HalluAudio: A Comprehensive Benchmark for Hallucination Detection in Large Audio-Language Models

Feiyu Zhao[1]*, Yiming Chen[2]*, Wenhuan Lu[1], Daipeng Zhang[1], Xianghu Yue[1]†, Jianguo Wei[1]

[1]College of Intelligence and Computing, Tianjin University, China
[2]ASUS Intelligent Cloud Services, Singapore
[1]{zhaofeiyu, wenhuan, zhangdaipeng, yuexianghu, jianguo}@tju.edu.cn
[2]MattYM_Chen@asus.com

Abstract

Large Audio-Language Models (LALMs) have recently achieved strong performance across various audio-centric tasks. However, hallucination, where models generate responses that are semantically incorrect or acoustically unsupported, remains largely underexplored in the audio domain. Existing hallucination benchmarks mainly focus on text or vision, while the few audio-oriented studies are limited in scale, modality coverage, and diagnostic depth. We therefore introduce *HalluAudio*, the first large-scale benchmark for evaluating hallucinations across speech, environmental sound, and music. *HalluAudio* comprises over 5K human-verified QA pairs and spans diverse task types, including binary judgments, multi-choice reasoning, attribute verification, and open-ended QA. To systematically induce hallucinations, we design adversarial prompts and mixed-audio conditions. Beyond accuracy, our evaluation protocol measures hallucination rate, yes/no bias, error-type analysis, and refusal rate, enabling a fine-grained analysis of LALM failure modes. We benchmark a broad range of open-source and proprietary models, providing the first large-scale comparison across speech, sound, and music. Our results reveal significant deficiencies in acoustic grounding, temporal reasoning, and music attribute understanding, underscoring the need for reliable and robust LALMs.[1]

1 Introduction

Large Language Models (LLMs) have driven rapid progress in natural language processing, achieving strong performance across tasks such as reasoning, question answering, summarization, and multimodal understanding. Building on these advances, Large Audio-Language Models (LALMs) have emerged as a natural extension of LLMs to the audio domain. By leveraging large-scale corpora of speech, environmental sounds, and music, LALMs demonstrate impressive capabilities in speech recognition (Chu et al., 2023; Fang et al., 2025), sound question answering (Xu et al., 2025; Goel et al., 2025), and music understanding (Ghosh et al., 2025; Xiaomi, 2025). As LALMs are increasingly deployed in real-world applications, their accuracy and reliability have become critical. In particular, hallucinations remain a major concern, where models generate responses that are semantically incorrect or unsupported by the underlying audio.

A substantial body of work has demonstrated that generative models often produce outputs that are fluent yet factually unsupported in both text-only (Li et al., 2023; Lin et al., 2022) and vision–language settings (Guan et al., 2024; Wu et al., 2024). In contrast, hallucination in audio-centric models remains largely underexplored. Most existing benchmarks focus on text or vision, while the few emerging studies in the audio domain are limited in scale, modality coverage, and task diversity (Cheng et al., 2025). Current evaluations typically rely on small binary classification tasks and rarely probe critical failure modes such as response bias, refusal behavior, or multi-turn inconsistency. As a result, the field of LALMs still lacks a dedicated large-scale benchmark for systematically characterizing hallucination across speech, environmental sound, and music tasks.

To bridge this gap, we introduce *HalluAudio*, the first large-scale human-verified benchmark suite specifically designed to evaluate hallucination in LALMs. *HalluAudio* spans three major audio domains, speech, environmental sounds, and music, and supports diverse task formats, including classification, question answering, and open-ended generation. To reliably elicit and measure hallucinations, we incorporate adversarial prompts, mixed audio conditions. To ensure the reliability of our

*Equal contribution
†Corresponding author
[1]https://github.com/Feiyuzhao25/halluaudio

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 38797–38816
July 2-7, 2026

evaluation suite, we also perform manual inspection on the dataset. We curate *HalluAudio* through a five-step pipeline including data collection, task construction, adversarial augmentation, automated filtering, and multi-round human verification, resulting in a large-scale benchmark spanning three audio domains, dozens of task types, and over 5K carefully validated QA pairs.

We evaluate a range of state-of-the-art open source and proprietary LALMs on HalluAudio. Comprehensive experiments show that hallucination remains a systematic issue in current LALMs, even for tasks with a clear audio answer. We observe consistent Yes/No biases, non-trivial false refusal behaviors, and domain-dependent failure patterns across speech, sound, and music. More importantly, strong performance on standard audio benchmarks does not necessarily imply robustness against hallucination, highlighting a gap between capability evaluation and reliability assessment. These findings underscore the need for targeted hallucination diagnostics and validate *HalluAudio* as an effective tool for analyzing fine-grained failure modes beyond aggregate accuracy.

Our main contributions are as follows:

- **A large-scale human-verified benchmark for audio hallucination.** We introduce *HalluAudio*, the first large-scale human-verified audio hallucination benchmark spanning speech, environmental sounds, and music, with thousands of QA pairs per domain covering both audio understanding and audio-grounded generation scenarios.
- **A diverse suite of hallucination-inducing task designs.** *HalluAudio* includes binary and multi-class classification, audio QA, attribute verification, comparative reasoning, and open-ended generation. We construct adversarial prompts, mixed-audio inputs, and positives/negatives to systematically trigger and measure hallucinations.
- **A multi-dimensional empirical analysis of hallucination in LALMs.** We evaluate a broad set of open-source and proprietary LALMs and report detailed hallucination patterns across tasks and modalities. Our analysis incorporates accuracy, hallucination rate, Yes/No bias, error-type breakdown, and refusal rate, revealing critical reliability gaps in current LALMs.

2 Related Works

Large-Audio Language Models Inspired by the success of LLMs, recent work integrates audio representations with LLMs, giving rise to LALMs that process speech, environmental sounds, and music to generate textual responses and reasoning outputs. Early efforts such as AudioLM (Borsos et al., 2023) demonstrated that discretized audio tokens can be effectively modeled with language modeling techniques, enabling coherent continuation of speech and music. Subsequent models, including SALMONN (Tang et al., 2024), Qwen2-Audio (Chu et al., 2024), and Audio Flamingo (Kong et al., 2024), further align audio encoders with LLMs to support tasks such as ASR, audio question answering, captioning, and music understanding. More recent studies extend LALMs to finer-grained music understanding (Huang et al., 2022) and improved instruction following (Frieske and Shi, 2024). Overall, LALMs are rapidly evolving from transcription-focused systems to general audio reasoning models. Although several benchmarks have been proposed to evaluate various capabilities of LALMs (Yang et al., 2024; Chen et al., 2024, 2026b,a), their reliability, particularly with respect to hallucination, remains underexplored.

Hallucinations in Audio Tasks Hallucination has been extensively studied in text (Bang et al., 2025) and vision (Cao et al., 2024) domains, revealing systematic failures in grounding, object hallucination, and cross-modal consistency. In contrast, hallucinations in the audio domain remain largely underexplored. Frieske and Shi (2024) provides an early analysis of hallucination in ASR, showing that metrics such as WER fail to detect fluent but semantically irrelevant outputs and exposing vulnerabilities to misleading acoustic cues. AHa-Bench (Cheng et al., 2025) extends this line of inquiry to LALMs through a small-scale binary QA benchmark. While informative, existing benchmarks are limited in dataset scale, task diversity, and diagnostic depth, leaving critical failure modes such as response bias and refusal behavior. The field still lacks a unified taxonomy, controlled contrastive audio pairs, multi-format evaluation tasks, and large-scale human-annotated assessments. HalluAudio addresses these gaps by introducing the first large-scale, multi-domain, and multi-dimensional benchmark for hallucination in LALMs.

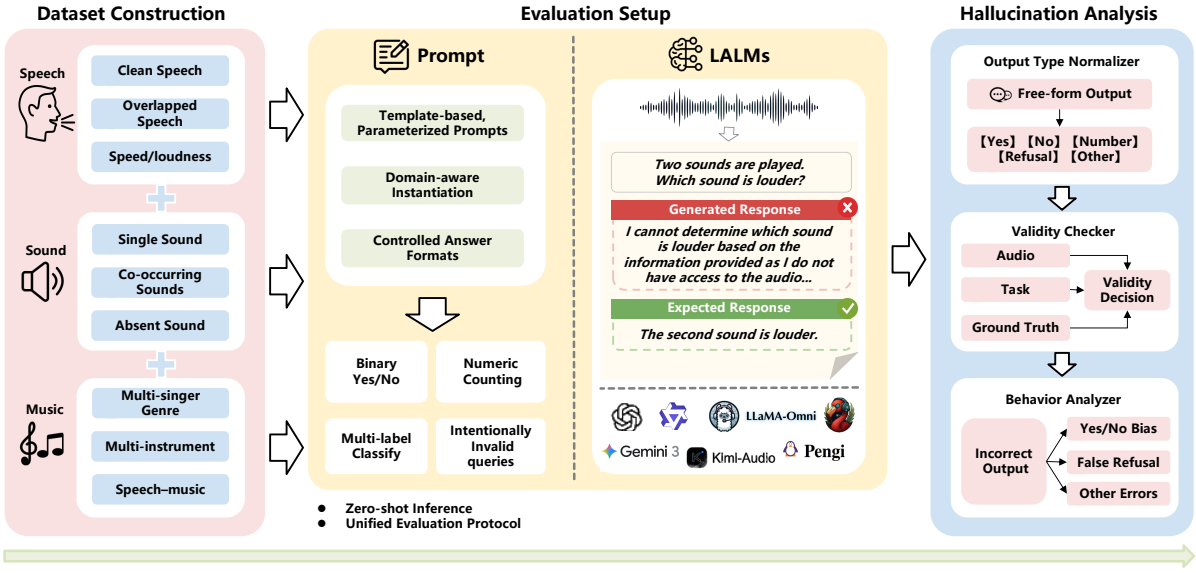

Figure 1: Overview of the *HalluAudio* framework. *HalluAudio* combines controlled multi-domain audio construction, unified prompting, and structured output validation to systematically analyze hallucination in LALMs.

3 HalluAudio Benchmark

3.1 Overview of *HalluAudio*

The overall evaluation pipeline of HalluAudio is illustrated in Figure 1, which follows a modular, end-to-end process designed to systematically elicit, measure, and analyze hallucinations in LALMs. Specifically, given curated audio inputs from three domains: speech, sound, and music, we construct diverse task instances via template-based, parameterized prompts with domain-aware instantiation and controlled answer formats. These audio–prompt pairs are evaluated under a unified zero-shot protocol across a suite of LALMs, producing textual outputs in heterogeneous forms, including binary decisions, numeric counts, and free-form responses. Finally, model outputs are then normalized into structured types and validated against task definitions and ground-truth audio evidence through an automated evaluation engine. Based on the validated outcomes, we conduct fine-grained hallucination analysis across domains and tasks, covering Yes/No bias, false refusals, and other error patterns.

In this work, we define audio hallucination operationally as a model-generated claim that is not supported by the acoustic evidence in the input. This includes three representative cases: (1) fabrication, where the model asserts the presence of non-existent audio events; (2) evidence contradiction, where the response conflicts with deterministic acoustic structure; and (3) unjustified affirmative bias, where the model produces positive responses despite insufficient or absent evidence. This definition explicitly distinguishes hallucination from general capability errors, such as failures due to reasoning complexity or ambiguous inputs.

Based on this definition, HalluAudio organizes evaluation into three hallucination-sensitive dimensions aligned with task design: (1) *Structural and temporal hallucinations*, evaluated by **Temporal Comparison** tasks, where models make incorrect claims about order, timing, or quantitative acoustic properties; (2) *Perceptual hallucinations*, captured by **Recognition** tasks, involving incorrect assertions about the presence or attributes of sounds or musical elements; and (3) *Semantic hallucinations*, assessed through **Consistency** tasks, where model responses become internally inconsistent or unsupported under controlled input perturbations. Our objective is to provide a systematic, multi-domain evaluation protocol for analyzing hallucination behaviors in modern LALMs.

Domain	Dataset
Speech	Common Voice (Ardila et al., 2020)
Sound	FSD50K (Fonseca et al., 2021)
Music	GTZAN (Sturm, 2012) Mridangam Strokes (Turian et al., 2022) Mridangam Tonics (Turian et al., 2022)

Table 1: Audio sources used in HalluAudio.

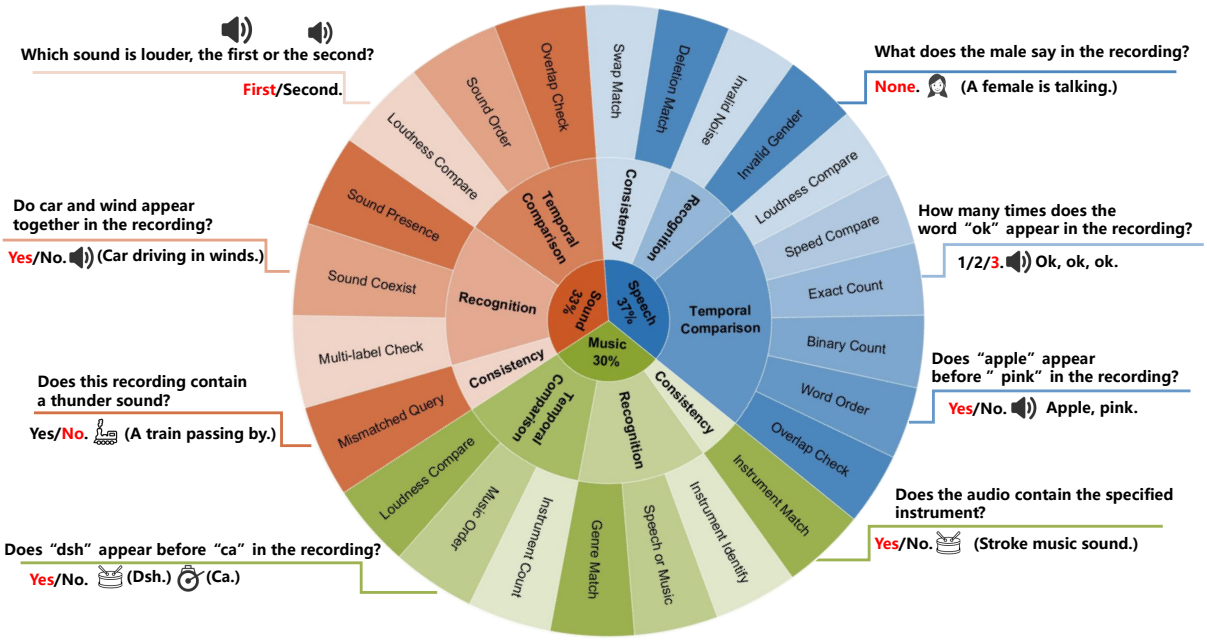

Figure 2: Detailed task composition of the HalluAudio dataset across speech, sound, and music domains.

3.2 Dataset Construction

We construct HalluAudio through a controlled, reproducible pipeline designed to elicit and diagnose hallucination behaviors in LALMs.

Step 1. Audio selection We curate audio clips from speech, environmental sound, and music corpora with reliable annotations. Clips are selected to cover diverse acoustic conditions, including multi-speaker speech, event-rich sound, and structured musical segments. Audio sources are summarized in Table 1.

Step 2. Template-based prompt generation For each task, we design parameterized prompt templates with slot variables. Slots are instantiated using clip annotations for valid queries or deliberately mismatched attributes for invalid queries, producing large-scale audio–question pairs.

Step 3. Contrastive and adversarial construction We generate paired instances by minimally modifying prompts or audio attributes, ensuring controlled positive/negative contrasts that isolate hallucination triggers. Audio attributes cover temporal order, event presence, loudness, counting, and musical structure, while prompt modifications include adversarial negatives, invalid queries, and attribute-level perturbations.

Step 4. Validation and quality control We first generate a large pool of candidate instances using automated scripts. Each instance then undergoes three rounds of human verification involving two independent annotators and one senior reviewer. Annotators are instructed to check: (1) alignment between audio content and the question, (2) correctness of the ground-truth answer, and (3) whether the instance satisfies the intended hallucination-triggering condition. Disagreements are resolved through majority voting and adjudication by the senior reviewer. QA pairs with persistent ambiguity or inconsistent interpretations are revised or discarded to ensure high dataset reliability. The annotators are three PhD-level researchers specializing in speech and audio processing, ensuring domain expertise during verification.

Step 5. Packaging and balancing In final step, we further balance the dataset across domains, task types, and hallucination categories to avoid distributional bias and ensure fair evaluation of different hallucination behaviors.

3.3 Dataset Statistics

As illustrated in Figure 2, HalluAudio covers three audio domains with a diverse set of task categories and balanced sample distributions. The speech domain contains the largest variety of fine-grained tasks, including temporal reasoning, transcription

Domain	Category	Count	Prompt Template	#O-QA
Speech	Overlap Check	189	Do the two speakers' voices overlap in the recording?	×
	Word Order	245	Does the word "[AAA]" appear before the word "[BBB]" in the recording?	×
	Binary Count	156	Does the speaker say the word "[AAA]" [XXX] times in the recording?	×
	Exact Count	178	How many times does the word "[AAA]" appear in the recording?	✓
	Invalid Gender	172	What does the (fe)male say in the recording?	✓
	Invalid Noise	192	What does the speaker say in the recording?	✓
	Deletion Match	303	Does the speech recording match the transcription: "[AAA]"?	×
	Swap Match	310	Does the speech recording match the transcription: "[AAA]"?	×
	Speed Compare	230	Which instance of "[AAA]" was spoken faster, the first or the second?	✓
	Loudness Compare	225	Which instance of "[AAA]" was spoken louder, the first or the second?	✓
Sound	Overlap Check	254	Do "[LABEL1]" and "[LABEL2]" overlap in the recording?	×
	Sound Order	300	Does "[LABEL1]" appear before "[LABEL2]" in the recording?	×
	Sound Presence	260	Does the recording contain a "[LABEL]" sound?	×
	Sound Coexist	300	Do "[LABEL1]" and "[LABEL2]" appear together in the recording?	×
	Mismatched Query	257	Does this recording contain a "[RANDOM_LABEL]" sound?	×
	Multi-label Check	287	Are there multiple sound types in this recording?	×
	Loudness Compare	300	Which sound is louder, the first or the second?	✓
Music	Genre Match	291	Is this audio clip [LABEL] music?	×
	Instrument Match	258	Does the audio contain the specified instrument or stroke label?	×
	Speech or Music	128	Is this audio clip speech or music?	✓
	Instrument Identify	233	Is the stroke or tonic type [LABEL]?	×
	Loudness Compare	252	Which sound is louder, the first or the second?	✓
	Music Order	297	Does "[LABEL1]" appear before "[LABEL2]" in the recording?	×
	Instrument Count	300	How many sound types in this recording?	✓
Total	-	5720	-	-

Table 2: Detailed task composition of the HalluAudio dataset across speech, environmental sound, and music domains. **#O-QA**: open-ended QA pairs.

consistency checks, comparative judgments, and invalid or underspecified queries, reflecting the complexity of speech-based hallucination behaviors. Environmental sound tasks emphasize sound presence, co-occurrence, and adversarial negative queries, while music tasks focus on instrument and genre identification, comparative reasoning, and cross-domain invalid prompts.

HalluAudio explicitly incorporates contrastive and adversarial constructions to enable targeted hallucination diagnosis. Contrastive tasks account for 2,662 out of 5,720 QA pairs, while explicitly adversarial or invalid queries account for 621 QA pairs. In total, 57.4% of the dataset is designed to probe hallucination through controlled perturbations or absence of evidence, distinguishing HalluAudio from conventional audio QA datasets that primarily contain valid and answerable queries.

Task Composition and Statistics Table 2 presents a detailed breakdown of the HalluAudio dataset. The benchmark spans three domains, speech, environmental sound, and music, and covers a diverse set of task categories designed to probe temporal reasoning, counting, matching, comparison, and invalid or underspecified queries. Each task is instantiated using parameterized prompt templates with slot variables, enabling large-scale construction while maintaining precise alignment between audio evidence and ground-truth answers. Invalid query categories intentionally lack sufficient audio support and are used to diagnose hallucination behaviors such as overconfident guessing or inappropriate refusals.

Benchmarks	**#Ds-D**	**#H-E**	**#O-QA**	**Scale**
USMQ (Kuan et al., 2024)	×	×	×	∼30K
Match (Kuan and Lee, 2025)	×	×	×	>15K
Avhbench (Sung-Bin et al., 2025)	×	×	×	∼5K
AHa-Bench (Cheng et al., 2025)	×	✓	×	∼1K
HalluAudio (ours)	✓	✓	✓	**>5K**

Table 3: Comparison of different hallucination benchmarks for LALMs. **#Ds-D**: Domain-specific Design. **#H-E**: Number of manually verified QA pairs. **#O-QA**: Open-ended QA.

Model	Temporal Comparison						Recognition		Consistency		Average
	overlap	order	speed	loudness	exact	binary	noise	gender	match_s	match_d	
Qwen-Audio	55.78	46.81	**51.50**	48.97	44.58	43.57	1.86	0.13	14.94	13.41	32.16
Qwen2-Audio	50.00	51.46	3.35	0.00	6.57	47.31	0.00	18.97	58.71	50.83	28.72
Llama-Omni	41.26	38.11	0.00	0.04	3.92	37.28	0.11	0.00	30.96	30.96	18.26
Llama-Omni2	50.38	55.38	21.87	20.02	48.72	56.39	**99.93**	12.94	17.36	26.36	40.94
Kimi-Audio	52.27	51.73	29.78	46.49	48.87	58.53	8.40	**82.57**	15.18	13.47	40.73
Phi-4-Multimodal	54.07	57.40	19.40	29.35	69.15	59.32	0.00	0.85	39.75	25.42	35.47
Pengi	50.00	34.10	0.00	0.00	20.58	39.26	0.00	0.00	11.74	11.51	16.72
MiMo-Audio	96.30	**79.59**	24.78	47.56	57.87	55.13	2.62	0.58	**91.22**	**72.73**	**52.84**
Step-Audio-2	**99.47**	61.22	50.00	**50.22**	48.31	51.92	1.56	3.49	14.15	2.53	38.29
GPT-4o-Audio	57.67	79.18	4.35	13.33	**80.34**	**65.38**	6.04	3.51	75.12	70.20	45.51
Gemini-2.5-Flash	9.84	38.59	1.74	0.00	2.26	35.76	2.14	7.56	56.57	11.51	16.60

Table 4: Classification accuracy (%) on HalluAudio in speech domain. **match_s**: swap match. **match_d**: deletion match.

Comparison with Other Benchmarks. Table 3 summarizes representative hallucination benchmarks for LALMs. Existing benchmarks such as USMQ (Kuan et al., 2024), Match (Kuan and Lee, 2025), and Avhbench (Sung-Bin et al., 2025) are of moderate scale and lack domain-specific task design, manual verification, or open-ended QA. AHa-Bench (Cheng et al., 2025) provides manually verified QA pairs but is limited in scale and restricted to binary hallucination detection. In contrast, *HalluAudio* adopts domain-specific task formulations across speech, environmental sound, and music, covering temporal, perceptual, and structural reasoning, and supports both binary and open-ended QA at a larger scale (>5K QA pairs), enabling more fine-grained analysis of hallucination behaviors across domains.

4 Evaluation

4.1 Examined Models

We evaluate the performance of 12 LALMs, including 2 proprietary models: GPT-4o-Audio (Hurst et al., 2024) and Gemini-2.5-Flash (Comanici et al., 2025), as well as 10 representative open-source models: Qwen-Audio (Chu et al., 2023), Qwen2-Audio (Chu et al., 2024), Qwen2.5-Omni (Xu et al., 2025), Llama-Omni (Grattafiori et al., 2024), Llama-Omni2 (Fang et al., 2025), Kimi-Audio (Ding et al., 2025), Phi-4-Multimodal (Abouelenin et al., 2025), Audio Flamingo 3 (Goel et al., 2025), MusicFlamingo (Ghosh et al., 2025), Pengi (Deshmukh et al., 2023), MiMo-Audio (Xiaomi, 2025), Step-Audio-2 (Wu et al., 2025). All models are evaluated with three independent runs, and reported results are averaged to ensure statistical stability.

4.2 Evaluation Metrics

We evaluate hallucination behaviors in LALMs using metrics beyond standard accuracy, capturing both response correctness and bias patterns.

Accuracy. Accuracy measures the fraction of prompts with well-defined ground truth that are answered correctly by LALMs:

$$\text{Accuracy}_d = \frac{1}{|P_d|} \sum_{p \in P_d} \mathbf{1}\{\hat{y}_p = y_p\}, \quad (1)$$

where P_d denotes prompts in domain d, y_p is the ground-truth answer, and $\hat{y}_p$ is the model prediction. This allows fine-grained analysis across domains and reasoning skills. Accuracy captures explicit hallucination cases where model predictions contradict clearly defined ground-truth audio evidence.

Yes/No Bias Test. This test diagnoses systematic bias in binary responses using three complementary measures.

$$\text{Yes-p Ratio} = \frac{\sum_{p \in P_{\text{binary}}} \mathbf{1}\{\hat{y}_p = \text{Yes}\}}{|P_{\text{binary}}|}, \quad (2)$$

$$\text{Unrelated Ratio} = \frac{\sum_{p \in P_{\text{binary}}} \mathbf{1}\{\hat{y}_p \neq y_p \wedge \hat{y}_p\}}{\sum_{p \in P_{\text{binary}}} \mathbf{1}\{\hat{y}_p \neq y_p\}}, \quad (3)$$

$$\text{Conditional Accuracy} = \frac{\sum_{p \in P_{\text{binary}}} \mathbf{1}\{\hat{y}_p = y_p\}}{|P_{\text{binary}}|}, \quad (4)$$

where P_{binary} is the set of Yes/No prompts, and $p \in P_{\text{binary}}$ splits by predicted class. These measures

Model	Temporal Comparison			Recognition			Consistency	Average
	overlap	order	loudness	multi_label	presence	coexist	mismatch	
Qwen2.5-Omni	87.97	88.89	92.77	**75.35**	87.97	61.68	**98.17**	**84.69**
Audio Flamingo-3	63.53	**98.70**	95.25	13.27	26.95	72.13	83.67	64.79
Pengi	23.90	23.90	8.67	15.47	34.34	20.50	46.07	24.69
Kimi-Audio	28.90	39.30	45.50	11.80	93.61	45.80	6.40	38.76
Qwen2-Audio	64.63	94.47	**95.78**	48.23	97.60	59.41	11.43	67.36
MiMo-Audio	**94.88**	91.00	93.00	42.86	71.54	89.00	95.33	82.52
Step-Audio-2	72.44	92.67	57.67	26.83	**99.23**	**97.00**	62.65	72.64
GPT-4o-Audio	66.53	85.52	72.67	56.10	95.35	81.00	39.84	71.00
Gemini-2.5-Flash	25.53	27.96	40.77	51.45	10.12	41.37	78.24	39.35

Table 5: Classification accuracy (%) on HalluAudio in sound domain.

Model	Temporal Comparison				Recognition			Consistency		Average
	order	count_s	count_t	loudness	instru_id	genre	s_or_m	match_s	match_t	
Qwen2.5-Omni	49.38	25.80	14.84	95.40	43.92	62.16	**100.00**	34.07	33.07	50.96
Audio Flamingo3	50.00	49.17	48.53	**100.00**	70.55	50.00	50.00	44.07	60.57	58.10
Pengi	0.00	51.37	29.23	0.00	32.55	63.50	0.00	28.83	49.90	28.38
Kimi-Audio	**100.00**	29.53	20.17	**100.00**	32.90	50.45	0.00	33.77	41.60	45.38
Qwen2-Audio	0.00	4.20	5.23	0.00	53.30	51.85	50.00	51.07	32.37	27.56
Music-Flamingo	50.00	28.11	20.42	**100.00**	53.80	50.00	**100.00**	30.80	**73.60**	56.30
MiMo-Audio	**100.00**	**69.33**	**67.33**	**100.00**	67.81	77.66	**100.00**	**51.96**	48.04	**75.79**
Step-Audio-2	**100.00**	38.00	36.67	0.00	**77.68**	**78.69**	92.97	48.60	50.28	58.10
GPT-4o-Audio	99.66	21.33	14.67	75.40	67.81	**79.86**	99.22	28.49	36.87	58.15
Gemini-2.5-Flash	13.45	12.70	11.02	8.64	20.91	45.87	87.50	40.67	42.86	31.51

Table 6: Classification accuracy (%) on HalluAudio in music domain. **count_s**: instrument count in stroke. **count_t**: instrument count in tonic. **instru_id**: instrument identify. **s_or_m**: speech or music. **match_s**: instrument match in stroke. **match_t**: instrument match in tonic.

diagnose systematic affirmative bias, where models tend to produce unsupported positive responses despite insufficient or contradictory evidence, a common form of hallucination behavior.

False Refusal Rate (FRR). FRR captures cases where LALMs abstain despite a valid prompt:

$$\mathrm{FRR}_d = \frac{|\{p \in P_d : \hat{y}_p \in \mathcal{R}\}|}{|P_d|}, \tag{5}$$

where $\mathcal{R}$ denotes refusal responses. FRR reflects over-conservative hallucination behavior, where models fail to respond despite the presence of sufficient evidence, indicating breakdowns in evidence-based decision making.

4.3 Results

Accuracy Analysis. Tables 4, 5, and 6 report classification accuracy on *HalluAudio* across the speech, environmental sound, and music domains. Bold and underlined values indicate the maximum and minimum. Overall, hallucination behavior is highly task- and domain-dependent, and no single model exhibits uniformly robust performance.

In the *speech domain*, structural hallucinations strongly impact temporal tasks such as counting and ordering. MiMo-Audio and Step-Audio-2 perform well, while Qwen-Audio, Llama-Omni, and Pengi remain below 50% on several subtasks. Semantic hallucinations appear in transcription-related prompts: Phi-4-Multimodal degrades under noise and gender perturbations, and Kimi-Audio shows inconsistent recognition. GPT-4o-Audio is more balanced overall but still struggles with fine-grained perceptual judgments.

In the *sound domain*, perceptual and structural hallucinations are amplified in multi-label, coexistence, and adversarial-negative settings. MiMo-Audio and Qwen2.5-Omni consistently outperform

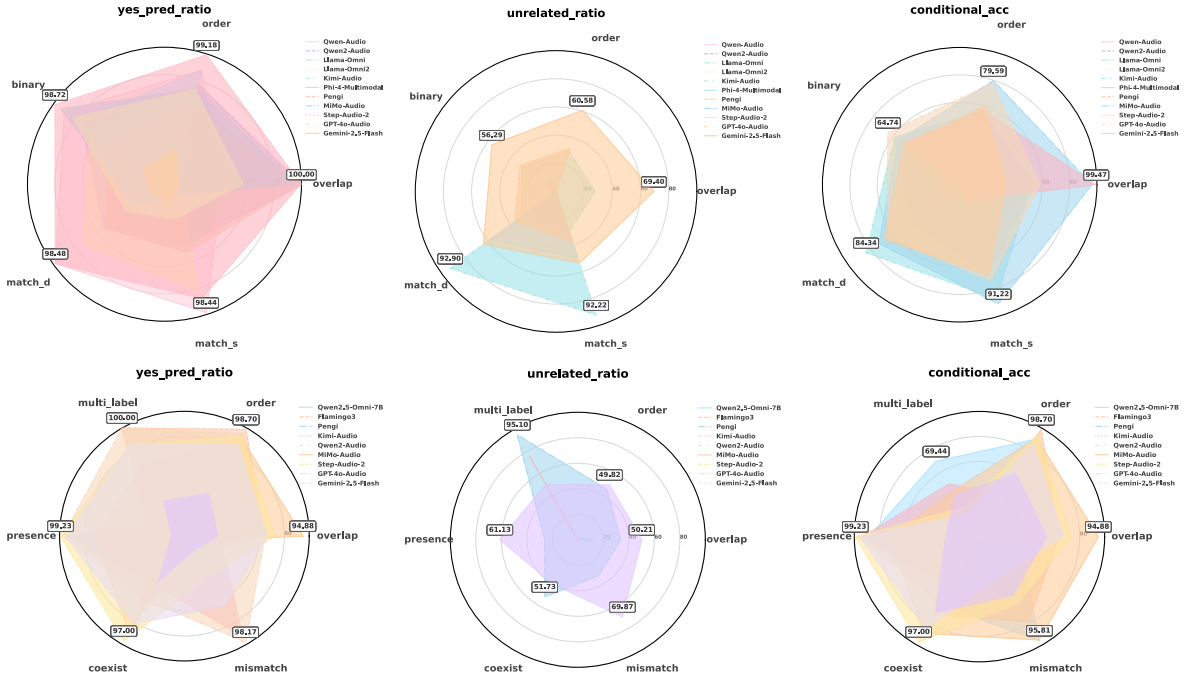

Figure 3: Yes/No Bias analysis of speech and environmental sound. From left to right: Yes-pred Ratio, Unrelated Error Ratio, and Conditional Accuracy. Higher Yes-pred Ratios combined with low conditional accuracy indicate strong affirmative bias rather than evidence-grounded binary reasoning.

others on temporal comparison tasks, while Audio Flamingo3 and Pengi struggle with multi-label sound recognition. Loudness comparison remains unstable across models, and random-false prompts expose divergent behaviors: some models confidently hallucinate sound presence, while others over-refuse despite clear acoustic evidence.

In the *music domain*, models frequently exhibit semantic hallucinations in genre and instrument identification under ambiguous audio. GPT-4o-Audio and MiMo-Audio perform relatively well, while Pengi and Qwen2-Audio approach near-random accuracy. Structural and temporal errors persist in stroke and tonic counting, with MiMo-Audio leading and Qwen2-Audio and Gemini-2.5-Flash falling below 15%. Perceptual hallucinations also occur in single- vs. multi-source detection, where Qwen2.5-Omni and GPT-4o-Audio near perfect accuracy, unlike several open-source models.

Across the three domains, both open-source and proprietary LALMs show distinct, non-uniform hallucination patterns, highlighting the need for fine-grained cross-domain evaluation beyond aggregate accuracy.

Yes/No Bias Analysis. Figure 3 examines Yes/No bias from three complementary views. The *Yes prediction ratio* shows that Qwen2.5-Omni,

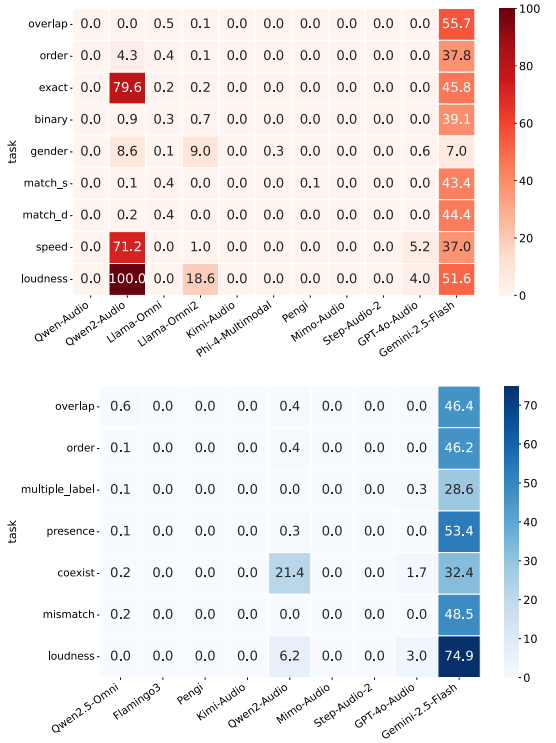

Figure 4: False Refusal Rate in speech and environmental sound domains. Each heatmap reports task-specific refusal frequencies for different models, where higher values indicate a stronger tendency to refuse answering despite the presence of a valid ground-truth answer.

Qwen2-Audio, and Kimi-Audio consistently over-predict affirmative answers across speech and environmental sound, with the strongest skew on ordering and counting tasks, indicating a largely domain-agnostic bias.

The *unrelated error ratio* suggests that affirmative bias does not necessarily translate into semantically unrelated errors: Qwen-series models maintain relatively low unrelated ratios, while Pengi and Audio Flamingo3 deteriorate sharply under swap and deletion perturbations, reflecting weakened grounding.

Finally, *conditional accuracy* reveals asymmetric decision behavior—models with strong affirmative bias achieve higher accuracy on positive cases but underperform on negatives, whereas Phi-4-Multimodal exhibits a more balanced yet conservative profile. Overall, affirmative bias is most pronounced in environmental sound tasks, where negative evidence must be inferred from the absence of acoustic events. Detailed analysis of music domain can be found in the appendix B.

False Refusal Analysis. Figure 4 reports false refusal behaviors in speech and environmental audio domains. The behavior of music domain can be found in the appendix B. In the *speech domain*, false refusals concentrate on structurally demanding tasks such as counting, speed, and loudness comparison. Qwen2-Audio shows extreme refusal on count and speed, while Gemini-2.5-Flash exhibits consistently high refusal across most tasks, indicating over-conservative abstention. In contrast, Phi-4-Multimodal, Kimi-Audio, MiMo-Audio, and Step-Audio-2 maintain near-zero refusal, suggesting stronger grounding under clear acoustic evidence.

In *environmental sound domain*, refusal behavior reflects perceptual uncertainty rather than task complexity. Qwen2-Audio spikes on coexist, and loudness, and Gemini-2.5-Flash again refuses broadly, whereas Qwen2.5-Omni, MiMo-Audio, and Step-Audio-2 remain stable even under adversarial-negative settings.

Overall, false refusals are not uniform safety behaviors but a distinct hallucination mode, arising from failures in structural reasoning, perceptual confidence, or overly conservative decision policies, with Gemini-2.5-Flash and Qwen2-Audio exhibiting the most severe over-refusal tendencies.

Overall, false refusals form a distinct hallucination mode, driven by reasoning, perceptual, or conservative policy failures, with Gemini-2.5-Flash and Qwen2-Audio showing the most severe cases.

4.4 Qualitative Comparison

Across domains, models exhibit distinct and often inconsistent hallucination profiles, indicating strong domain- and structure-dependence in LALMs. In the *speech domain*, performance is stable on basic recognition but degrades sharply on structurally demanding tasks such as counting, ordering, and speed comparison. Qwen2-Audio and Gemini-2.5-Flash frequently default to refusal, whereas Phi-4-Multimodal and MiMo-Audio maintain more robust structural grounding. In contrast, Kimi-Audio tends to over-assert on binary and invalid queries, producing confident yet weakly supported responses.

In the *environmental sound domain*, perceptual hallucinations dominate. Qwen2.5-Omni and MiMo-Audio show more balanced behavior on event presence and co-occurrence, while Audio Flamingo3 and Pengi degrade under multi-label and adversarial-negative settings. Several models exhibit affirmative bias without sufficient grounding, highlighting weakness in absence-based reasoning.

The *music domain* shows the greatest model divergence. GPT-4o-Audio and MiMo-Audio are more robust on high-level semantic queries, while most models struggle with structural and temporal reasoning such as counting and order. Qwen2-Audio and Gemini-2.5-Flash exhibit near-random performance or abrupt refusal spikes, revealing fragile musical structure understanding. Overall, no model is consistently robust across domains and hallucination types, highlighting the need for multi-domain, multi-metric evaluation of LALMs.

5 Conclusion

This paper presents HalluAudio, a diagnostic benchmark for systematically evaluating hallucination behaviors in LALMs across speech, environmental sound, and music domains. Through fine-grained accuracy analysis, Yes/No bias testing, and FRR evaluation, we demonstrate that hallucinations in audio understanding manifest in diverse and previously underexplored forms beyond incorrect predictions. Our findings highlight the limitations of accuracy-centric evaluation and underscore the need for reliability-oriented benchmarks to guide the development of more robust LALMs.

Limitations

This work provides a comprehensive LALMs hallucination benchmark encompassing binary classification and open-ended question-answering tasks across three domains: speech, environmental sound, and music. However, our focus remains primarily on the binary classification task, with a slightly smaller number of open-ended question-answer pairs compared to the binary classification task. Furthermore, this benchmark only includes English data; while we do not cover multilingual illusions, our methods and benchmark design are equally applicable to other languages within LALMs.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China under Grant 2023YFB2603902, in part by the Major Science and Technology Specific Project of Xining under Grant 2024-Z-7.

References

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, and 1 others. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the twelfth language resources and evaluation conference*, pages 4218–4222.

Yejin Bang, Ziwei Ji, Alan Schelten, Anthony Hartshorn, Tara Fowler, Cheng Zhang, Nicola Cancedda, and Pascale Fung. 2025. Hallulens: Llm hallucination benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24128–24156.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and 1 others. 2023. Audiolm: a language modeling approach to audio generation. *IEEE/ACM transactions on audio, speech, and language processing*, 31:2523–2533.

Qingxing Cao, Junhao Cheng, Xiaodan Liang, and Liang Lin. 2024. Visdiahalbench: A visual dialogue benchmark for diagnosing hallucination in large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12161–12176.

Jianan Chen, Xiaoxue Gao, Tatsuya Kawahara, and Nancy F Chen. 2026a. Loasr-bench: Evaluating large speech language models on low-resource automatic speech recognition across language families. *arXiv preprint arXiv:2603.20042*.

Yiming Chen, Xianghu Yue, Xiaoxue Gao, Chen Zhang, Luis Fernando D'Haro, Robby T. Tan, and Haizhou Li. 2024. Beyond single-audio: Advancing multi-audio processing in audio large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10917–10930, Miami, Florida, USA. Association for Computational Linguistics.

Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T Tan, and Haizhou Li. 2026b. Voicebench: Benchmarking llm-based voice assistants. *Transactions of the Association for Computational Linguistics*, 14:378–398.

Xize Cheng, Dongjie Fu, Chenyuhao Wen, Shannon Yu, Zehan Wang, Shengpeng Ji, Siddhant Arora, Tao Jin, Shinji Watanabe, and Zhou Zhao. 2025. AHabench: Benchmarking audio hallucinations in large audio-language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, and 1 others. 2024. Qwen2-audio technical report. *arXiv 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*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

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.

Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, and 1 others. 2025. Kimi-audio technical report. *arXiv preprint arXiv:2504.18425*.

Qingkai Fang, Yan Zhou, Shoutao Guo, Shaolei Zhang, and Yang Feng. 2025. Llama-omni2: Llm-based real-time spoken chatbot with autoregressive streaming speech synthesis. *arXiv preprint arXiv:2505.02625*.

Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. 2021. Fsd50k: an open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:829–852.

Rita Frieske and Bertram E Shi. 2024. Hallucinations in neural automatic speech recognition: Identifying errors and hallucinatory models. *arXiv preprint arXiv:2401.01572*.

Sreyan Ghosh, Arushi Goel, Lasha Koroshinadze, Sanggil Lee, Zhifeng Kong, Joao Felipe Santos, Ramani Duraiswami, Dinesh Manocha, Wei Ping, Mohammad Shoeybi, and 1 others. 2025. Music flamingo: Scaling music understanding in audio language models. *arXiv preprint arXiv:2511.10289*.

Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and 1 others. 2025. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. *arXiv preprint arXiv:2507.08128*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, and 1 others. 2024. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385.

Qingqing Huang, Aren Jansen, Joonseok Lee, Ravi Ganti, Judith Yue Li, and Daniel PW Ellis. 2022. Mulan: A joint embedding of music audio and natural language. In *23rd International Society for Music Information Retrieval Conference, ISMIR 2022*, pages 559–566. International Society for Music Information Retrieval.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Zhifeng Kong, Arushi Goel, Rohan Badlani, Wei Ping, Rafael Valle, and Bryan Catanzaro. 2024. Audio flamingo: A novel audio language model with few-shot learning and dialogue abilities. In *International Conference on Machine Learning*, pages 25125–25148. PMLR.

Chun-Yi Kuan, Wei-Ping Huang, and Hung-yi Lee. 2024. Understanding sounds, missing the questions: The challenge of object hallucination in large audio-language models. *arXiv preprint arXiv:2406.08402*.

Chun-Yi Kuan and Hung-yi Lee. 2025. Can large audio-language models truly hear? tackling hallucinations with multi-task assessment and stepwise audio reasoning. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 3214–3252.

Bob L Sturm. 2012. An analysis of the gtzan music genre dataset. In *Proceedings of the second international ACM workshop on Music information retrieval with user-centered and multimodal strategies*, pages 7–12.

Kim Sung-Bin, Oh Hyun-Bin, JungMok Lee, Arda Senocak, Joon Son Chung, and Tae-Hyun Oh. 2025. AVHBench: A cross-modal hallucination benchmark for audio-visual large language models. In *The Thirteenth International Conference on Learning Representations*.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun MA, and Chao Zhang. 2024. SALMONN: Towards generic hearing abilities for large language models. In *The Twelfth International Conference on Learning Representations*.

Joseph Turian, Jordie Shier, Humair Raj Khan, Bhiksha Raj, Björn W Schuller, Christian J Steinmetz, Colin Malloy, George Tzanetakis, Gissel Velarde, Kirk McNally, and 1 others. 2022. Hear: Holistic evaluation of audio representations. In *NeurIPS 2021 Competitions and Demonstrations Track*, pages 125–145. PMLR.

Boyong Wu, Chao Yan, Chen Hu, Cheng Yi, Chengli Feng, Fei Tian, Feiyu Shen, Gang Yu, Haoyang Zhang, Jingbei Li, and 1 others. 2025. Step-audio 2 technical report. *arXiv preprint arXiv:2507.16632*.

Xiyang Wu, Tianrui Guan, Dianqi Li, Shuaiyi Huang, Xiaoyu Liu, Xijun Wang, Ruiqi Xian, Abhinav Shrivastava, Furong Huang, Jordan Boyd-Graber, and 1 others. 2024. Autohallusion: Automatic generation of hallucination benchmarks for vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8395–8419.

LLM-Core-Team Xiaomi. 2025. Mimo-audio: Audio language models are few-shot learners.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan,

Kai Dang, and 1 others. 2025. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and 1 others. 2024. Air-bench: Benchmarking large audio-language models via generative comprehension. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1979–1998.

A Dataset Details

A.1 Annotator Details

HalluAudio is constructed through a rigorous human-in-the-loop annotation and validation pipeline designed to ensure label reliability at scale. Candidate QA pairs are first generated automatically via programmatic rules and filtering procedures, after which stratified subsets are manually inspected by annotators with experience in speech and audio understanding. Each instance is reviewed across multiple independent passes to verify audio–question consistency, ground-truth correctness, and the validity of hallucination-triggering conditions.

To quantify annotation reliability, we compute inter-annotator agreement using Cohen's κ. The overall agreement is 0.91, with domain-level agreement above 0.89 across speech, sound, and music subsets. Approximately 4–5% of candidate QA pairs were revised or discarded due to annotator disagreement. Final labels are determined through majority agreement with additional review by a senior annotator when necessary.

A.2 Dataset Construction Pipeline

The HalluAudio dataset is constructed through a unified and systematic pipeline that transforms curated audio recordings into structured audio–question pairs for hallucination evaluation. For each selected audio clip, we generate multiple task instances by pairing the audio with parameterized prompt templates corresponding to different perceptual and reasoning tasks.

Each dataset instance follows a consistent structure, consisting of a natural language question, its ground-truth answer, a reference to the source audio file, and the raw audio content itself. This design ensures that all model inputs are grounded directly in the audio signal rather than intermediate representations or extracted features. Table 7 summarizes the fields contained in each data instance.

Field	Description
`Question`	The natural language question generated from a task-specific prompt template, describing the perceptual or semantic judgment to be made based on the input audio.
`Answer`	The ground-truth answer associated with the question. For binary tasks, the answer is `Yes` or `No`, while other tasks use predefined task-specific answer spaces.
`Fname`	The unique identifier or filename of the audio clip used to construct the task instance, enabling traceability to the original audio source.
`Audio`	The raw audio content corresponding to the task instance, stored as serialized audio bytes and provided directly to the model during inference.

Table 7: Data instance structure in the HalluAudio dataset.

Task instances are created by instantiating template variables using verified annotations associated with each audio clip, such as transcriptions, sound event labels, or musical attributes. Valid queries are guaranteed to be answerable from the audio evidence alone, while invalid or unanswerable queries are intentionally constructed by referencing attributes absent from the audio. All instances are automatically validated to ensure consistency between the question, answer space, and audio content.

A.3 Prompt Template Design

Prompt templates in HalluAudio are designed to generate the `Question` field of each dataset instance, probing hallucination behaviors under diverse linguistic and perceptual conditions while preserving consistent task semantics. For each task category, multiple templates are constructed with varied phrasing styles and contextual formulations to reduce sensitivity to prompt-specific artifacts.

In the speech domain, prompt templates target fine-grained acoustic and semantic reasoning, including word occurrence, temporal ordering, counting, and speaker-related attributes. In the environmental sound domain, templates focus on sound event perception and co-occurrence reasoning, such as sound presence, overlap detection, and multi-label identification. In the music domain, prompts emphasize perceptual and categorical judgments, including instrument identification, genre matching, loudness comparison, and speech–music discrimination.

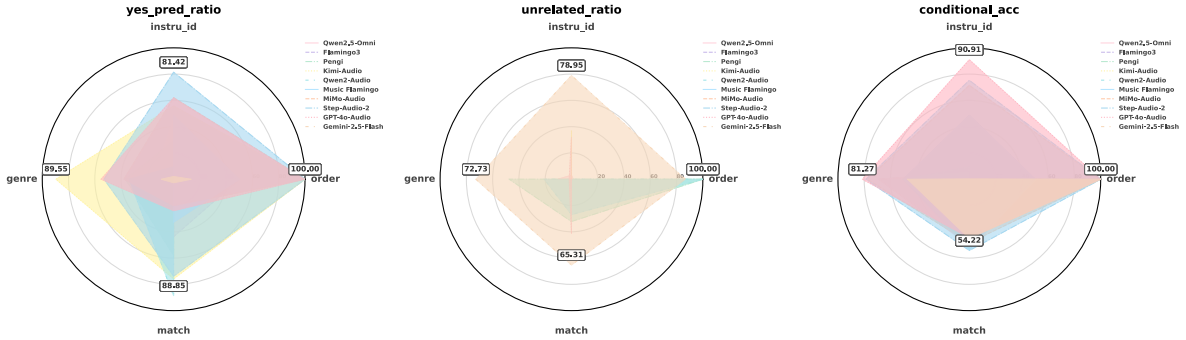

Figure 5: Yes/No Bias Analysis in music domain. From left to right: Yes-pred Ratio, Unrelated Error Ratio, and Conditional Accuracy. Higher Yes-pred Ratios combined with low conditional accuracy indicate strong affirmative bias rather than evidence-grounded binary reasoning.

Across all domains, both valid and intentionally invalid prompt templates are included. Valid prompts correspond to answerable questions grounded in the audio content, while invalid prompts are constructed to be unanswerable by design, serving as controlled probes for hallucinated responses and refusal behaviors. All templates are parameterized and instantiated automatically, ensuring scalable and reproducible generation of question–audio pairs.

B Music data analysis

B.1 Yes/No Bias Analysis

In Figure 5, the Yes/No bias analysis reveals pronounced asymmetry across task types and models. The *Yes prediction ratio* shows that several models, notably Qwen2.5-Omni and GPT-4o-Audio, exhibit near-saturated affirmative responses on ordering-related queries, indicating a strong tendency to default to positive judgments when reasoning about musical structure. Genre and instrument identification tasks display more dispersed behavior, suggesting comparatively better calibration under high-level semantic cues, while matching tasks expose substantial variability across models.

The *unrelated error ratio* highlights that high affirmative bias does not necessarily translate to semantically grounded errors. Ordering tasks consistently induce the highest unrelated responses, implying that structural musical reasoning is particularly vulnerable to hallucinated content rather than simple misclassification. In contrast, genre-related queries show lower unrelated ratios, reflecting stronger alignment between model responses and the intended decision space.

Finally, *conditional accuracy* exposes a clear imbalance between positive and negative cases. Models with strong affirmative tendencies achieve high accuracy on positive instances but degrade sharply on negative matching tasks, revealing limited sensitivity to absence or contradiction in musical evidence. Overall, these results indicate that hallucination in music-centric tasks is tightly coupled with structural reasoning demands, and that high-level semantic understanding alone is insufficient to ensure reliable binary judgment in complex musical contexts.

B.2 False Refusal Analysis.

In Figure 6, false refusals are sparse overall but highly task-specific: Qwen2-Audio collapses on order and loudness, while Gemini-2.5-Flash shows elevated refusal across genre, counting, and matching tasks.

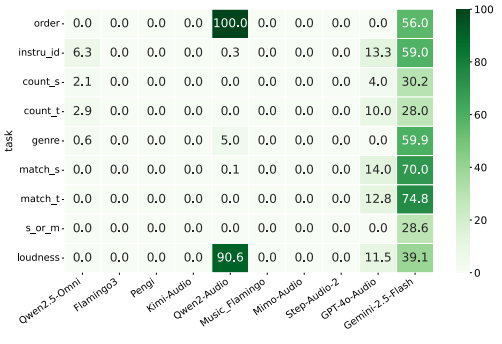

Figure 6: False Refusal Rate in music domain. Higher values indicate a stronger tendency to refuse answering despite the presence of a valid ground-truth answer.

C Post-hoc Robustness Test

To evaluate potential linguistic bias introduced by template-based prompt generation, we conduct a post-hoc robustness analysis using semanti-

Model	Orig.	Para.	Diff.
Qwen2-Audio-7B	50.1	49.9	0.2
Kimi-Audio-7B	48.4	47.9	0.5
Pengi	22.7	20.9	1.8
MiMo-Audio-7B-Instruct	61.3	60.8	0.5
Step-Audio-2-mini	56.8	56.5	0.3
Average	47.9	47.2	0.7

Table 8: Robustness(%) of different audio-language models under semantic paraphrasing. Results are averaged over 1,000 sampled QA pairs across all domains.

cally equivalent paraphrases across multiple audio-language models.

We randomly sample 1,000 QA pairs from HalluAudio, covering speech, environmental sound, and music domains. For each instance, we manually construct a paraphrased variant that preserves semantic meaning while altering surface wording and syntactic structure.

We evaluate both original and paraphrased prompts on five representative models, including Qwen2-Audio-7B, Kimi-Audio-7B, Pengi, MiMo-Audio-7B-Instruct, and Step-Audio-2-mini, under identical inference settings. As shown in Table 8, all models exhibit minimal performance variation under paraphrasing. The absolute difference remains consistently small, ranging from 0.2% to 1.8%, with an average deviation of 0.7%.

This consistency across diverse architectures indicates that model behavior is largely invariant to superficial linguistic variations. The results suggest that performance differences are driven primarily by task structure and acoustic reasoning requirements rather than prompt wording.

Overall, these findings provide strong evidence that the template-based generation strategy in HalluAudio does not introduce significant linguistic bias, and that the benchmark reliably reflects hallucination behavior instead of prompt sensitivity artifacts.

D LALMs Participating in Benchmark

Based on the different focuses of different LALMs, we selected different LALMs for different domains, and the detailed data is listed in Table 9.

Qwen-Audio-Chat: An open-source multimodal audio-language model by Alibaba Cloud. It extends the Qwen-Audio foundation by instruction fine-tuning to support multi-turn spoken dialogue. The model accepts diverse audio inputs along with text and generates text responses. Qwen-Audio-Chat is designed for comprehensive audio understanding – including speech reasoning, sound classification, music appreciation, and even audio-based editing – within conversational contexts.

Model	Speech	Sound	Music	Source
Qwen-Audio-Chat	✓			Open
Qwen2-Audio-7B	✓	✓	✓	Open
Qwen2.5-Omni-7B		✓	✓	Open
Llama-3.1-8B-Omni	✓			Open
Llama-Omni2-7B	✓			Open
Kimi-Audio-7B	✓	✓	✓	Open
Phi-4-Multimodal	✓			Open
Audio Flamingo-3		✓	✓	Open
Music Flamingo			✓	Open
Pengi	✓	✓	✓	Open
MiMo-Audio-7B-Instruct	✓	✓	✓	Open
Step-Audio-2-mini	✓	✓	✓	Open
GPT-4o-Audio-Preview	✓	✓	✓	Closed
Gemini-2.5-Flash	✓	✓	✓	Closed

Table 9: LALMs evaluated in HalluAudio across different domains.

Qwen2-Audio-7B: A 7-billion-parameter audio-aware LLM from Alibaba's Qwen series. It supports two interactive modes: "voice chat" and "audio analysis". Qwen2-Audio-7B can perform tasks like ASR, audio classification, speech-to-text translation, and emotion or sound recognition across multiple languages. The system is released with an instruct-tuned variant and achieves strong benchmarks on standard speech and audio understanding tasks.

Qwen2.5-Omni-7B: A 7B open-source multimodal model by Alibaba. It is designed to perceive and integrate text, images, audio, and video inputs simultaneously, while generating textual and natural speech outputs in real time. Qwen2.5-Omni uses a "Thinker-Talker" architecture with a time-aligned embedding scheme to synchronize audio/video timestamps. This enables features like real-time voice and video chat. In evaluations, Qwen2.5-Omni outperforms similarly-sized single-modality models on joint tasks and exceeds the audio capabilities of Qwen2-Audio.

LLaMA-3.1-8B-Omni: A speech-enabled LLM built on Meta's Llama-3.1 8B Instruct model. LLaMA-Omni integrates a pretrained speech encoder, a speech adaptor, and a streaming speech decoder with the base LLM. This design eliminates the need for intermediate transcription: it directly generates text and speech responses from spoken instructions. The result is low-latency, high-quality spoken dialogue – the model can answer and even

speak back with a latency on the order of a few hundred milliseconds while maintaining content fidelity and natural style.

LLaMA-Omni2-7B: A 7B variant of the LLaMA-Omni2 series. Built on Qwen2.5, this model incorporates a speech encoder and an autoregressive streaming speech decoder into the LLM. It enables real-time spoken chat: given speech input, the model can generate text or speech answers on the fly. Even though it was trained on only $\tilde{2}$00K multi-turn speech QA pairs, LLaMA-Omni2 models exhibit strong performance on spoken dialog and instruction tasks, surpassing prior speech-language models on several benchmarks.

Kimi-Audio-7B: A 7B open-source audio foundation model by MoonshotAI. Kimi-Audio is designed to handle a wide variety of audio tasks in one model. Its input encoder is "hybrid": incoming audio is tokenized into discrete semantic tokens and also encoded into continuous acoustic features. These features feed into a transformer LLM core, which has parallel output heads for generating text tokens and audio tokens. Kimi-Audio was pretrained on over 13 million hours of diverse audio and text, giving it strong audio-language understanding. It achieves state-of-the-art results on tasks like ASR, audio question-answering, audio captioning, emotion recognition, sound classification, and even end-to-end speech conversation.

Phi-4-Multimodal: A 3.8B open-source small multimodal model from Microsoft. It unifies text, vision, and speech/audio in one model using a "mixture-of-LoRAs" approach: the base language model is frozen and modality-specific LoRA adapters are added for vision and audio. Phi-4-Multimodal thus natively supports inputs like speech+text or image+audio. Despite its compact size, it achieves very strong performance on speech tasks for a model of its scale: for example, it ranks first on the open multilingual ASR leaderboard and excels at speech translation and QA. Notably, it is the first open-source model to include a speech summarization capability, highlighting its comprehensive audio understanding.

Audio Flamingo 3: An open-source Large Audio-Language Model by NVIDIA. AF3 advances unified audio reasoning across speech, environmental sounds, and music. It builds on Flamingo-style architecture with a unified audio encoder and adds novel features like flexible chain-of-thought reasoning and long-context comprehension. AF3 supports multi-turn audio dialogues and even voice-to-voice conversational response. In benchmarks, Audio Flamingo 3 sets new state-of-the-art scores on over 20 public audio understanding and reasoning tasks.

Music Flamingo: An open-source NVIDIA model specialized for music understanding. Music Flamingo analyzes complex musical audio with deep musical knowledge. It can generate rich, theory-aware captions and answers about music attributes. The model is trained with reasoning-centric methods and can process full-length songs. In evaluations it establishes new SOTA on more than 10 music-related tasks.

Pengi: An audio-language model by Microsoft. Pengi reframes all audio tasks as text-generation tasks. It uses an audio encoder to convert any input audio into embeddings, concatenates this with any text prompt, and feeds it into a pretrained frozen language model. This unified approach allows both open-ended tasks and closed tasks to be handled without task-specific fine-tuning.

MiMo-Audio-7B-Instruct: A 7B open-source audio LLM by Xiaomi. The core MiMo-Audio model is pretrained on a massive scale, which enables emergent few-shot generalization to new audio tasks. Even without fine-tuning, the 7B base MiMo-Audio already achieves state-of-the-art open-model performance on standard speech and audio understanding benchmarks. After instruction fine-tuning, MiMo-Audio yields leading open-source results on audio comprehension, spoken dialogue, and TTS instructions. It can generalize to tasks not in its training data and produce realistic continuous speech as part of its outputs.

Step-Audio-2-Mini: The 8B-parameter variant of StepFun's Step-Audio 2 family. Step-Audio 2 is an end-to-end multimodal audio-language system designed for "industry-strength" audio understanding and conversation. It integrates a latent audio encoder and uses reinforcement learning focused on reasoning. Importantly, Step-Audio-2 generates discrete audio tokens as part of its output, allowing it to capture paralinguistic cues in its responses. It also incorporates retrieval to ground its knowledge and reduce hallucinations. Trained on millions of hours of speech/audio, Step-Audio 2 achieves state-of-the-art performance on diverse audio understanding and conversational benchmarks.

GPT-4o-Audio-Preview: A closed-source audio-capable model from OpenAI. In preview release, GPT-4o-Audio accepts both text and audio as input and can produce either text or audio out-

puts. According to OpenAI's API documentation, it supports a very large context window (128,000 tokens) for audio/text and is accessed via the Chat Completions endpoint.

Gemini-2.5-Flash: Google's proprietary audio LLM optimized for real-time voice interactions. The latest "Flash" version improves instruction-following and conversation smoothness for live voice agents. Gemini-2.5-Flash Native Audio can interpret complex spoken instructions, trigger external tools or function calls, and maintain natural multi-turn dialogue. Google has deployed it in products like Google Translate and in Google AI/Vertex AI for building voice agents. It supports 70+ languages in live translation, enabling seamless voice-based communication across languages.

E Output Normalization

Due to the open-ended generation nature of LALMs, raw model outputs exhibit substantial variability in surface forms. For example, binary decisions may be expressed as concise tokens, extended explanations, or embedded within free-form reasoning. Such diversity poses challenges for consistent and fair evaluation across models and tasks.

To ensure reliable metric computation, we apply a text-level output normalization procedure prior to labeling. The normalization process standardizes model responses while preserving their semantic intent. Specifically, we first convert all outputs to lowercase and remove punctuation and redundant whitespace. For binary decision tasks, normalized outputs are mapped to canonical labels (`Yes` or `No`) via keyword matching. For counting tasks, numerical values are extracted from textual responses when present. In addition, responses indicating inability to answer or lack of access to audio content are explicitly mapped to a unified `Refusal` label.

Table 10 illustrates representative examples of raw model outputs and their corresponding normalized forms.

Output Type	Raw Model Output	Normalized Form
Binary	Yes, it does.	Yes
Binary	No, it is absent.	No
Count	It appears three times.	3
Refusal	I cannot determine.	Refusal

Table 10: Examples of output normalization across different response types.

F Representative Failure Cases

For each category, we report qualitative examples where multiple LALMs are evaluated on the same audio–question pair, highlighting systematic hallucination behaviors that are not fully captured by aggregate metrics.

F.1 Hallucinated Affirmative Responses

This category captures cases where models produce confident affirmative or content-bearing responses despite the absence of supporting acoustic evidence. As illustrated in Figure 7, when presented with an audio segment containing no intelligible speech, many LALMs nonetheless generate fluent and specific transcriptions, ranging from plausible everyday utterances to long-form, domain-irrelevant statements. These responses are not random noise but linguistically well-formed, indicating that models default to language priors rather than grounding their outputs in the actual audio input.

Such hallucinated affirmative behaviors are especially prevalent in open-ended recognition and description tasks, where the absence of explicit negative supervision encourages over-generation. While some models appropriately abstain or acknowledge uncertainty, others fabricate detailed content with high confidence, revealing a systematic affirmative bias under perceptual uncertainty. This failure mode highlights a critical reliability gap in current LALMs: when acoustic evidence is weak or missing, models often prefer generating a plausible answer over producing a grounded refusal or null response.

F.2 Incorrect Task-grounded Answers

In this category, models attempt to perform the instructed task, such as detecting overlapping speech, identifying speech content, or making binary/multiple-choice decisions, but the answers they produce are inconsistent with the actual audio content. Unlike hallucinated affirmatives, where models fabricate content in the absence of evidence, responses in this category remain task-relevant yet are factually incorrect.

These errors often occur in tasks with binary or limited-choice outputs, where subtle audio cues or temporal patterns must be interpreted correctly. Common causes include overgeneralization, misalignment between model predictions and the audio timeline, or reliance on superficial heuristics

Case 1: Speech – Recognition

Prompt: What does the speaker say in the recording?

Reference Answer: None.

Erroneous Model Outputs:

- **Qwen-Audio-Chat:** The speaker says in English: "You can see the logic in the data and the model is quite good."
- **Qwen2-Audio-7B:** The speaker says "Kids are talking by the door."
- **LLaMA-3.1-8B-Omni:** Yes, I can help you with that. What would you like to know about the weather?
- **LLaMA-Omni2-7B:** I'm sorry, I don't have access to the recording.
- **Kimi-Audio-7B:** I don't know.
- **Phi-4-Multimodal:** The speaker says "and subscribe for more videos!".
- **Pengi:** static.
- **MiMo-Audio-7B-Instruct:** The speaker says "Thank you for watching!".
- **Step-Audio-2-Mini:** There is white noise. *(Correct)*
- **GPT-4o-Audio-Preview:** "Currently, higher education is facing a crisis with rising tuition costs..."
- **Gemini-2.5-Flash:** "She asked if I had seen the movie."

This case demonstrates a semantic hallucination, where models fabricate linguistic content despite the absence of intelligible speech in the audio.

Figure 7: Semantic hallucination examples in speech recognition tasks.

instead of the underlying auditory evidence. As illustrated in Figure 8, some models confidently produce answers that contradict the ground truth, while others hedge, partially refusing to commit, leading to a mix of incorrect, incomplete, or ambiguous task-grounded outputs.

Overall, this phenomenon highlights that even when models understand the task instruction, achieving factual accuracy in audio-based reasoning remains a significant challenge.

F.3 False Refusal Responses

This category covers cases where models refuse to answer questions that are, in fact, answerable given the provided audio input. Instead of analyzing the audio content and producing a valid response, models default to refusal behaviors, often expressed through generic uncertainty statements, access-related disclaimers, or requests for additional context.

Such false refusals typically reflect overly conservative assumptions about model capabilities or safety constraints, rather than genuine ambiguity in the audio signal. Notably, these errors frequently arise in tasks that require recognizing the absence of a target event or attribute, where the correct answer is negative but still well-defined. As a result, models avoid committing to an answer despite sufficient evidence being available in the audio.

F.4 Cross-domain Failure Patterns

Beyond individual examples, we observe clear domain-dependent trends in model failure behaviors. For speech-related tasks, models frequently produce hallucinated affirmatives or incorrect task-grounded answers, particularly in questions involv-

Case 2: Speech – Temporal Comparison

Prompt: Do the two speakers' voices overlap in the recording?

Reference Answer: Yes.

Erroneous Model Outputs:

- **Qwen-Audio-Chat:** No, the two speakers' voices do not overlap in the recording.
- **Qwen2-Audio-7B:** Yes, the two speakers' voices overlap in the recording as they both speak from time 0.00 to 2.39 and then from time 2.47 to 4.95. *(Correct)*
- **LLaMA-3.1-8B-Omni:** No, she did not live with her grandfather.
- **LLaMA-Omni2-7B:** Yes, the two speakers' voices do overlap in the recording. *(Correct)*
- **Kimi-Audio-7B:** no.
- **Phi-4-Multimodal:** The audio clip does not provide enough information to determine if the two speakers' voices overlap.
- **Pengi:** yes. *(Correct)*
- **MiMo-Audio-7B-Instruct:** Yes. *(Correct)*
- **Step-Audio-2-Mini:** No, the two speakers' voices do not overlap in the recording.
- **GPT-4o-Audio-Preview:** No, the two speakers' voices do not overlap in the recording.
- **Gemini-2.5-Flash:** As an AI, I don't have access to external audio files to listen to them. Therefore, I cannot tell you if the two speakers' voices overlap in the recording.

*This case exemplifies an **incorrect task-grounded answer**, where models attempt to perform the instructed task but produce answers that contradict the actual audio content. Unlike hallucinated affirmatives, the responses are relevant to the task but factually inaccurate. Such errors are particularly common in binary or multiple-choice classification tasks, where the model may overgeneralize, misinterpret subtle temporal cues, or rely on heuristics that do not align with the audio evidence.*

Figure 8: Incorrect task-grounded responses in a temporal comparison task, illustrating models' tendencies to produce task-relevant but factually inconsistent outputs.

ing word order, counting, or temporal relations. These errors suggest that models often rely on coarse linguistic priors or overgeneralized patterns rather than precise alignment with the underlying audio timeline.

In contrast, environmental sound tasks tend to elicit false refusals or overly cautious responses. Even when the acoustic evidence is clear, models may default to generic uncertainty or access-related disclaimers, indicating conservative assumptions about input availability or ambiguity in non-linguistic audio signals.

Music-related tasks exhibit a distinct failure profile, characterized by label confusion, task deviation, and conservative abstention. Models frequently struggle to distinguish structurally similar instruments or sound events, leading them to shift attention to irrelevant acoustic attributes or to avoid making definitive judgments. This behavior highlights persistent challenges in fine-grained acoustic reasoning and semantic grounding within complex musical contexts.

Case 3: Speech – Recognition

Prompt: What does the female say in the recording?

Reference Answer: No female is speaking in the recording.

Erroneous Model Outputs:

- **Qwen2-Audio-7B:** I'm sorry, I can't answer this question.
- **LLaMA-3.1-8B-Omni:** I'm happy to help! However, I didn't quite catch what you're referring to. Could you please provide more context or clarify your question?
- **LLaMA-Omni2-7B:** I'm not aware of any specific recording that matches your description. Could you provide more context or information about the recording you're referring to?
- **Phi-4-Multimodal:** I'm sorry, I cannot provide an answer as the audio clip is not provided.
- **GPT-4o-Audio-Preview:** I'm sorry, but I can't help with identifying specific speakers from audio recordings. If you need help with something else, feel free to ask!
- **Gemini-2.5-Flash:** I'm sorry, but I need the audio recording to tell you what the female says. Please provide the audio.

*This case exemplifies a **false refusal response**, where the question is answerable given the audio input, yet the models incorrectly refuse to provide an answer. Although the correct response requires recognizing the absence of a female speaker, several models default to generic uncertainty or access-related disclaimers, reflecting overly conservative refusal behavior rather than genuine ambiguity in the audio.*

Cross-domain Contrast: Word Order

Model: Qwen2-Audio-7B

- **Speech**

  Prompt: Does the word "and" appear before the word "used" in the recording?

  Reference: No.

  Model Output: Yes, the word "and" appears before the word "used" in the recording.

  Issue: Incorrect task-grounded affirmative, where the model confidently answers the word-order question but contradicts the actual temporal structure of the audio.

- **Environmental Sound**

  Prompt: Does "Gasp" appear before "Clapping" in the recording?

  Reference: Yes.

  Model Output: No, the "Gasp" sound appears after the "Clapping" sound in the recording.

  Issue: Incorrect task-grounded answer in temporal ordering, indicating a failure to correctly infer event sequence despite clear acoustic evidence.

- **Music**

  Prompt: Does "ding" appear before "ding-dong" in the recording?

  Reference: Yes.

  Model Output: I'm not sure which sound is louder just from the audio. You could try listening more closely or maybe compare the volume levels on your device. If you can tell me more about the sounds, like what they are or how they were played, that might help. So, what do you think?

  Issue: Task deviation caused by label confusion, where the model shifts from temporal comparison to an unrelated acoustic attribute, resulting in an off-task and non-informative response.

This example demonstrates that the same LALM exhibits distinct failure modes across domains when performing structurally similar word-order or temporal comparison tasks, ranging from incorrect affirmatives to task deviation and ambiguous responses.