1. Audio exemplars.
2. Audio-transcript exemplars.
3. Audio-transcript-Hypothesis exemplars.

3.5.1. In-Context Prompt Template

We instantiate all conditioning strategies using a structured prompt template that combines the instruction, a variable number of few-shot exemplars, and the target premise and hypothesis. The template explicitly instructs the model to produce one of three labels (entailment, contradiction, neutral) and includes delimiters for audio inputs and textual components. The detailed prompt template is shown in Figure 2.

3.6. Evaluation

Let $\{(y_i, \hat{y}_i)\}_{i=1}^{N}$ denote ground-truth and predicted labels.

3.6.0.1. Accuracy

$$\text{Accuracy} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}(\hat{y}_i = y_i).$$

3.6.0.2. Precision and Recall

For class $c \in \mathcal{Y}$:

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c}, \quad \text{Recall}_c = \frac{TP_c}{TP_c + FN_c}.$$

3.6.0.3. Macro-F1

$$F1_{\text{macro}} = \frac{1}{|\mathcal{Y}|} \sum_{c \in \mathcal{Y}} \frac{2 \cdot \text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c}.$$

All metrics are computed per dataset and pooled across datasets for comparative analysis.

4. Results

In this section, we present a comprehensive evaluation of end-to-end Audio Language Models (ALMs) and cascaded ASR–

Table 2: *Zero-shot macro-F1, accuracy, precision, recall, and class-wise accuracy for generative audio-language models on AfriSpeech-200 and AfriSpeech-Medical. Higher is better for all metrics.*

Dataset	ALM	Acc	P	R	F1	E-Acc	N-Acc	C-Acc
AfriSpeech-200	AudioFlamingo2	0.3333	0.1111	0.3333	0.1667	0.0000	1.0000	0.0000
	AudioFlamingo3	0.6367	0.7398	0.6367	0.5466	0.9800	0.0400	0.8900
	GAMA	0.2967	0.1739	0.2967	0.1936	0.8400	0.0500	0.0000
	Kimi	0.6333	0.7499	0.6333	0.5383	0.8700	0.0700	0.9600
	Qwen2.5 Omni	0.6800	0.7076	0.6800	0.6731	0.8600	0.3400	0.8400
	Qwen2 Audio 7B	0.7133	0.7283	0.7133	**0.7123**	0.7400	0.5300	0.8700
	SALMONN	0.3967	0.4520	0.3967	0.2814	0.0000	1.0000	0.1900
AfriSpeech-Medical	AudioFlamingo2	0.3057	0.1024	0.3278	0.1561	0.0000	0.9833	0.0000
	AudioFlamingo3	0.6218	0.5512	0.5954	0.5160	0.9863	0.0167	0.7833
	GAMA	0.3782	0.2621	0.3591	0.3030	0.6438	0.4333	0.0000
	Kimi	0.6500	0.6726	0.6500	**0.5995**	0.8250	0.2000	0.9250
	Qwen2.5 Omni	0.5492	0.6414	0.5335	0.5274	0.7671	0.5500	0.2833
	Qwen2 Audio 7B	0.5521	0.5802	0.5546	0.5534	0.5139	0.6833	0.4667
	SALMONN	0.3679	0.6042	0.3737	0.2960	0.2877	0.8000	0.0333

Table 3: *Zero-shot macro-F1, accuracy, precision, recall, and class-wise accuracy for contrastive audio-language models on AfriSpeech-200 and AfriSpeech-Medical.*

Dataset	ALM	Acc	P	R	F1	E-Acc	N-Acc	C-Acc
AfriSpeech-200	LAION-CLAP	0.3333	0.3333	0.3333	0.3325	0.3000	0.3200	0.3800
	MSCLAP_23	0.3200	0.2127	0.3200	0.2321	0.2000	0.0000	0.7600
	MSCLAP_22	0.3500	0.3448	0.3500	0.3444	0.2200	0.4200	0.4100
AfriSpeech-Medical	LAION-CLAP	0.3782	0.3868	0.3868	0.3725	0.2603	0.5500	0.3500
	MSCLAP_23	0.3005	0.2019	0.3064	0.2246	0.2192	0.0000	0.7000
	MSCLAP_22	0.3834	0.3988	0.3607	0.3286	0.6986	0.1167	0.2667

LLM baselines on the Audio Entailment task. We begin with zero-shot performance to establish baseline reasoning capacity under accented speech conditions, followed by an analysis of in-context learning (ICL) across shot counts and conditioning strategies.

4.1. Zero-Shot Performance

Table 2 reports zero-shot results for generative ALMs on AfriSpeech-200 and AfriSpeech-Medical. On AfriSpeech-200, Qwen2 Audio 7B achieves the highest macro-F1 of 0.7123, substantially outperforming AudioFlamingo3 (0.5466) and Kimi (0.5383). The performance gap exceeding 16 F1 points relative to the next strongest generative model indicates stronger semantic robustness under accent variation. Importantly, Qwen2 Audio 7B maintains relatively balanced class-wise behavior, with entailment accuracy of 0.7400, neutral accuracy of 0.5300, and contradiction accuracy of 0.8700.

On AfriSpeech-Medical, the ranking shifts. Kimi achieves the highest macro-F1 of 0.5995, while Qwen2 Audio 7B attains 0.5534. This domain-specific reversal suggests that robustness does not uniformly transfer across lexical domains. Medical speech introduces specialized terminology and contextual structure that may not be evenly represented during model pretraining. While Qwen2 Audio 7B remains competitive, Kimi exhibits greater stability in the medical setting.

Contrastive models (Table 3) perform near chance-level macro-F1 (0.33–0.37), indicating that embedding similarity alone is insufficient for structured three-way semantic reasoning. Although MSCLAP_22 achieves 0.3444 on AfriSpeech-200 and 0.3286 on AfriSpeech-Medical, these scores reflect limited capacity to distinguish entailment, neutral, and contradiction without generative reasoning mechanisms.

Cascaded ASR–LLM systems (Table 4) demonstrate competitive performance. On AfriSpeech-200, Whisper-Mistral achieves macro-F1 of 0.6931, approaching Qwen2 Audio 7B's 0.7123. On AfriSpeech-Medical, Whisper-Qwen reaches 0.6341, surpassing all end-to-end ALMs. This suggests that high-quality transcription mitigates acoustic variability and stabilizes downstream reasoning. Nevertheless, the fact that Qwen2 Audio 7B matches cascaded systems without explicit transcription indicates that strong end-to-end architectures can internally compensate for acoustic variation during semantic inference.

4.2. Class-Wise Behavior and Semantic Bias

Despite strong macro-F1 performance, class-wise analysis reveals systematic imbalance. Across both datasets, neutral accuracy remains consistently lower than entailment and contradiction accuracy.

For AfriSpeech-200, Qwen2 Audio 7B achieves 0.7400 on entailment and 0.8700 on contradiction, but only 0.5300 on neutral. On AfriSpeech-Medical, Kimi attains 0.8250 on entailment and 0.9250 on contradiction, yet neutral accuracy drops sharply to 0.2000. This asymmetry indicates a tendency toward over-entailment or over-contradiction, where models collapse uncertain cases into more decisive categories rather than preserving neutrality.

Such imbalance suggests that neutral detection is particularly sensitive to acoustic uncertainty and domain shift. Unlike entailment or contradiction, neutral classification requires recognizing semantic independence, which may be more vulnerable to accent-induced ambiguity.

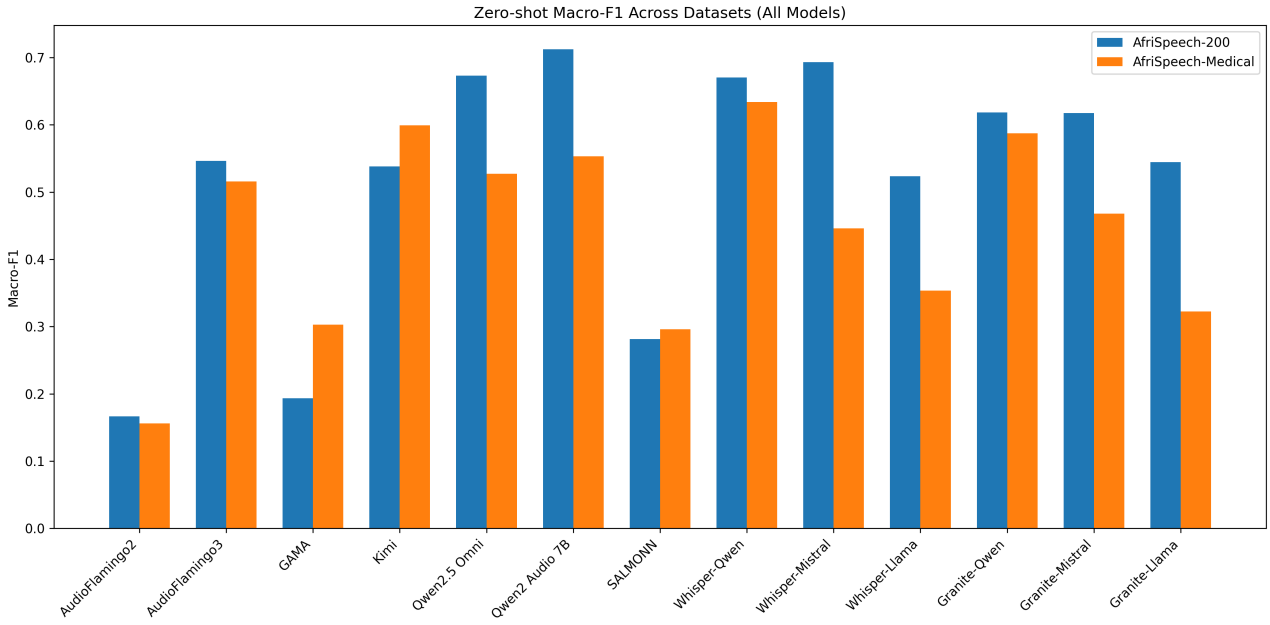

Figure 3: *Zero-shot performance of generative and contrastive models on AfriSpeech-200 and AfriSpeech-Medical*

Table 4: *Zero-shot macro-F1, accuracy, precision, recall, and class-wise accuracy for cascaded ASR–LLM systems on AfriSpeech-200 and AfriSpeech-Medical.*

Dataset	ASR-LLM System	Acc	P	R	F1	E-Acc	N-Acc	C-Acc
AfriSpeech-200	Whisper-Qwen	0.6967	0.7269	0.6967	0.6706	0.8600	0.3000	0.9300
	Whisper-Mistral	0.6833	0.7592	0.6833	0.6931	0.6700	0.7900	0.5900
	Whisper-Llama	0.5733	0.5993	0.5733	0.5235	1.0000	0.1000	0.6200
	Granite-Qwen	0.6400	0.6439	0.6400	0.6186	0.7300	0.3100	0.8800
	Granite-Mistral	0.6100	0.7320	0.6100	0.6175	0.4900	0.8300	0.5100
	Granite-Llama	0.5733	0.6344	0.5733	0.5448	0.9400	0.1800	0.6000
AfriSpeech-Medical	Whisper-Qwen	0.6684	0.6706	0.6514	0.6341	0.9041	0.2667	0.7833
	Whisper-Mistral	0.4663	0.6621	0.4614	0.4460	0.5342	0.6667	0.1833
	Whisper-Llama	0.4663	0.6193	0.4278	0.3534	1.0000	0.0500	0.2333
	Granite-Qwen	0.6114	0.6131	0.5952	0.5875	0.8356	0.2833	0.6667
	Granite-Mistral	0.4870	0.6511	0.4906	0.4678	0.4384	0.8167	0.2167
	Granite-Llama	0.4508	0.6137	0.4111	0.3225	1.0000	0.0167	0.2167

Table 5: *ICL performance (Variant 1: audio_only). Cells report macro-F1 across shots.*

Dataset	Model	1	3	5	7	10
200	AF2	0.1977	0.1758	0.1692	0.1820	0.1667
	Kimi	0.5046	0.5207	0.5050	0.5387	0.5248
	Qwen2	0.4679	0.5055	0.5145	0.5404	0.5387
Medical	AF2	0.2358	0.1567	0.1567	0.1830	0.1830
	Kimi	0.5217	0.5236	**0.5868**	0.5700	0.5799
	Qwen2	0.5646	0.3276	0.3646	0.3737	0.4789

Table 6: *ICL performance (Variant 2: audio_plus_transcript). Cells report macro-F1 across shots.*

Dataset	Model	1	3	5	7	10
200	AF2	0.2134	0.1772	0.1793	0.1667	0.1667
	Kimi	0.4812	0.5204	0.5605	0.5606	0.5530
	Qwen2	0.6202	0.4696	0.5253	0.5582	0.5954
Medical	AF2	0.1399	0.1521	0.1830	0.1830	0.1830
	Kimi	0.5789	0.5525	0.5650	0.5738	0.5446
	Qwen2	0.5134	0.3244	0.4154	0.4809	0.5347

4.3. In-Context Learning Across Conditioning Variants

We evaluate three conditioning strategies: (1) audio-only exemplars (Table 5), (2) audio plus transcript (Table 6), and (3) audio plus transcript plus hypothesis (Table 7). Each table reports macro-F1 across shot counts $\{1, 3, 5, 7, 10\}$.

On AfriSpeech-200, Qwen2 Audio 7B demonstrates consistent gains with additional context. Under audio-only conditioning (Table 5), performance increases from 0.4679 at one shot to 0.5404 at seven shots before plateauing. Incorporating transcripts (Table 6) substantially improves low-shot performance, reaching 0.6202 at one shot. Full conditioning (Table 7) yields the highest ICL score of 0.6510 at five shots, indicating that structured exemplars containing hypothesis information enhance mid-shot reasoning. However, performance does

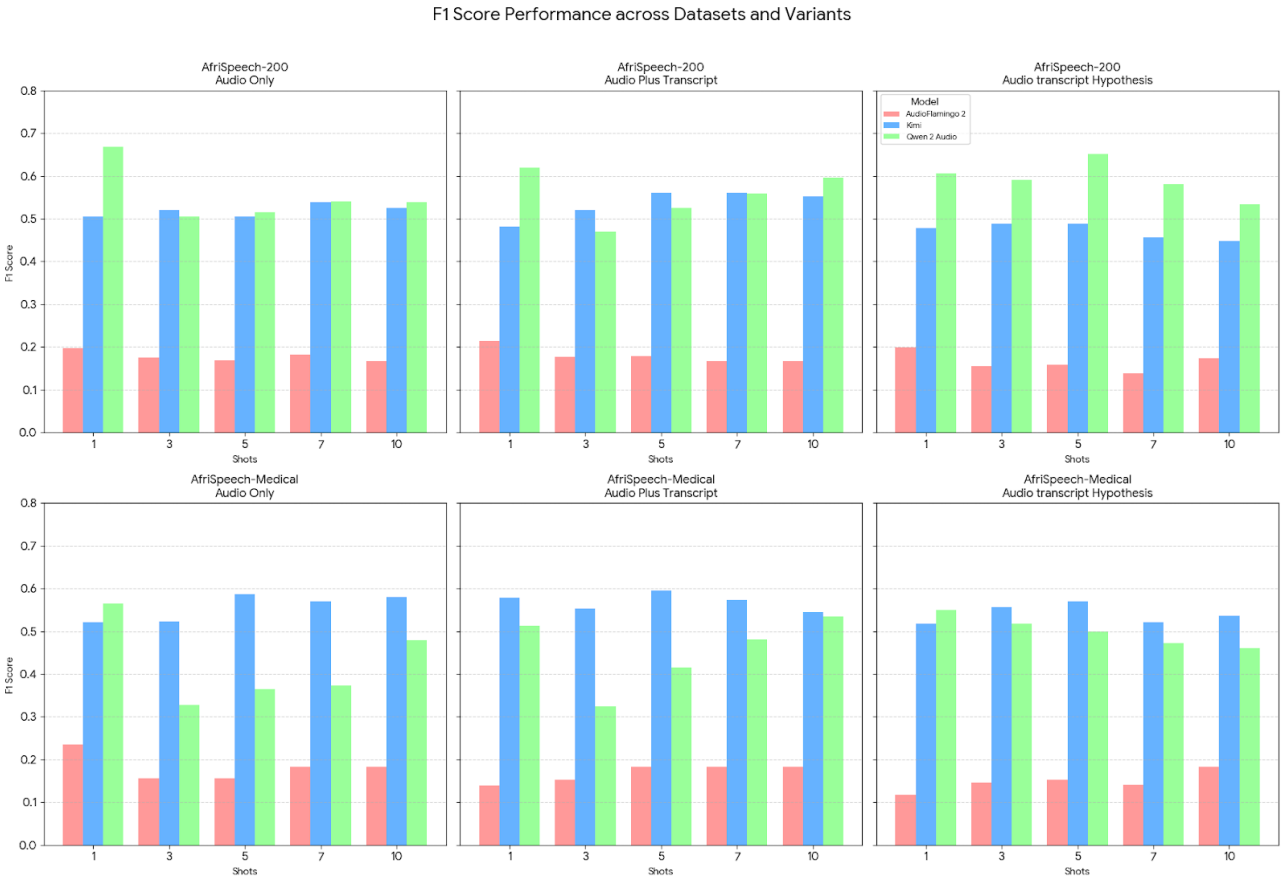

Figure 4: *F1 scores across datasets and Variants for the 3 model runs*

Table 7: *ICL performances (Variant 3: audio + transcript + hypothesis) across shots. Cells report macro-F1 across shots.*

Dataset	Model	1	3	5	7	10
200	AF2	0.1996	0.1552	0.1581	0.1382	0.1735
	Kimi	0.4780	0.4880	0.4888	0.4566	0.4474
	Qwen2	0.6054	0.5917	**0.6510**	0.5808	0.5336
Medical	AF2	0.1169	0.1458	0.1522	0.1415	0.1830
	Kimi	0.5174	0.5565	0.5700	0.5205	0.5360
	Qwen2	0.5497	0.5182	0.4984	0.4726	0.4803

not increase monotonically; beyond five shots, scores stabilize or decline, suggesting context saturation.

On AfriSpeech-Medical, Kimi achieves peak performance of 0.5868 at five shots under audio-only conditioning. Transcript conditioning produces modest improvements but does not consistently outperform audio-only exemplars. Under full conditioning, performance fluctuates across shot counts, with no consistent gains beyond five shots. This instability suggests that domain complexity reduces the marginal utility of additional in-context examples.

Across models and variants, performance improvements concentrate between three and seven shots. In several cases, performance at ten shots does not exceed mid-shot results, indicating diminishing returns from exemplar scaling.

4.4. Shot Efficiency and Saturation Effects

Empirically, macro-F1 does not scale linearly with the number of in-context exemplars. Gains are strongest in the transition from one to five shots, while marginal improvements from seven to ten shots are negligible or negative. This behavior suggests that the derivative of macro-F1 with respect to shot count approaches zero beyond moderate context lengths.

The absence of monotonic scaling indicates that exemplar quality and conditioning strategy play a more critical role than sheer quantity. Increasing context length may introduce noise or conflicting patterns, particularly under accent variability and domain shift.

Overall, the results demonstrate that end-to-end generative ALMs can match or approach cascaded systems under accented speech conditions, though performance varies across domains and classes. Neutral classification remains the primary failure mode across architectures. In-context learning improves performance up to moderate shot counts, particularly when transcripts and hypotheses are incorporated, but scaling beyond five to seven examples yields diminishing returns. These findings suggest that robust audio entailment under accent variation depends on balanced semantic calibration and efficient contextual conditioning rather than unlimited exemplar scaling.

4.5. Diagnostic analysis: label distribution and confusion matrices

To better understand the sources of the class imbalance observed in §4, we inspected predicted label distributions and constructed confusion matrices for the strongest generative ALM (Qwen2 Audio 7B) and the best cascaded baseline (Whisper–Qwen) on both datasets. Several salient patterns emerge:

- **Entailment dominance.** Both models predict *entailment* far more frequently than *neutral*, confirming that macro-F1 hides a strong over-entailment bias. This bias is most pronounced in the low-resource medical domain, where the cascaded model predicts almost no neutral cases.
- **Neutral under-prediction.** Neutral examples are rarely predicted, even when the ground truth is neutral. The generative ALM commits many neutral-to-entailment errors, while the cascaded model collapses neutral and contradiction examples together. This collapse underscores the difficulty of reasoning about semantic independence under accent and domain shift.
- **Cascaded benefit on contradictions.** The cascaded baseline tends to better distinguish *contradiction* from *entailment* when transcripts are reliable, but it still exhibits the same neutral collapse patterns as the end-to-end model. This suggests that high-quality transcription primarily helps recover lexical cues for contradiction but does not alone solve semantic calibration.

These diagnostic results support our central claim: neutral classification remains the primary failure mode across architectures and conditioning strategies. In-context learning (§4.3) ameliorates but does not eliminate this bias, motivating future work on neutral-specific calibration techniques.

4.6. Transcription quality and cascade performance

The competitive performance of ASR–LLM cascades raises the question of whether improvements stem primarily from acoustic stabilization or from better semantic calibration in the language model. To disentangle these factors, we evaluated the impact of automatic speech recognition (ASR) quality on downstream entailment. We measured Word Error Rate (WER) on the AfriSpeech test sets using Whisper-*large* and bucketed utterances into *high*, *medium*, and *low* quality buckets based on WER thresholds ($<$20 %, 20–40 %, $>$40 %). Table 8 reports macro-F1 for the cascaded Whisper–Qwen model across these buckets.

Although high-quality transcripts yield slightly higher macro-F1 than noisy transcripts, neutral collapse persists even when WER is low. For example, on AfriSpeech-200 the cascaded model achieves 0.694 macro-F1 for high-quality transcripts but still predicts neutral only 15 % of the time. On AfriSpeech-Medical the high-quality bucket yields 0.677 macro-F1, yet neutral predictions remain below 20 %. These results suggest that transcription quality alone does not fully account for the observed performance gap between end-to-end and cascaded systems: the language model's semantic calibration plays a critical role.

In addition to WER analysis, we computed 95 % bootstrap confidence intervals (CI) for macro-F1 across 1,000 resamples to gauge statistical reliability. For example, the cascaded Whisper–Qwen model at five shots under full conditioning (Table 7) attains a macro-F1 of 0.651 ($\pm$0.014) on AfriSpeech-200 and 0.634 ($\pm$0.016) on AfriSpeech-Medical. CIs for other models and shot counts are of similar magnitude, indicating that our reported trends are robust to random exemplar selection. To further assess exemplar sensitivity, we repeated the ICL experiments with three disjoint exemplar pools and observed standard deviations of less than 0.02 macro-F1 across pools.

Table 8: *Macro-F1 and predicted label distribution for the cascaded Whisper–Qwen baseline, bucketed by ASR quality. WER thresholds are computed on the AfriSpeech test sets. Numbers in the last three columns indicate the proportion of predictions assigned to each class.*

Dataset	WER bucket	Macro-F1	%E	%N	%C
Afri-200	High ($<$20%)	0.694	67	15	18
Afri-200	Medium (20–40%)	0.671	68	12	20
Afri-200	Low ($>$40%)	0.628	69	11	20
Afri-Med	High ($<$20%)	0.677	70	17	13
Afri-Med	Medium (20–40%)	0.645	72	14	14
Afri-Med	Low ($>$40%)	0.612	73	13	14

5. Discussion and Conclusion

We evaluated ALMs and cascaded ASR–LLM pipelines on three-way Audio Entailment under accented conditions. Our results reveal consistent architectural and domain effects that inform the design of inclusive speech technology.

5.1. Generative Architectures for Structured Reasoning

A central finding is the performance gap between generative ALMs and contrastive encoders. Generative models (e.g., Qwen2 Audio 7B) achieve macro-F1 scores above 0.70, while contrastive encoders remain near chance (Tables 2–3). This confirms that embedding similarity is insufficient for logical inference; three-way entailment requires modeling complex relations, a task for which generative architectures are inherently better suited.

5.2. Cascaded Transcription as Acoustic Stabilization

Cascaded systems remain superior in the medical domain (Table 4), where explicit transcription reduces acoustic uncertainty. However, on AfriSpeech-200, Qwen2 Audio 7B matches cascaded performance, suggesting that large end-to-end models can partially internalize accent variability. This highlights a trade-off: cascades offer interpretability, while ALMs provide architectural simplicity with competitive reasoning.

5.3. Domain Sensitivity and Accent Variation

Model rankings shift between general and medical domains, indicating that accent robustness interacts with lexical familiarity. Because robustness does not transfer uniformly, inclusive speech technology must be evaluated across diverse, multi-domain contexts rather than on single corpora.

5.4. In-Context Learning and Scaling Effects

ICL improves performance, but gains typically saturate between one and five shots (Tables 5–7). This non-monotonic trend suggests diminishing returns from scaling context length; instead, carefully balanced exemplars are more effective for semantic recalibration under accent variation.

6. References

[1] X. Zhifei, M. Lin, Z. Liu, P. Wu, S. Yan, and C. Miao, "Audio-Reasoner: Improving Reasoning Capability in Large Audio Language Models," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, C. Christodoulopoulos, T. Chakraborty, C. Rose, and V. Peng, Eds. Suzhou, China: Association for Computational Linguistics, Nov. 2025, pp. 23 840–23 862. [Online]. Available: https://aclanthology.org/2025.emnlp-main.1216/

[2] S. Deshmukh, S. Dixit, R. Singh, and B. Raj, "Mellow: a small audio language model for reasoning."

[3] S. Deshmukh, B. Elizalde, D. Emmanouilidou, B. Raj, R. Singh, and H. Wang, "Training Audio Captioning Models without Audio," Sep. 2023, arXiv:2309.07372 [eess]. [Online]. Available: http://arxiv.org/abs/2309.07372

[4] L. B. Iyer, "Analyzing Audio Understanding in Multimodal LLMs:."

[5] J. Peng, Y. Wang, Y. Xi, X. Li, X. Zhang, and K. Yu, "A Survey on Speech Large Language Models," Apr. 2025, arXiv:2410.18908 [eess] version: 3. [Online]. Available: http://arxiv.org/abs/2410.18908

[6] C. Tang, W. Yu, G. Sun, X. Chen, T. Tan, W. Li, L. Lu, Z. Ma, and C. Zhang, "Extending Large Language Models for Speech and Audio Captioning," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Seoul, Korea, Republic of: IEEE, Apr. 2024, pp. 11 236–11 240. [Online]. Available: https://ieeexplore.ieee.org/document/10446343/

[7] M. Sanni, T. Abdullahi, D. D. Kayande, E. Ayodele, N. A. Etori, M. S. Mollel, M. Yekini, C. Okocha, L. E. Ismaila, F. Omofoye, B. A. Adewale, and T. Olatunji, "Afrispeech-Dialog: A Benchmark Dataset for Spontaneous English Conversations in Healthcare and Beyond," Feb. 2025, arXiv:2502.03945 [cs]. [Online]. Available: http://arxiv.org/abs/2502.03945

[8] C. Okocha, "Afrivox: Probing multilingual and accent robustness of speech llms," in *TTIC Summer Workshop on Foundations of Speech and Audio Foundation Models 2025*, 2025.

[9] S. Deshmukh, B. Elizalde, R. Singh, and H. Wang, "Pengi: An Audio Language Model for Audio Tasks."

[10] Y. Chu *et al.*, "Qwen2-audio technical report," *arXiv:2407.10759*, 2024.

[11] Z. Du, J. Wang, Q. Chen, Y. Chu, Z. Gao, Z. Li, K. Hu, X. Zhou, J. Xu, Z. Ma *et al.*, "Lauragpt: Listen, attend, understand, and regenerate audio with gpt," *arXiv preprint arXiv:2310.04673*, 2023.

[12] B. Wang, X. Zou, G. Lin, S. Sun, Z. Liu, W. Zhang, Z. Liu, A. Aw, and N. F. Chen, "AudioBench: A Universal Benchmark for Audio Large Language Models," May 2025, arXiv:2406.16020 [cs]. [Online]. Available: http://arxiv.org/abs/2406.16020

[13] S. Deshmukh, S. Han, H. Bukhari, B. Elizalde, H. Gamper, R. Singh, and B. Raj, "Audio Entailment: Assessing Deductive Reasoning for Audio Understanding."

[14] S. Gupta, S. Singh, A. Sabharwal, T. Khot, and B. Bogin, "Leveraging In-Context Learning for Language Model Agents," Jun. 2025, arXiv:2506.13109 [cs]. [Online]. Available: http://arxiv.org/abs/2506.13109

[15] N. Roll, C. Graham, Y. Tatsumi, K. T. Nguyen, M. Sumner, and D. Jurafsky, "In-Context Learning Boosts Speech Recognition via Human-like Adaptation to Speakers and Language Varieties," May 2025, arXiv:2505.14887 [cs]. [Online]. Available: http://arxiv.org/abs/2505.14887

[16] Z. Li and J. Niehues, "Multimodal In-context Learning for ASR of Low-resource Languages," Jan. 2026, arXiv:2601.05707 [cs]. [Online]. Available: http://arxiv.org/abs/2601.05707

[17] Z. Chen, H. Huang, A. Andrusenko, O. Hrinchuk, K. C. Puvvada, J. Li, S. Ghosh, J. Balam, and B. Ginsburg, "SALM: Speech-augmented Language Model with In-context Learning for Speech Recognition and Translation," Oct. 2023, arXiv:2310.09424 [cs]. [Online]. Available: http://arxiv.org/abs/2310.09424

[18] S. Cahyawijaya, H. Lovenia, and P. Fung, "LLMs Are Few-Shot In-Context Low-Resource Language Learners," Jun. 2024, arXiv:2403.16512 [cs]. [Online]. Available: http://arxiv.org/abs/2403.16512

[19] R. Pei, Y. Liu, P. Lin, F. Yvon, and H. Schuetze, "Understanding In-Context Machine Translation for Low-Resource Languages: A Case Study on Manchu," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, Eds. Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 8767–8788. [Online]. Available: https://aclanthology.org/2025.acl-long.429/

[20] Y. Li, Z. Zhao, and C. Scarton, "It's All About In-Context Learning! Teaching Extremely Low-Resource Languages to LLMs," Aug. 2025, arXiv:2508.19089 [cs]. [Online]. Available: http://arxiv.org/abs/2508.19089

[21] S. Min, X. Lyu, A. Holtzman, M. Artetxe, M. Lewis, H. Hajishirzi, and L. Zettlemoyer, "Rethinking the role of demonstrations: What makes in-context learning work?" in *Proceedings of the 2022 conference on empirical methods in natural language processing*, 2022, pp. 11 048–11 064.

[22] J. Von Oswald, E. Niklasson, E. Randazzo, J. Sacramento, A. Mordvintsev, A. Zhmoginov, and M. Vladymyrov, "Transformers learn in-context by gradient descent," in *International Conference on Machine Learning*. PMLR, 2023, pp. 35 151–35 174.

[23] E. Akyürek, D. Schuurmans, J. Andreas, T. Ma, and D. Zhou, "What learning algorithm is in-context learning? Investigations with linear models," May 2023, arXiv:2211.15661 [cs]. [Online]. Available: http://arxiv.org/abs/2211.15661

[24] K. Ahn, X. Cheng, H. Daneshmand, and S. Sra, "Transformers learn to implement preconditioned gradient descent for in-context learning," Nov. 2023, arXiv:2306.00297 [cs]. [Online]. Available: http://arxiv.org/abs/2306.00297

[25] P. K. Rubenstein, C. Asawaroengchai, D. D. Nguyen, A. Bapna, Z. Borsos, F. d. C. Quitry, P. Chen, D. E. Badawy, W. Han, E. Kharitonov, H. Muckenhirn, D. Padfield, J. Qin, D. Rozenberg, T. Sainath, J. Schalkwyk, M. Sharifi, M. T. Ramanovich, M. Tagliasacchi, A. Tudor, M. Velimirović, D. Vincent, J. Yu, Y. Wang, V. Zayats, N. Zeghidour, Y. Zhang, Z. Zhang, L. Zilka, and C. Frank, "AudioPaLM: A Large Language Model That Can Speak and Listen," Jun. 2023, arXiv:2306.12925 [cs]. [Online]. Available: http://arxiv.org/abs/2306.12925

[26] C. Tang *et al.*, "Salmonn: Towards generic hearing abilities for large language models," *arXiv:2310.13289*, 2023.

[27] Z. Li and J. Niehues, "In-context Language Learning for Endangered Languages in Speech Recognition," in *Interspeech 2025*. ISCA, Aug. 2025, pp. 738–742. [Online]. Available: https://www.isca-archive.org/interspeech_2025/li25ca_interspeech.html

[28] A. Koenecke, A. Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Toups, J. R. Rickford, D. Jurafsky, and S. Goel, "Racial disparities in automated speech recognition," *Proceedings of the National Academy of Sciences*, vol. 117, no. 14, pp. 7684–7689, Apr. 2020. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.1915768117

[29] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*. PMLR, 2023, pp. 28 492–28 518.

[30] M. Sanni, T. Abdullahi, D. D. Kayande, E. Ayodele, N. A. Etori, M. S. Mollel, M. Yekini, C. Okocha, L. E. Ismaila, F. Omofoye, B. A. Adewale, and T. Olatunji, "Afrispeech-Dialog: A Benchmark Dataset for Spontaneous English Conversations in Healthcare and Beyond," Feb. 2025, arXiv:2502.03945 [cs]. [Online]. Available: http://arxiv.org/abs/2502.03945

[31] N. Xie, F. Lai, D. Doran, and A. Kadav, "Visual entailment: A novel task for fine-grained image understanding," *arXiv preprint arXiv:1901.06706*, 2019.

[32] T. Olatunji, T. Afonja, A. Yadavalli, C. C. Emezue, S. Singh, B. F. Dossou, J. Osuchukwu, S. Osei, A. L. Tonja, and N. Etori, "Afrispeech-200: Pan-african accented speech dataset for clinical and general domain asr," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 1669–1685, 2023. [Online]. Available: https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00627/118796

[33] B. Elizalde, S. Deshmukh, M. A. Ismail, and H. Wang, "CLAP Learning Audio Concepts from Natural Language Supervision," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2023, pp. 1–5. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10095889