Afrispeech Semantics: Evaluating Audio–Semantic Reasoning in Spoken Language Models Across Domains and Accents

Anonymous ACL submission

Abstract

Audio language models (ALMs) are increasingly used for speech-based understanding; yet, their ability to perform semantic reasoning beyond transcription, Text-to-Audio Retrieval, Captioning, and Question-Answering accuracy remains insufficiently benchmarked. In particular, the effects of accent variation, domain shift, and semantic over-inference on audio reasoning are poorly understood. We evaluate audio language models across five semantic and paralinguistic reasoning tasks: entailment, consistency, plausibility, accent drift, and accent restraint. Collectively, these tasks assess a model's ability to reason over spoken audio as the primary evidence source, including whether a textual hypothesis can be inferred, contradicted, or left undetermined by the audio, whether statements align or conflict with spoken content, whether claims are plausible given the discourse, and whether model predictions remain stable or appropriately constrained across accent variation. These findings highlight critical limitations in current audio reasoning evaluations and hope to provide guidance for more robust and equitable ALM design and assessment.

1 Introduction

Recent multimodal models, commonly referred to as audio language models (ALMs), are trained on large collections of audio–text pairs using either contrastive learning (Elizalde et al., 2023) or next-token prediction objectives (Chu et al., 2024; Tang et al., 2023; Goel et al., 2025; KimiTeam et al., 2025). Once trained, ALMs can be prompted to perform a wide range of tasks grounded in audio, including captioning, retrieval, and question answering, and have demonstrated strong performance across many established benchmarks (Sakshi et al., 2024; Wang et al., 2024).

Despite these advances, most existing evaluations primarily focus on surface-level correctness rather than semantic reasoning grounded in the audio signal (Wang et al., 2024; Yang et al., 2025). In open-ended settings, ALMs are often rewarded for producing plausible responses, even when those responses rely on contextual assumptions or linguistic priors rather than evidence present in the audio (Chiang et al., 2025). This limitation becomes particularly problematic for interactive and reasoning-oriented applications, where models are expected to infer what can—and cannot—be concluded from what is heard (Sanni et al., 2025).

To address this gap, prior work introduced audio entailment as a focused task for evaluating deductive reasoning in audio language models (Deshmukh et al.), framing the problem as determining whether a textual hypothesis is entailed, contradicted, or unsupported by an audio premise. While this formulation provides a principled starting point, it captures only a narrow slice of the reasoning challenges faced by ALMs. Existing benchmarks are limited in domain diversity and do not explicitly test failure modes such as semantic over-inference, robustness to unfamiliar named entities, or sensitivity to accent and pronunciation variation (Shi et al., 2024; Wang et al., 2024; Sakshi et al., 2024).

In this work, we expand the evaluation of audio reasoning beyond a single entailment task or deductive reasoning task. We introduce a unified semantic reasoning framework comprising several domain-diverse speech datasets, each paired with task formulations tailored to the semantic properties of the domain.

Using this framework, we benchmark state-of-the-art contrastive and next-token prediction audio language models under a controlled inference protocol. Our results reveal consistent patterns of over-entailment, domain-specific reasoning failures, and accent-conditioned semantic drift, even when transcription quality is high. These findings suggest that current benchmarks substantially underesti-

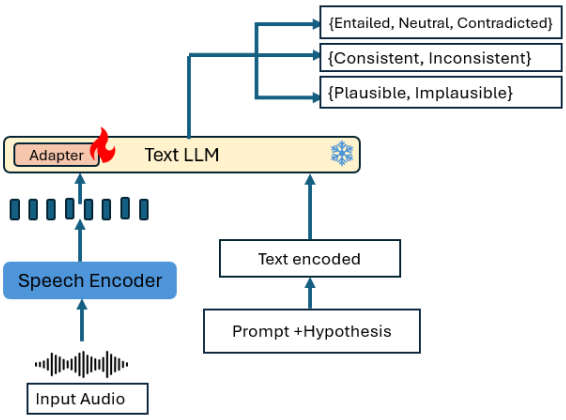

Figure 1: Overview of Afrispeech Semantics.

mate reasoning errors in audio language models, underscoring the need for more comprehensive, domain-aware evaluation of semantic reasoning from audio.

In this work, we study logical and semantic reasoning for ALMs. Our contributions are as follows:

- We introduce four African-accented domain-diverse datasets that include linguistically diverse and accented audio, enabling the study of semantic reasoning beyond standard speech conditions.
- We propose multiple task formulations designed to test whether models preserve meaning under pronunciation variation rather than relying on contextual priors.
- Hypotheses are generated using LLaMA model and systematically verified and corrected by human annotators, ensuring semantic validity and grounding in the audio evidence, including accented and unfamiliar speech patterns.
- We benchmark both contrastive and next-token prediction audio language models under a unified inference protocol, revealing consistent over-entailment and accent-sensitive reasoning failures that are not captured by existing benchmarks.

2 Related works

Audio–language models combine acoustic perception with language modeling to support a broad range of audio understanding tasks (Ghosh et al., 2025, 2024). Early approaches focused primarily on contrastive learning to align audio and text embeddings, enabling strong performance on retrieval and classification tasks (Elizalde et al., 2023). More recent work has explored next-token prediction frameworks that treat audio understanding as a conditional text generation problem, allowing models to perform open-ended tasks such as captioning, dialogue, and question answering (Deshmukh et al., 2023). These advances have substantially expanded the functional scope of ALMs; however, evaluation has largely emphasized task completion or linguistic plausibility rather than the semantic validity of the inferred conclusions (Peng et al., 2025).

As ALMs have become more capable of generating free-form responses, concerns have emerged regarding their reliance on contextual priors and language statistics rather than evidence grounded in the audio signal (Sakshi et al., 2024). Existing audio question answering and captioning benchmarks often permit multiple acceptable outputs, making it difficult to distinguish between correct inference and plausible hallucination (Chu et al., 2023). As a result, models may appear to perform well despite systematically over-interpreting or misattributing audio events (Shi et al., 2024; Kubis et al., 2025). These limitations have motivated interest in more principled evaluation frameworks that explicitly test reasoning behavior rather than surface-level alignment (Kubis et al., 2025).

Audio entailment was introduced as a structured task to evaluate deductive reasoning in audio–language models by determining whether a textual hypothesis is entailed, contradicted, or unsupported by an audio premise (Deshmukh et al.). By framing audio understanding as a three-way inference problem, this work provided an important step toward disentangling reasoning from generation quality and revealed significant reasoning deficiencies in state-of-the-art models. However, the focus of audio entailment is intentionally narrow: it evaluates a single reasoning formulation over a limited set of domains and does not explicitly probe how reasoning failures vary across domain types, semantic phenomena, or speech characteristics such as accent variation.

Related work in vision–language and text-based inference has shown that single-task benchmarks can underestimate reasoning errors and mask systematic failure modes (Sarkar et al., 2025; Cardoso et al.). Multi-task and domain-aware evaluation has proven essential for uncovering issues such as over-generalization, reliance on world knowledge,

Task	Reasoning	Complexity
Entailment	Deductive semantic	■■■■
Consistency	Logical compatibility	■■
Plausibility	Pragmatic semantic	■■
Accent Drift	Robustness	■■■
Accent Restraint	Causal restraint	■■■■■

Table 1: Task taxonomy by reasoning type and the complexity of each reasoning type.

Dataset	Hours (h)	Accent Coverage
AfriSpeech-200[1]	200	13
AfriSpeech-General[2]	2.07	8
Afri-Names[3]	8.92	12
Afrispeech - Medical[4]	4.20	6

Table 2: Overview of the four speech corpora used in this work. Durations are measured after pruning unusable segments. Detailed statistics and split methodology are provided in Appendix A.

and sensitivity to spurious correlations (Olatunji et al., 2023; Ogundepo et al., 2023). In the audio domain, however, comparable multi-task semantic reasoning evaluations remain underdeveloped. Prior benchmarks do not systematically test semantic restraint, robustness to unfamiliar named entities, or accent-conditioned meaning drift phenomena that are especially relevant for spoken language understanding in diverse, real-world settings (Wang et al., 2025; Sakshi et al., 2024).

3 Semantic Reasoning Tasks for Audio Language Models

3.1 Problem Setup and Notation

We study semantic reasoning in audio language models through tasks that require models to draw conclusions from spoken audio. Each task is formulated as a relationship between an audio recording and a textual hypothesis, as illustrated in fig. 1. Let a denote an audio recording serving as the *premise*, and let h denote a natural language *hypothesis*. Given a pair (a, h), the model predicts the semantic relationship between the hypothesis and the content expressed in the audio. An overview of the reasoning types associated with each task is provided in Table 1.

[1] https://huggingface.co/datasets/intronhealth/afrispeech-200

[2] https://huggingface.co/datasets/intronhealth/afrispeech-dialog

[3] https://huggingface.co/datasets/intronhealth/afri-names

[4] https://huggingface.co/datasets/intronhealth/med-convo-nig

Following prior work on audio entailment (Deshmukh et al.), we restrict inference to information supported by the audio signal itself. Models are not permitted to assume unstated facts or rely on external world knowledge beyond what can be reasonably inferred from the spoken content.

3.2 Audio Entailment

We adopt audio entailment as the core semantic reasoning task. The objective is to determine whether a hypothesis is supported, contradicted, or left undetermined by an audio premise. Specifically, each (a, h) pair is assigned one of three labels:

- **Entailment (E):** The audio provides sufficient evidence to support the hypothesis.
- **Neutral (N):** The audio does not provide enough information to determine the truth of the hypothesis.
- **Contradiction (C):** The audio provides sufficient evidence to refute the hypothesis.

Formally, the task is defined as

$$f(a, h) \rightarrow y, \quad y \in \{E, N, C\}.$$

This task evaluates deductive semantic reasoning grounded in spoken language. Unlike text-based entailment, audio entailment requires models to reason over acoustic realizations of meaning, including speaker variation and prosodic cues, while maintaining strict evidence-based inference.

3.3 Plausibility and Consistency

Beyond entailment, models may conflate semantic compatibility with evidential support. We therefore introduce plausibility and consistency tasks that further probe the boundaries of inference.

In the **plausibility** task, hypotheses are constructed to be reasonable given commonsense knowledge or discourse context, but are neither stated nor implied by the audio. The task assesses whether models incorrectly accept such hypotheses based solely on plausibility.

In the **consistency** task, hypotheses are either semantically compatible or incompatible with the spoken content. Unlike entailment, this task does not admit a neutral option; instead, it focuses on detecting agreement or contradiction with the audio premise. Together, these tasks assess whether models rely on semantic evidence rather than commonsense priors when making judgments.

3.4 Accent-Conditioned Semantic Drift and Accent Restraint

Spoken language varies substantially across speakers and accents. To evaluate robustness under such variation, we introduce tasks that examine accent-conditioned semantic drift. In this setting, audio recordings differ in accent or pronunciation while preserving equivalent semantic content, and hypotheses remain unchanged. The task assesses whether accent variation systematically alters semantic predictions.

We further introduce an accent restraint task to test whether models appropriately suppress accent-based cues when accent is semantically irrelevant. Here, accent functions as a nuisance variable, and correct behavior requires invariant semantic judgments across accented realizations.

4 Dataset Construction and Annotation

In this section we describe how we curated and annotated the data used throughout our evaluation. Rather than treating this step as a minor implementation detail, we view careful dataset construction as central to any fair benchmark. Our goal was to build a suite of tasks that capture real-world variation in both content and pronunciation, while ensuring that every hypothesis is grounded in the acoustic evidence. Table 2 gives a high-level overview of the four underlying corpora. A more detailed description of the speaker demographics, transcript quality and split methodology can be found in appendix A. We provide qualitative examples of these tasks in Figure 2.

4.1 Audio Premises

Each dataset is built around a collection of audio premises a paired with one or more textual hypotheses. The corpora span distinct domains including conversational telephone speech, read speech focused on named entities, and clinical dialogues to elicit a variety of reasoning behaviours. Importantly, the recordings cover a wide range of African accents and dialects, meaning models must contend with both linguistic and paralinguistic variation rather than idealised laboratory speech. We selected datasets whose original annotations were created from scratch by listening to the audio, thereby avoiding the confounding effect of textual metadata. Further details on duration, speaker demographics and accent distribution are provided in appendix A.

4.2 Hypothesis Generation

For every audio premise we create a small set of hypotheses that probe different semantic relationships: whether a statement is entailed by the audio, contradicted by it, merely plausible under common sense, or tests for accent-induced drift. To scale this process, we first use an LLM to propose candidate hypotheses under carefully crafted prompts that forbid unsupported negation and extraneous world knowledge. These prompts are reproduced verbatim in appendix I. Crucially, the LLM proposals are only a starting point.

4.3 Human Verification and Correction

After automatic generation, each candidate hypothesis is vetted by trained human annotators. Annotators listen to the audio in full and judge whether the proposed statement is entailed, contradicted, or unsupported by the recording. If a hypothesis contains hallucinated details or ambiguous phrasing, annotators edit or replace it to ensure that the final set of hypotheses is both semantically precise and audibly grounded. This manual verification is particularly critical when dealing with accented speech or rare proper names, where large language models often introduce subtle semantic drift. Our annotation protocol—including guidelines, quality control procedures and examples—is described in appendix B. By combining LLM-assisted generation with human oversight we strike a balance between scalability and data integrity.

5 Experimental Setup

This section describes the models, inference protocols, and evaluation procedures used to assess semantic reasoning in audio language models. Our setup is designed to enable fair comparison across model families while isolating reasoning behavior from transcription or generation quality.

5.1 Audio Language Models

We evaluate two broad classes of audio language models (ALMs): *contrastive* models and *next-token prediction* models. Contrastive ALMs learn joint audio–text representations and are commonly used for retrieval and classification tasks (Elizalde et al., 2023). Next-token prediction ALMs condition on audio and text inputs to generate free-form textual responses (Deshmukh et al., 2023). All models used are opensource models and also state of the art models.

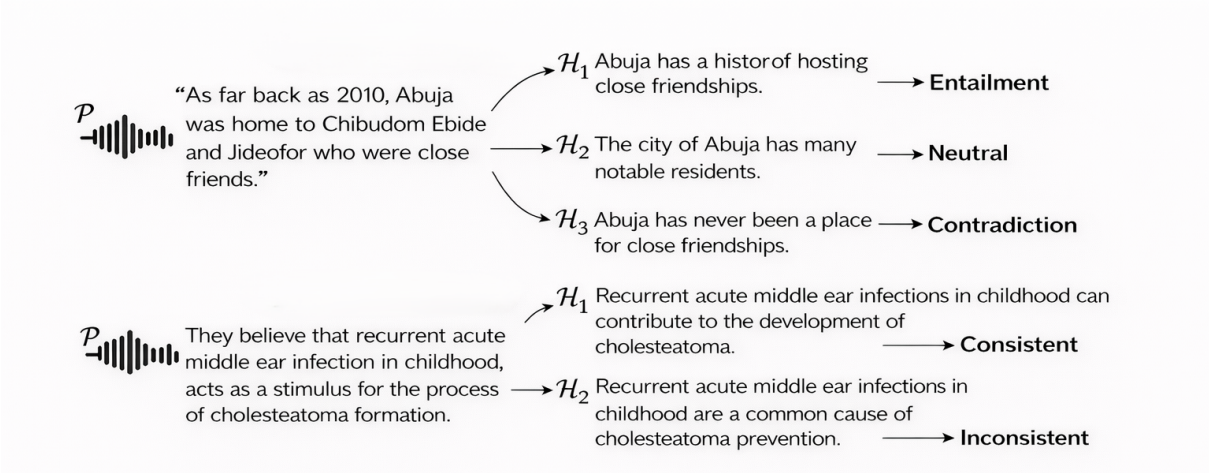

Figure 2: Examples of two AfriSpeech-derived audio reasoning tasks. Each example uses an audio premise P (shown as a waveform icon and the verbatim transcript) paired with hypotheses H_1, H_2, and H_3 (or fewer, depending on the task). For Spoken Entailment, hypotheses are labeled ENTAILMENT, NEUTRAL, or CONTRADICTION. For Medical Consistency, hypotheses are labeled CONSISTENT or INCONSISTENT.

For contrastive models, the audio premise and hypothesis are encoded separately, and their similarity is used to predict the semantic relationship between them. For next-token prediction models, the audio premise and hypothesis are provided as input, and the model generates a textual response indicating its judgment. We evaluate ten Audio language models in total, three of which are constractive models and seven are next-token prediction models.

We evaluate all models in a zero-shot setting unless otherwise stated. Model specifications and checkpoints are summarised in Table 12, and inference hyperparameters are listed in Table 14. These tables provide the necessary context to reproduce our experiments.

5.2 Prompt Selection

We observed that some models are sensitive to prompt phrasing for the semantic consistency task. To avoid cherry-picking, we pre-defined three prompt variants and selected the variant that achieved the best macro F1 on a held-out development subset, aggregated across datasets. Table 17 reports the comparison, and Appendix I lists the full prompt templates used. We use the selected prompt variant for all evaluations.

5.3 Inference Protocol

To ensure comparability across model types, we adopt a unified inference protocol aligned with the task formulation in Section 3. For contrastive models, similarity scores between audio and hypothesis embeddings are mapped to entailment, neutral, or contradiction labels using fixed thresholds derived from a validation set. Threshold selection procedures follow standard practice and are detailed in the appendix I.

5.4 Evaluation and Label Mapping

Evaluating free-form outputs from next-token prediction models presents challenges due to variability in phrasing and instruction-following behavior. To address this, we adopt a model-based evaluation strategy in which a lightweight language model maps generated responses to discrete labels. This approach is validated against a human-annotated subset and achieves high agreement.

For contrastive models, predictions are obtained directly from thresholded similarity scores, eliminating the need for additional post-processing. All evaluations are conducted at the hypothesis level and aggregated at the dataset and task level.

5.5 Metrics

We report standard classification metrics, including accuracy, precision, recall, and macro-averaged F1 score. For multi-hypothesis settings, metrics are computed across all hypotheses associated with an audio premise. In addition, we report class-wise performance to analyze asymmetric error patterns, such as over-entailment and under-detection of contradiction.

To examine how reasoning performance varies across task types, we report results separately for entailment, semantic restraint, plausibility consis-

tency, and accent-conditioned semantic drift tasks. Detailed breakdowns and statistical summaries are provided in the appendix C.

5.6 Reproducibility

All datasets, fixed splits, inference prompts, and evaluation scripts will be released to support reproducibility. Hyperparameters and implementation details necessary to replicate our results are documented in the table 14.

6 Results

To interpret the scores listed in the following tables, we first provide a high-level summary of model behaviour. We evaluate eleven generative audio language models and three contrastive models across four corpora and five semantic reasoning tasks. Our primary metrics are accuracy and macro–averaged F1, supplemented by class-specific accuracies to detect asymmetric error patterns. In the aggregate, generative models substantially outperform contrastive approaches on tasks that require nuanced reasoning, though contrastive baselines offer useful lower bounds.

6.1 Overall performance across tasks

Across all datasets, we observe a wide spread of performance. On the AfriSpeech-200 corpus, the strongest generative models (e.g., Qwen2.5Omni and Qwen2AudioInstruct) achieve macro F1 scores above 0.82 on plausibility and consistency tasks, while weaker baselines such as SALMONN lag markedly. Results on the AfriSpeech-Gen and Medical sets show similar trends, with Qwen2-based models achieving F1 scores around 0.85–0.95 in many settings. The contrastive models (LAION-CLAP and MSCLAP variants) perform near chance on most reasoning tasks, highlighting the limitations of embedding-only approaches for fine-grained semantics. Detailed per-task and per-model results are tabulated below. Qualitative examples comparing model outputs on these tasks are presented in Table 13 (Appendix E).

6.2 Task-wise analysis

Audio Plausibility and Consistency. The audio plausibility task measures whether models can distinguish statements that are plausible but unsupported by the audio from those that are implausible. As shown in Tables 3 and 7, generative models tend to accept plausible but unsupported statements, with accuracy ranges from 0.41 to 0.97. Qwen2.5Omni and AudioFlamingo models often over-accept plausible statements, indicating a propensity for hallucination. Consistency, on the other hand, probes whether a hypothesis aligns or conflicts with the audio without a neutral option. Generative models again dominate, with Kimi-Audio and Qwen2.5Omni achieving F1 scores above 0.8 on AfriSpeech-Gen (Table 5). Full results for contrastive baselines on these tasks are provided in Table 7 and Table 8.

Audio Entailment. The audio entailment task asks whether a hypothesis is entailed, contradicted, or unsupported by the audio premise. Table 4 shows that even the best models struggle on this three-way classification: F1 scores rarely exceed 0.71 on AfriSpeech-200, with frequent confusion between neutral and contradiction. Contrastive models perform poorly, reflecting their limited capacity to discriminate fine-grained semantic relations (Table 6). Qwen2-based models outperform others, but still exhibit a tendency to over-entail, as evidenced by lower neutral and contradiction accuracies.

Accent Drift and Restraint. The AfriNames subset isolates reasoning under accent variation and low-semantic-content utterances. Table 9 reports bias rates (fraction of accent-sensitive inferences) and accept rates (fraction of accent-invariant predictions). Qwen2AudioInstruct shows the lowest bias (33.25%) while AudioFlamingo2 exhibits substantial accent bias (96.75%). The restraint evaluation in Table 10 underscores that models often hallucinate unsupported content; SALMONN and LAION-CLAP produce hallucinations in more than 60% of cases. Models such as Qwen2.5Omni strike a better balance between suppressing hallucination and affirming supported content, achieving a semantic restraint accuracy (SRA) of 66.0%.

7 Discussion

The empirical results prompt a number of takeaways about the current state of audio–semantic reasoning. First, even the best-performing models exhibit systematic over-entailment: hypotheses that are plausible but not entailed by the audio are frequently marked as true. This suggests that models rely heavily on linguistic priors and commonsense knowledge rather than strictly grounding their predictions in the acoustic evidence. Second, domain shift remains a challenge. Performance on the Med-

Dataset	ALM	Acc ↑	P ↑	R ↑	F1 ↑	Acc_P ↑	Acc_I ↑
Afri-200	AudioFlamingo2	0.2585	0.2548	0.2573	0.2560	0.5146	0.0000
	AudioFlamingo3	0.9512	0.9520	0.9511	**0.9512**	0.9709	0.9314
	GAMA	0.4390	0.8211	0.4374	0.4469	0.7670	0.1078
	Kimi	0.6927	0.7878	0.6941	0.6659	0.4078	0.9804
	Qwen2.5 Omni	0.8293	0.8723	0.8301	0.8244	0.6602	1.0000
	Qwen2 Audio 7B	0.8146	0.8722	0.8155	0.8180	0.6408	0.9902
	SALMONN	0.6439	0.7630	0.6455	0.5999	0.3107	0.9804
Afri-Gen	AudioFlamingo2	0.3707	0.2590	0.3707	0.3050	0.7414	0.0000
	AudioFlamingo3	0.9138	0.9552	0.9138	0.9339	0.9310	0.8966
	GAMA	0.4741	0.8597	0.4741	0.5529	0.7414	0.2069
	Kimi	0.8103	0.9194	0.8103	0.8506	0.9655	0.6552
	Qwen2.5 Omni	0.9569	0.9570	0.9569	**0.9569**	0.9483	0.9655
	Qwen2 Audio 7B	0.8103	0.8523	0.8103	0.8138	0.6724	0.9483
	SALMONN	0.6724	0.8021	0.6724	0.6330	0.3448	1.0000

Table 3: Zero-shot performance on Audio Plausibility (Generative Models). ↑ indicates higher is better.

Dataset	ALM	Acc ↑	P ↑	R ↑	F1 ↑	E-Acc ↑	N-Acc ↑	C-Acc ↑
Afri-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
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 4: Zero-shot performance on Audio Entailment (Generative Models). ↑ indicates higher is better.

ical corpus, which contains spontaneous clinical dialogues with colloquial phrasing, lags behind performance on the read-speech AfriSpeech-Gen set. Models that excel on one domain may falter on another, indicating that domain-specific fine-tuning or broader training corpora may be necessary.

Accent variation further complicates semantic reasoning. The accent-drift experiments reveal that models often infer different meanings from semantically equivalent utterances when those utterances are spoken with different accents. Such bias underscores the risk of deploying audio systems in multilingual, multicultural settings without careful calibration. The accent restraint task, which measures hallucination on low-semantic-content utterances, shows that most models readily invent content when little is present. Among generative models, Qwen2-based variants strike a better balance between sensitivity and restraint, but there is ample room for improvement.

Finally, the stark contrast between generative and contrastive models highlights the limits of embedding-only approaches for semantic reasoning. Although contrastive models perform well on retrieval-style tasks, they struggle to make fine-grained entailment judgments. Future work might explore hybrid architectures or training objectives that explicitly align representations with reasoning labels. We hope that the benchmark and analyses presented here will spur more nuanced evaluation and lead to audio language models that reason as carefully as they speak.

8 Caption-before-Reasoning

To explore whether structured prompting strategies improve audio–semantic reasoning, we experimented with caption-before-reasoning. In the caption-before-reasoning setup, models were first prompted to generate an explicit audio caption before performing the downstream semantic reasoning task. Contrary to our expectations, this strategy did not yield consistent performance gains. Error

Dataset	ALM	Acc ↑	P ↑	R ↑	F1 ↑	Acc_C ↑	Acc_I ↑
Afri-200	AudioFlamingo2	0.6019	0.6493	0.6019	0.5677	0.3204	0.8835
	AudioFlamingo3	0.5000	0.2500	0.5000	0.3333	0.0000	1.0000
	GAMA	0.2573	0.7802	0.2573	0.2819	0.4951	0.0194
	Kimi	0.8301	0.8545	0.8301	0.8271	0.6990	0.9612
	Qwen2.5 Omni	0.7136	0.8179	0.7136	0.6880	1.0000	0.4272
	Qwen2 Audio 7B	0.8447	0.8452	0.8447	**0.8446**	0.8641	0.8252
	SALMONN	0.7379	0.7393	0.7379	0.7375	0.7767	0.6990
Afri-Gen	AudioFlamingo2	0.8793	0.9022	0.8793	0.8815	0.7759	0.9828
	AudioFlamingo3	0.5000	0.2500	0.5000	0.3333	0.0000	1.0000
	GAMA	0.2759	0.8684	0.2759	0.3562	0.4828	0.0690
	Kimi	0.8190	0.8418	0.8190	0.8159	0.6897	0.9483
	Qwen2.5 Omni	0.9138	0.9265	0.9138	**0.9131**	1.0000	0.8276
	Qwen2 Audio7B	0.8276	0.8625	0.8276	0.8233	0.6724	0.9828
	SALMONN	0.9052	0.9053	0.9052	0.9052	0.8966	0.9138
Medical	AudioFlamingo2	0.8125	0.8175	0.8125	**0.8118**	0.7500	0.8750
	AudioFlamingo3	0.5000	0.2500	0.5000	0.3333	0.0000	1.0000
	GAMA	0.3500	0.7429	0.3500	0.4210	0.5500	0.1500
	Kimi	0.7375	0.8279	0.7375	0.7181	0.4750	1.0000
	Qwen2.5Omni	0.7250	0.8226	0.7250	0.7025	1.0000	0.4500
	Qwen2 Audio	0.7625	0.7640	0.7625	0.7622	0.7250	0.8000
	SALMONN	0.7750	0.7757	0.7750	0.7749	0.8000	0.7500

Table 5: Zero-shot performance on Audio Consistency (Generative Models). ↑ indicates higher is better.

Dataset	ALM	Acc ↑	P ↑	R ↑	F1 ↑	E-Acc ↑	N-Acc ↑	C-Acc ↑
Afri-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
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

Table 6: Zero-shot performance on audio entailment (Contrastive Models). ↑ indicates higher is better.

analysis revealed two primary causes: (1) model refusals, where models declined to provide a response or reasoning (e.g., "I cannot provide that analysis"), and (2) missing labels, where models produced a caption but failed to output the required classification label, rendering the response unusable for evaluation.

9 Conclusion

We have introduced a comprehensive benchmark for evaluating semantic reasoning in audio language models across diverse domains, tasks, and accents. By combining large-scale hypothesis generation with careful human verification, our datasets isolate distinct reasoning phenomena including entailment, plausibility, consistency, accent-conditioned drift and semantic restraint that challenge current models beyond transcription quality. Our results show that next-token prediction models, particularly Qwen2-based variants and AudioFlamingo, substantially outperform contrastive baselines on most reasoning tasks, yet still exhibit significant over-entailment and susceptibility to commonsense bias. The benchmark also surfaces performance gaps under domain shift and accent variation, underscoring the need for robustness to linguistic diversity. We hope that these resources and analyses will catalyze the development of audio language models that ground their inferences in acoustic evidence and handle the richness of spoken language with care.

Limitations

This work has several limitations. First, although we curated diverse datasets across multiple African accents and domains, the benchmark still reflects a finite selection of languages, topics and speakers; performance may not generalize to other dialects, spontaneous discourse genres or low-resource languages outside our coverage. Second, while our human verification protocol mitigates hallucinated hypotheses, annotator judgments and edits remain subjective; some subtle distinctions could be missed or misclassified. Third, the reliance on

Dataset	ALM	Acc ↑	P ↑	R ↑	F1 ↑	Acc_P ↑	Acc_I ↑
Afri-200	LAION-CLAP	0.4976	0.4968	0.4970	0.4906	0.6117	0.3824
	MSCLAP_23	0.5122	0.5124	0.5124	**0.5116**	0.4757	0.5490
	MSCLAP_22	0.4976	0.4987	0.4997	0.3841	0.0680	0.9314
Afri-Gen	LAION-CLAP	0.5603	0.5626	0.5603	**0.5564**	0.6552	0.4655
	MSCLAP_23	0.5172	0.5176	0.5172	0.5149	0.5862	0.4483
	MSCLAP_22	0.5431	0.5550	0.5431	0.5169	0.3103	0.7759
Medical	LAION-CLAP	0.5875	0.6068	0.5875	**0.5680**	0.8000	0.3750
	MSCLAP_23	0.4000	0.3990	0.4000	0.3985	0.4500	0.3500
	MSCLAP_22	0.5375	0.5419	0.5375	0.5250	0.3750	0.7000

Table 7: Zero-shot performance on audio Plausibility (Contrastive Models). ↑ indicates higher is better.

Dataset	ALM	Acc ↑	P ↑	R ↑	F1 ↑	Acc_C ↑	Acc_I ↑
Afri-200	LAION-CLAP	0.5049	0.5050	0.5049	**0.5018**	0.4272	0.5825
	MSCLAP_23	0.4660	0.4467	0.4660	0.4128	0.7670	0.1650
	MSCLAP_22	0.5340	0.5514	0.5340	0.4908	0.2427	0.8252
Afri-Gen	LAION-CLAP	0.5603	0.5618	0.5603	**0.5577**	0.6379	0.4828
	MSCLAP_23	0.4483	0.4483	0.4483	0.4483	0.4483	0.4483
	MSCLAP_22	0.5000	0.5000	0.5000	0.3967	0.0862	0.9138
Medical	LAION-CLAP	0.5750	0.5758	0.5750	0.5739	0.5250	0.6250
	MSCLAP	0.6250	0.6374	0.6250	**0.6164**	0.7750	0.4750
	MSCLAP_22	0.5000	0.5000	0.5000	0.4972	0.5750	0.4250

Table 8: Zero-shot performance on Audio Consistency (Contrastive Models). ↑ indicates higher is better.

Model	Bias Rate ↓	Accept Rate ↑
AudioFlamingo2	96.75%	97.75%
AudioFlamingo3	81.50%	**98.25%**
GAMA	58.75%	68.25%
Kimi	55.25%	84.50%
Qwen2.5Omni	41.50%	95.50%
Qwen2AudioInstruct	33.25%	88.75%
SALMONN	1.50%	3.50%

Table 9: AfriNames accent-drift evaluation across models. ↓ indicates lower is better; ↑ indicates higher is better.

Model	HLU Rate ↓	SPRT Rate ↑	SRA ↑
AudioFlamingo3	88.00%	100.00%	66.00%
GAMA	96.83%	98.50%	**72.62%**
Kimi	17.17%	80.50%	12.88%
Qwen2.5Omni	2.33%	62.50%	1.75%
Qwen2AudioInstruct	39.17%	97.50%	29.38%
SALMONN	27.17%	35.50%	20.38%
LAION-CLAP	60.17%	68.00%	45.12%

Table 10: AfriNames restraint evaluation across models. The SRA (↑) metric balances the model's ability to identify supported content against its tendency toward hallucination (↓). HLU stands for Hallucination and SPRT stands for support

large language models for hypothesis generation and label mapping introduces their own biases and may not capture all error modes. Finally, our evaluation focuses on zero-shot inference; fine-tuning or prompting strategies may yield different patterns of success and failure. Future work should expand the dataset to additional languages and domains, incorporate multiple annotation phases for quality control, and explore training-time interventions to reduce over-entailment and accent sensitivity.

References

Henrique Lopes Cardoso, Rui Sousa-Silva, Maarit Koponen, and Antonio Pareja-Lora. Proceedings of 2nd LUHME Workshop.

Cheng-Han Chiang, Xiaofei Wang, Chung-Ching Lin, Kevin Lin, Linjie Li, Radu Kopetz, Yao Qian, Zhendong Wang, Zhengyuan Yang, Hung-yi Lee, and Lijuan Wang. 2025. Audio-Aware Large Language Models as Judges for Speaking Styles. *arXiv preprint*. ArXiv:2506.05984 [eess].

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-audio technical report. *Preprint*, arXiv:2407.10759.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-Audio: Advancing Universal Audio Understanding via Unified Large-Scale Audio-Language Models. *arXiv preprint*. ArXiv:2311.07919 [eess].

Soham Deshmukh, Benjamin Elizalde, Rita Singh, and Huaming Wang. 2023. Pengi: An Audio Language Model for Audio Tasks. *Advances in Neural Information Processing Systems*, 36:18090–18108.

Soham Deshmukh, Shuo Han, Hazim Bukhari, Benjamin Elizalde, Hannes Gamper, Rita Singh, and Bhiksha Raj. Audio Entailment: Assessing Deductive Reasoning for Audio Understanding.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. CLAP Learning Audio Concepts from Natural Language Supervision. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. ISSN: 2379-190X.

Sreyan Ghosh, Zhifeng Kong, Sonal Kumar, S. Sakshi, Jaehyeon Kim, Wei Ping, Rafael Valle, Dinesh Manocha, and Bryan Catanzaro. 2025. Audio Flamingo 2: An Audio-Language Model with Long-Audio Understanding and Expert Reasoning Abilities. *arXiv preprint*. ArXiv:2503.03983 [cs].

Sreyan Ghosh, Sonal Kumar, Ashish Seth, Chandra Kiran Reddy Evuru, Utkarsh Tyagi, S. Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. 2024. GAMA: A Large Audio-Language Model with Advanced Audio Understanding and Complex Reasoning Abilities. *arXiv preprint*. ArXiv:2406.11768 [cs].

Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and Bryan Catanzaro. 2025. Audio Flamingo 3: Advancing Audio Intelligence with Fully Open Large Audio Language Models. *arXiv preprint*. ArXiv:2507.08128 [cs].

KimiTeam, Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, Zhengtao Wang, Chu Wei, Yifei Xin, Xinran Xu, Jianwei Yu, Yutao Zhang, Xinyu Zhou, Y. Charles, and 21 others. 2025. Kimi-Audio Technical Report. *arXiv preprint*. ArXiv:2504.18425 [eess].

Marek Kubis, Paweł Skórzewski, Iwona Christop, Mateusz Czyżnikiewicz, Jakub Kubiak, Łukasz Bondaruk, and Marcin Lewandowski. 2025. Preservation of Language Understanding Capabilities in Speech-aware Large Language Models. *arXiv preprint*. ArXiv:2509.12171 [cs].

Odunayo Ogundepo, Tajuddeen R. Gwadabe, Clara E. Rivera, Jonathan H. Clark, Sebastian Ruder, David Ifeoluwa Adelani, Bonaventure F. P. Dossou, Abdou Aziz DIOP, Claytone Sikasote, Gilles Hacheme, Happy Buzaaba, Ignatius Ezeani, Rooweither Mabuya, Salomey Osei, Chris Emezue, Albert Njoroge Kahira, Shamsuddeen H. Muhammad, Akintunde Oladipo, Abraham Toluwase Owodunni, and 33 others. 2023. AfriQA: Cross-lingual Open-Retrieval Question Answering for African Languages. *arXiv preprint*. ArXiv:2305.06897 [cs].

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

Jing Peng, Yucheng Wang, Yu Xi, Xu Li, Xizhuo Zhang, and Kai Yu. 2025. A Survey on Speech Large Language Models. *arXiv preprint*. ArXiv:2410.18908 [eess] version: 3.

S. Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. 2024. MMAU: A Massive Multi-Task Audio Understanding and Reasoning Benchmark. *arXiv preprint*. ArXiv:2410.19168 [eess].

Mardhiyah Sanni, Tassallah Abdullahi, Devendra D. Kayande, Emmanuel Ayodele, Naome A. Etori, Michael S. Mollel, Moshood Yekini, Chibuzor Okocha, Lukman E. Ismaila, Folafunmi Omofoye, Boluwatife A. Adewale, and Tobi Olatunji. 2025. Afrispeech-Dialog: A Benchmark Dataset for Spontaneous English Conversations in Healthcare and Beyond. *arXiv preprint*. ArXiv:2502.03945 [cs].

Ayushman Sarkar, Mohd Yamani Idna Idris, and Zhenyu Yu. 2025. Reasoning in Computer Vision: Taxonomy, Models, Tasks, and Methodologies. *arXiv preprint*. ArXiv:2508.10523 [cs].

Jiatong Shi, Shih-Heng Wang, William Chen, Martijn Bartelds, Vanya Bannihatti Kumar, Jinchuan Tian, Xuankai Chang, Dan Jurafsky, Karen Livescu, Hung-yi Lee, and Shinji Watanabe. 2024. ML-SUPERB 2.0: Benchmarking Multilingual Speech Models Across Modeling Constraints, Languages, and Datasets. *arXiv preprint*. ArXiv:2406.08641 [cs].

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2023. Salmonn: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2310.13289*.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F Chen. 2024. Audiobench: A universal benchmark for audio large language models. *arXiv preprint arXiv:2406.16020*.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F. Chen. 2025. AudioBench: A Universal Benchmark for Audio Large Language Models. *arXiv preprint*. ArXiv:2406.16020 [cs].

Wanqi Yang, Yanda Li, Yunchao Wei, Meng Fang, and Ling Chen. 2025. SpeechR: A Benchmark for Speech Reasoning in Large Audio-Language Models. *arXiv preprint*. ArXiv:2508.02018 [cs] version: 1.

A Dataset Details

We utilize a diverse collection of Pan-African speech and text datasets to evaluate semantic reasoning across clinical and general domains. Table 11 provides a consolidated overview of the dataset statistics and accessibility.

Dataset	Primary Domain
AfriSpeech-200[5]	Clinical/General ASR
AfriSpeech-Dialog[6]	Conversational ASR/SLU
Afri-Names[7]	Named Entity/Commands
Afrispeech-Medical[8]	Clinical Dialogues

Table 11: Summary of datasets. These resources support evaluation across entailment, consistency, plausibility, and accent-conditioned reasoning tasks.

AfriSpeech-200 This corpus contains 67,577 audio–transcript pairs from 2,463 unique speakers across 13 African countries. It serves as the benchmark for reasoning tasks that require high-fidelity transcription of diverse African accents in both medical and everyday contexts.

AfriSpeech-Dialog This dataset consists of simulated spontaneous conversations. Unlike read-speech corpora, it captures the nuances of natural disfluencies and conversational flow, specifically curated to test the robustness of spoken language understanding (SLU) models.

Afri-Names A specialized read-speech dataset containing 6,307 samples. It focuses on high-precision segments—specifically numbers, African named entities, and voice commands—spanning 12 accents across 4 countries. It is used primarily to evaluate accent-conditioned reasoning and NER.

Med-Convo-Nig Sourced from real-world Nigerian doctor–patient tele-consultations, this dataset provides high-stakes clinical dialogues. It is critical for evaluating model performance on local linguistic variations in healthcare-specific semantic tasks, such as clinical consistency and entailment.

[5] https://huggingface.co/datasets/intronhealth/afrispeech-200

[6] https://huggingface.co/datasets/intronhealth/afrispeech-dialog

[7] https://huggingface.co/datasets/intronhealth/afri-names

[8] https://huggingface.co/datasets/intronhealth/med-convo-nig

B Human Verification Protocol

All candidate hypotheses produced by the language model were audited by a team of three trained annotators with backgrounds in linguistics and speech technology. The annotators followed a detailed guideline document that emphasised:

- Listening to the entire audio premise, including pauses and prosodic cues, before reading the hypothesis.
- Classifying each hypothesis as supported, contradicted or unsupported based on the audio alone, without using external world knowledge.
- Editing hypotheses that contained hallucinated details, ambiguous wording, or reliance on unstated background information. Edits were constrained to preserve the intended semantic relationship and remain faithful to the original audio.
- Flagging any problematic audio segments (e.g., very low signal-to-noise ratio) for exclusion from the benchmark.

Each hypothesis was seen by two annotators; disagreements were resolved through discussion with a third annotator acting as arbiter. Random spot checks were performed by one of the authors to ensure consistency across the entire corpus. This human verification step proved especially valuable for accented speech and unfamiliar proper names, where automatic generation alone often led to subtle semantic drift.

C Evaluation Metrics

We evaluate audio–semantic reasoning performance using a combination of classification, calibration, and robustness-oriented metrics, following prior work on audio entailment and audio–language model evaluation. Metrics are computed at the hypothesis level and aggregated per model and dataset.

C.1 Primary Classification Metrics

For multi-class and binary decision tasks, we report standard classification metrics derived from the confusion matrix.

Accuracy (ACC). Accuracy measures the proportion of correctly classified instances:

$$\mathrm{ACC} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively.

Accuracy is reported for all tasks but may obscure class imbalance effects, especially in entailment-style settings.

Precision (P). Precision measures the proportion of predicted positives that are correct:

$$\mathrm{P} = \frac{TP}{TP + FP} \tag{2}$$

Recall (R). Recall measures the proportion of true positives that are correctly identified:

$$\mathrm{R} = \frac{TP}{TP + FN} \tag{3}$$

F1 Score (F1). The F1 score is the harmonic mean of precision and recall:

$$\mathrm{F1} = \frac{2 \cdot \mathrm{P} \cdot \mathrm{R}}{\mathrm{P} + \mathrm{R}} \tag{4}$$

For multi-class tasks (e.g., entailment, neutral, contradiction), we report macro-averaged precision, recall, and F1 unless otherwise stated.

C.2 Entailment-Aware Accuracy Metrics

Following prior audio entailment benchmarks, we additionally report label-specific accuracies to capture asymmetric error behavior.

Entailment Accuracy (EACC).

$$\mathrm{EACC} = \frac{\text{\# correctly predicted entailment instances}}{\text{\# entailment instances}} \tag{5}$$

Neutral Accuracy (NACC).

$$\mathrm{NACC} = \frac{\text{\# correctly predicted neutral instances}}{\text{\# neutral instances}} \tag{6}$$

Contradiction Accuracy (CACC).

$$\mathrm{CACC} = \frac{\text{\# correctly predicted contradiction instances}}{\text{\# contradiction instances}} \tag{7}$$

These metrics are especially informative for diagnosing model bias toward entailment or over-rejection behavior.

C.3 Task Applicability of Metrics

The applicability of each metric varies by task:

- **Audio Entailment.** We report ACC, P, R, F1, along with EACC, NACC, and CACC.
- **Audio–Text Semantic Consistency (Binary).** We report ACC, P, R, and F1. Label-specific accuracies reduce to positive and negative class accuracy.
- **Semantic Plausibility Judgment.** We report ACC and F1. Precision and recall are emphasized for the *implausible* class to measure hallucination susceptibility.

C.3.1 Robustness and Over-Inference Metrics (AfriNames)

For AfriNames-style datasets, which lack rich semantic content, standard NLI metrics are insufficient. Instead, we focus on restraint and robustness.

Semantic Restraint Accuracy (SRA). Semantic restraint accuracy measures the proportion of cases in which the model correctly refrains from asserting unsupported semantic content:

$$\mathrm{SRA} = \frac{\text{\# correct neutral or abstain predictions}}{\text{\# total AfriNames instances}} \tag{8}$$

C.3.2 Interpretation Across Models

Together, these metrics allow us to characterize:

- Logical inference capability (NLI, consistency),
- Pragmatic and intent-level reasoning,
- Commonsense integration from speech,
- Hallucination propensity and over-inference,
- Robustness to accent and low-semantic-content inputs.

This multi-metric evaluation provides a comprehensive and task-aligned assessment of audio–language models across diverse spoken domains.

C.4 Handling Multiple Hypotheses per Label

For several datasets and tasks, including Audio–Text Semantic Consistency and Spoken Natural Language Inference, we generate multiple hypotheses per semantic label (e.g., two consistent and two inconsistent hypotheses per audio instance). This

design choice reflects the inherent variability in how a single semantic relation can be linguistically expressed, while maintaining a fixed underlying label.

Evaluation Strategy. We adopt a unified evaluation strategy in which all hypotheses sharing the same semantic label are treated equivalently. Specifically, each hypothesis–audio pair is evaluated independently by the model, and the resulting predictions are aggregated at the label level. For binary tasks such as semantic consistency, hypotheses labeled as *consistent* are treated as positive instances, while those labeled as *inconsistent* are treated as negative instances. For ternary tasks such as entailment, hypotheses are grouped into *entails*, *neutral*, and *contradicts* categories.

This approach allows us to preserve hypothesis-level diversity without requiring separate inference runs or task-specific evaluation pipelines.

Implications for Metrics and Tables. All reported metrics (Accuracy, Precision, Recall, F1, and class-conditional accuracies) are computed over the full set of hypothesis–audio pairs. As a result, performance tables (e.g., Table 4) summarize model behavior at the *task level*, rather than at the individual hypothesis variant level. Each table corresponds to a single task (e.g., entailment or consistency), aggregating over all hypothesis realizations within that task.

This design ensures that evaluation remains comparable across datasets and models, while avoiding fragmentation into multiple near-duplicate tables.

Duplicate Hypothesis and Correlation Considerations. Although multiple hypotheses are generated from the same transcript, they are not duplicates in the lexical sense. Each hypothesis differs in phrasing, abstraction level, or semantic emphasis, even when sharing the same label. Consequently, models must generalize across paraphrastic variation rather than exploit surface-level repetition.

To mitigate concerns about correlation between hypotheses derived from the same audio, we emphasize that: (i) hypotheses are evaluated independently, (ii) metrics are reported in aggregate rather than per-audio averages, and (iii) our primary goal is to assess semantic robustness under linguistic variability rather than estimate per-instance uncertainty.

This protocol follows established practices in natural language inference and paraphrase-based evaluation, and is particularly appropriate for probing robustness in audio–semantic reasoning under accent, domain, and acoustic variation.

Summary. Treating multiple hypotheses per label as equivalent instances enables efficient inference, stable metric estimation, and controlled linguistic diversity, while preserving the integrity and interpretability of task-level evaluation.

D Model Architectures Table

E Different LLMs and their Generated Hypotheses

To illustrate qualitative differences among next-token prediction models, we include a set of manually curated examples in Table 13. Each block shows the original transcript followed by hypotheses generated by Llama 3.1 8B, Mistral Large 3 and Qwen 2 7B, annotated according to the corresponding semantic reasoning task. These examples highlight how different model architectures navigate plausibility, entailment, consistency, accent-induced drift and restraint.

Model	Model Size	LLM 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

Table 12: Architectural specifications of evaluated Audio Language Models.

This table specifies the underlying LLM decoder and the total parameter count (where known) for the 10 models.

F Hyperparameters

This table captures your specific sampling settings and the retry logic. Note that for CLAP-based models, generation parameters are not applicable because they are similarity-based. All Experiments were ran on a single Nvidia Blackwell B200 and 80GB of RAM

G Prompts

In section we show all the prompts used by the Large Language Model and prompts used by the Audio Language model.

H CLAP Architecture and Results Format

In this section, we describe the Clap Evaluation framework shown in Figure 3.

I Prompt Variants (v)

(v1).

```
Audio statement verification task.
Does the following statement match what is said in the audio?

Statement: {hypothesis}

Answer with exactly one word - either CONSISTENT or INCONSISTENT:
```

(v2).

```
Listen carefully to the audio recording. Then read the statement below.
Your task: Decide if the statement MATCHES or CONTRADICTS what is said in the audio.
- If the statement accurately reflects what is said in the audio, respond: CONSISTENT
- If the statement contradicts or misrepresents what is said in the audio, respond: INCONSISTENT
Think about what you heard in the audio before answering.

STATEMENT: {hypothesis}
Your answer (CONSISTENT or INCONSISTENT):
```

(v3).

```
You are given an audio recording and a text statement.
Determine whether the text is consistent with the meaning conveyed in the audio.
Respond with one of the following labels only:
CONSISTENT
INCONSISTENT

STATEMENT:
{hypothesis}
```

J Prompt Selection

Table 17 compares three prompt variants for the semantic consistency task. We report overall accuracy, precision, recall, and macro F1, along with class-specific accuracies for the CONSISTENT and INCONSISTENT labels.

Semantic Restraint / Over-Inference Prompt

Purpose: Generate hypotheses to test whether models correctly withhold inference when audio content is minimal.
Methodological Value: Measures hallucination/over-inference rates and provides a clean signal for semantic restraint.

System Prompt: You are a careful evaluator of spoken language meaning. You are given a verbatim transcript of a very short spoken audio recording. This audio contains LIMITED semantic content. Your task is to generate hypotheses that test whether a model can correctly WITHHOLD inference.
Task: Generate the following:
- **SUPPORTED hypotheses (2x):** Clearly supported by the audio; minimal and cautious.
- **UNSUPPORTED hypotheses (2x):** Likely to tempt a model to over-infer; not supported by this specific audio.

Constraints:
- Do not quote/paraphrase; no meta-language ("the speaker says").
- Do not add roles, topics, intent, or domains unless explicitly supported.
- **Prohibited:** Explicit negation terms ("no", "not", "never", "without").

Output Format (JSON): `{ "supported": ["...", "..."], "unsupported": ["...", "..."] }`

Accent-Conditioned Semantic Drift Prompt

Purpose: Test if models infer different meanings from similar utterances spoken with different accents.
Methodological Value: Measures fairness gaps and semantic drift triggered by non-semantic audio features.

System Prompt: You are evaluating semantic stability. This task tests whether models infer DIFFERENT meanings from similar short utterances spoken with different accents or pronunciations.
Task:
- **ACCENT-INVARIANT (4x):** Truth value depends only on semantics, not speaker identity.
- **ACCENT-SENSITIVE LURES (4x):** Incorrect inferences likely caused by accent or training bias.

Constraints:
- Do not assume nationality, ethnicity, gender, or location.
- Do not use explicit negation terms; no quoting the transcript.

Output Format (JSON): `{ "accent_invariant": [...], "accent_sensitive_lures": [...] }`

Audio Entailment

System Prompt: You are a helpful assistant with expert knowledge in spoken language understanding and parliamentary discourse. Generate hypotheses for a Spoken Natural Language Inference (NLI) task based on the provided institutional transcript.
Instructions: Generate three hypotheses for each:
- **Entailment:** Definitely true given the meaning of the spoken audio.
- **Neutral:** Might be true but cannot be determined from audio alone.
- **Contradiction:** Definitely false given the meaning of the audio.

Constraints:
- Reflect semantic inference, not surface wording; do not quote the transcript.
- **Prohibited:** Explicit negation terms ("no", "not", "never").

Output Format (JSON): `{ "entailment": [...], "neutral": [...], "contradiction": [...] }`

Audio–Text Semantic Consistency Prompt

System Prompt: You are assisting in the construction of an audio–text semantic consistency dataset using parliamentary audio transcripts.
Task: Generate four text statements:
- **Consistent (2x):** Semantically aligned with the audio.
- **Inconsistent (2x):** Semantically conflicting with the audio.

Constraints:
- Concern institutional roles, procedures, or formal discourse.
- Do not quote or restate the transcript; do not use explicit negation words.

Output Format (JSON): `{ "consistent": ["...", "..."], "inconsistent": ["...", "..."] }`

Dataset / Task	Model	Transcript and Generated Hypotheses
AfriNames — Accent Drift (Over-Inference)		
Transcript: "In the environmental assessment conducted over 34 months, researchers identified 24 endangered species and proposed a . . . "		
	Llama 3.1 8B	**Accent-Invariant:** The assessment identified multiple endangered species requiring protection.
		Accent-Sensitive Lure: The researchers are likely affiliated with a prestigious Western university.
	Mistral Large 3	**Accent-Invariant:** The study revealed significant threats to biodiversity.
		Accent-Sensitive Lure: The findings were shaped by international conservation agendas.
	Qwen 2 7B	**Accent-Invariant:** Several species were found to be at risk of extinction.
		Accent-Sensitive Lure: The research was conducted in an urbanized region.
AfriNames — Accent Restraint (Semantic Minimality)		
Transcript: "The 1921 census projects a population surge to 488 million, indicating a 4% growth rate influenced by immigration."		
	Llama 3.1 8B	**Supported:** The population is projected to increase.
		Unsupported: The census results will directly influence national economic policy.
	Mistral Large 3	**Supported:** Immigration contributes to population growth.
		Unsupported: The census was designed to justify housing reforms.
	Qwen 2 7B	**Supported:** A population surge is anticipated.
		Unsupported: The growth rate is unusually high for a developed country.
AfriSpeech-200 — Semantic Consistency		
Transcript: "This study builds on her previous work examining what qualities and skills make for stronger middle-grade science teachers."		
	Llama 3.1 8B	**Consistent:** The study focuses on identifying characteristics of effective science teachers.
		Inconsistent: The study explores the impact of dance training on teachers.
	Mistral Large 3	**Consistent:** The research examines skills linked to strong science instruction.
		Inconsistent: The work analyzes athletic performance in schools.
	Qwen 2 7B	**Consistent:** The paper investigates qualities that improve science teaching.
		Inconsistent: The research evaluates music education programs.
AfriSpeech-200 — Audio Entailment		
Transcript: "Snow arriving from the north later tomorrow, moving south through Thursday morning may lead to some travel disruption."		
	Llama 3.1 8B	**Entailment:** Weather conditions may disrupt travel over the next two days.
		Contradiction: Travel will not be affected by weather conditions.
	Mistral Large 3	**Entailment:** Snowfall could interfere with transportation schedules.
		Contradiction: Clear weather will ensure uninterrupted travel.
	Qwen 2 7B	**Entailment:** Adverse weather may impact travel plans.
		Contradiction: No travel disruptions are expected.
AfriSpeech-200 — Semantic Plausibility		
Transcript: "Not everyone can run a business effectively, so it is important that younger shoemakers are trained to be recruited by big companies."		
	Llama 3.1 8B	**Plausible:** Training increases employment opportunities for younger shoemakers.
		Implausible: Large companies avoid hiring trained shoemakers.
	Mistral Large 3	**Plausible:** Professional training prepares shoemakers for industry roles.
		Implausible: Shoemaking skills are irrelevant to corporate recruitment.
	Qwen 2 7B	**Plausible:** Skill development improves recruitment prospects.
		Implausible: Shoemaking is no longer valued by employers.

Table 13: Qualitative examples illustrating how different large language models generate task-specific hypotheses from the same spoken input. Each block presents the original transcript followed by model-generated hypotheses corresponding to the semantic reasoning task.

Model	Max New Tokens	Temp	Top-K	Top-P	Audio Temp/K
Qwen2.5-Omni	512	0.0	5	-	- / -
Qwen2-Audio	512	0.0	5	-	- / -
SALMONN	512	-	-	-	- / -
Kimi-Audio	512	0.0	5	-	0.8 / 10
GAMA	400	0.1	-	0.95	- / -
AudioFlamingo2	512	0.0	5	-	- / -
AudioFlamingo3	512	0.0	5	-	- / -
LAION-CLAP	-	-	-	-	- / -
MSCLAP 2023	-	-	-	-	- / -
MSCLAP 2022	-	-	-	-	- / -

Table 14: Inference hyperparameters. Generative models employ a greedy retry strategy (top_k=1) if initial sampling fails. CLAP models use similarity-based retrieval.

Prompt 1: Semantic Restraint / Over-Inference Generation

```
You are a careful evaluator of spoken language meaning.
You are given a verbatim transcript of a very short spoken audio recording.
This audio contains LIMITED semantic content. Your task is to generate hypotheses that test
whether a model can correctly WITHHOLD inference when the audio does not provide enough evidence.
```

TASK: Generate the following:

1. **SUPPORTED hypotheses (2 items)** – Statements clearly supported; minimal and cautious.
2. **UNSUPPORTED hypotheses (6 items)** – Statements that sound reasonable but are NOT supported; designed to tempt over-inference.

CONSTRAINTS:

- Do NOT quote/paraphrase transcript or describe what the speaker "says."
- Do NOT add roles, topics, locations, intent, or domains unless explicitly supported.
- Avoid explicit negation terms ("no," "not," "never").

GOAL: Unsupported hypotheses should test for: Invented topics, intent, social roles, domains, and situational context.

OUTPUT FORMAT (JSON):

```
{ "supported": ["...", "..."], "unsupported": ["...", "...", "...", "...", "...", "..."] }
```

TRANSCRIPT: {{TRANSCRIPT}}

Table 15: Prompt used to generate diagnostic items for evaluating model hallucination and over-inference.

Prompt 2: Accent-Conditioned Semantic Drift Generation

```
You are evaluating semantic stability in spoken language understanding.
IMPORTANT: This task tests whether models infer DIFFERENT meanings from similar short utterances
spoken with different accents or pronunciations.
```

TASK: Generate the following:

1. **ACCENT-INVARIANT hypotheses (4 items)** – Truth should NOT depend on accent or speaker identity; rely only on semantic content.
2. **ACCENT-SENSITIVE LURE hypotheses (4 items)** – Statements models might incorrectly infer due to accent, pronunciation, or training bias.

CONSTRAINTS:

- Do NOT assume speaker nationality, ethnicity, gender, age, or location.
- Do NOT quote/paraphrase. Each hypothesis must be one complete sentence.
- Avoid explicit negation terms.

GOAL: Lures should test for: Assigned social roles, geographic/cultural backgrounds, and injected domain meanings.

OUTPUT FORMAT (JSON):

```
{ "accent_invariant": [...], "accent_sensitive_lures": [...] }
```

TRANSCRIPT: {{TRANSCRIPT}}

Table 16: Prompt used to generate diagnostic items for evaluating fairness and stability under accent variation.

Audio Semantic Plausibility Prompt

System Prompt: You are given a transcript of a spoken medical audio recording.
Task: Generate four statements:
- **Plausible (2x):** Align with the medical context and clinical commonsense.
- **Implausible (2x):** Unlikely given the audio context and medical norms.

Constraints:
- Plausibility judgments may rely on medical commonsense.
- Do not quote the transcript; do not use explicit negation.

Output Format (JSON): `{ "plausible": ["...", "..."], "implausible": ["...", "..."] }`

ALM Prompt 1: Audio Entailment (Zero-Shot Classification)

Context: Evaluation of logical inference between a raw audio signal and a text hypothesis.

System Prompt: You are an Audio Reasoning Agent. You are given an audio recording and a hypothesis. Determine the logical relationship between them.
Task: Determine whether the hypothesis is:
- **Entailed:** The hypothesis is a direct logical consequence of the audio.
- **Contradicted:** The hypothesis is logically impossible given the audio.
- **Neutral:** The hypothesis is plausible but neither supported nor refuted by the audio.

Constraint: Respond with exactly one label from the following set: {`entailment`, `contradiction`, `neutral`}.

Input Format: `Audio: <audio_signal> | Hypothesis: {{HYPOTHESIS}}`

ALM Prompt 2: Semantic Consistency (Binary Alignment)

Context: Determining the semantic alignment between auditory evidence and a descriptive statement.

System Prompt: Given an audio recording and a statement, determine whether the statement is semantically consistent with the audio.
Constraint: Respond with exactly one label: {`consistent`, `inconsistent`}. Do not provide explanations.

Input Format: `Audio: <audio_signal> | Statement: {{STATEMENT}}`

ALM Prompt 3: Semantic Plausibility (Contextual Reasoning)

Context: Assessing if a statement is likely/reasonable within the situational context of the audio.

System Prompt: Given an audio recording and a statement, determine whether the statement is plausible given the audio context and general world knowledge.
Constraint: Respond with exactly one label: {`plausible`, `implausible`}.

Input Format: `Audio: <audio_signal> | Statement: {{STATEMENT}}`

Dataset	Prompt Variant	Acc.	Prec.	Rec.	F1	Acc. (Con.)	Acc. (Inc.)
AfriSpeech-200	AudioFlamingo3 (v1)	0.50	0.13	0.25	0.17	0.00	1.00
AfriSpeech-200	AudioFlamingo3 (v2)	0.53	0.52	0.36	**0.28**	0.07	1.00
AfriSpeech-200	AudioFlamingo3 (v3)	0.53	0.31	0.21	0.16	0.06	1.00
AfriSpeech-General	AudioFlamingo3 (v1)	0.50	0.17	0.33	0.22	0.00	1.00
AfriSpeech-General	AudioFlamingo3 (v2)	0.54	0.24	0.16	0.14	0.09	1.00
AfriSpeech-General	AudioFlamingo3 (v3)	0.76	0.42	0.38	**0.37**	0.52	1.00
Medical	AudioFlamingo3 (v1)	0.50	0.17	0.33	0.22	0.00	1.00
Medical	AudioFlamingo3 (v2)	0.53	0.31	0.21	0.16	0.05	1.00
Medical	AudioFlamingo3 (v3)	0.64	0.79	0.64	**0.58**	0.28	1.00

Table 17: Prompt ablation for AudioFlamingo3 on the semantic consistency task. Best macro F1 per dataset is bolded.

CLAP Evaluation Framework

Methodology: Unlike generative ALMs, CLAP evaluates semantic reasoning through similarity-based alignment between audio embeddings and templated text hypotheses.

Templated Hypothesis Generation: For each task, CLAP generates task-specific templates to transform labels into natural language assertions:
- **Consistency:** *"Given the audio, the following statement is consistent: {hypothesis}"*
- **Entailment:** *"Given the audio, the following statement is true: {hypothesis}"*
- **Plausibility:** *"Given the audio, the following statement is plausible: {hypothesis}"*

Inference Mechanism: The model computes cosine similarity scores $S(a, t_i)$ between the audio embedding (a) and each templated hypothesis embedding (t_i). The final prediction is determined by:

$$\text{Label} = \arg\max_i S(a, t_i)$$

Data Schema (Output JSON):
- `scores`: Vector of similarity scores for each label option (e.g., *Consistent* vs. *Inconsistent*).
- `best_match_idx`: Index of the hypothesis template with the highest similarity score.
- `best_match_hypothesis`: The full raw text of the highest-scoring templated statement.
- `best_match_label`: The categorical label extracted from the winning template.
- `pred`: The binary/multiclass prediction indicator for the sample.

Figure 3: The CLAP evaluation pipeline. This framework adapts contrastive audio-text models for reasoning tasks by wrapping hypotheses in semantic templates and selecting labels based on embedding similarity.

ALM	Acc ↑	P	R	F1	E-Acc	N-Acc	C-Acc
AudioFlamingo2	0.3109	0.1036	0.3333	0.1581	0.0000	1.0000	0.0000
AudioFlamingo3	0.4663	0.5368	0.4802	0.4872	0.2740	0.4333	0.7333
Kimi	0.3057	0.6168	0.3238	0.2490	0.0548	0.8167	0.1000
Qwen2.5-Omni	0.4767	0.6568	0.4795	0.4673	0.4384	0.7667	0.2333
Qwen2-Audio	0.3523	0.5423	0.3580	0.3036	0.2740	0.7500	0.0500

Table 18: Zero-shot performance on Afrispeech Medical Entailment. All metrics are macro-averaged where applicable. ↑ indicates higher is better. The models struggle to caption and also reason all in the same prompt.

Prompt: Spoken Natural Language Inference

You are evaluating Spoken Natural Language Inference (Spoken NLI).

You are given: (1) an audio recording (premise), (2) a text hypothesis.

Step 1 — AUDIO CAPTION (evidence only):
Write a short caption of what is explicitly said in the audio.
- Use 1–2 sentences.
- Do NOT infer unstated details or add background knowledge.
- If audio is unclear, say "unclear" rather than guessing.

Step 2 — REASONING (use caption as evidence):
Using ONLY the caption from Step 1 as evidence, determine whether the hypothesis is: ENTAILMENT, CONTRADICTION, or NEUTRAL.

Output format:
CAPTION: <your caption>
LABEL: <ENTAILMENT|CONTRADICTION|NEUTRAL>

Hypothesis: "<HYPOTHESIS_TEXT>"