

Pardon? Evaluating Conversational Repair in Large Audio-Language Models

Shuanghong Huang^{1*}, Jinlei Xu^{1*}, Youchao Zhou¹,
Yanghao Zhou¹, Xuan Zhao¹, Chong Feng^{1†}, Wenxuan Zhang^{2†}

¹Beijing Institute of Technology, ²Singapore University of Technology and Design

{shuanghong, xujinlei, fengchong}@bit.edu.cn, wxzhang@sutd.edu.sg

Abstract

Large Audio-Language Models (LALMs) have demonstrated strong performance in spoken question answering (QA), with existing evaluations primarily focusing on answer accuracy and robustness to acoustic perturbations. However, such evaluations implicitly assume that spoken inputs remain semantically answerable, an assumption that often fails in real-world interaction when essential information is missing. In this work, we introduce a repair-aware evaluation setting that explicitly distinguishes between answerable and unanswerable audio inputs. We define answerability as a property of the input itself and construct paired evaluation conditions using a semantic-acoustic masking protocol. Based on this setting, we propose the Evaluability Awareness and Repair (EAR) score, a non-compensatory metric that jointly evaluates task competence under answerable conditions and repair behavior under unanswerable conditions. Experiments on two spoken QA benchmarks across diverse LALMs reveal a consistent gap between answer accuracy and conversational reliability: while many models perform well when inputs are answerable, most fail to recognize semantic unanswerability and initiate appropriate conversational repair. These findings expose a limitation of prevailing accuracy-centric evaluation practices and motivate reliability assessments that treat unanswerable inputs as cues for repair and continued interaction. The core code and dataset are publicly available at <https://github.com/sheunghung/EAR>.

1 Introduction

Large Audio-Language Models (Cui et al., 2025; Arora et al., 2025) (LALMs) have rapidly evolved from passive speech recognition systems into interactive conversational agents capable of reasoning and decision-making directly from spoken in-

put (Tang et al., 2024; Xu et al., 2025). Recent models such as GPT-4o (Hurst et al., 2024), Gemini 2.5 (Comanici et al., 2025) and DeSTA2.5-Audio (Lu et al., 2026) demonstrate strong performance on spoken question answering (Gong et al., 2024) (QA) and related tasks, enabling a wide range of applications including virtual assistants (Anastassiou et al., 2024), real-time translation (Barrault et al., 2023), and multimodal dialogue systems (Fang et al., 2025). As LALMs increasingly operate in open-world conversational settings, their reliability hinges not only on recognizing speech accurately but also on responding appropriately to varying input conditions.

As illustrated in Figure 1, typical evaluations of LALMs primarily assess performance through answer accuracy under clean speech inputs, focusing on whether models can produce the correct response to a given query (Lipping et al., 2022; Yang et al., 2024a; Sakshi et al., 2025). To further test their reliability, more recent benchmarks extend evaluation to robustness under adverse acoustic conditions, such as noise, disfluency, or signal corruption (Ma et al., 2025b; Liu et al., 2025). In these settings, degraded audio is treated as a more challenging but still semantically answerable variant of the original input, and model outputs are evaluated against the fixed ground-truth answer.

This robustness-centered paradigm has been effective for measuring speech understanding under adverse conditions. However, it implicitly assumes a crucial premise: that the input remains semantically *answerable*. In real-world spoken communication, this assumption often breaks down. Audio inputs may become incomplete due to noise, transmission loss, or overlapping speech, leading to the loss of answer-critical semantic information. In such cases, the problem extends beyond reduced accuracy, as the input becomes intrinsically *unanswerable* even for a human listener. Rather than guessing or disengaging, human interlocutors re-

*Equal contribution.

†Corresponding author.

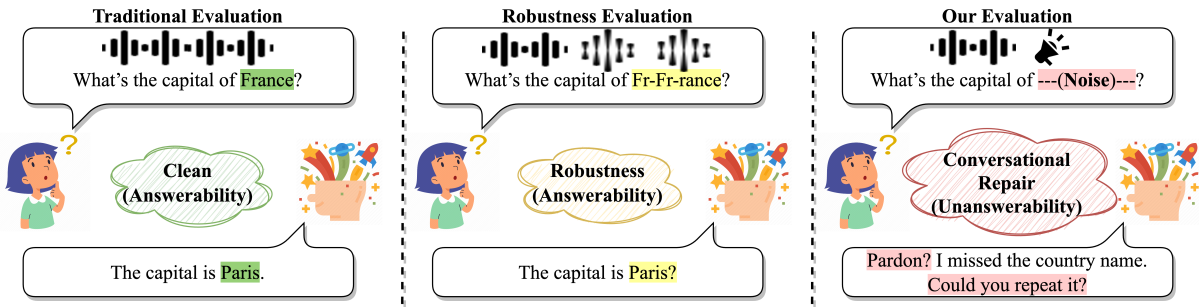


Figure 1: **Left:** Traditional evaluation measures answer correctness under clean, answerable inputs. **Middle:** Robustness evaluation tests whether correct answers are maintained under acoustic perturbations that preserve answerability. **Right:** Our repair-aware evaluation masks answer-critical information to create unanswerable inputs and assesses whether models shift from answering to conversational repair.

spond through conversational repair, signaling the breakdown (e.g., “Pardon?”) and requesting clarification to address the issue. Reliable spoken interaction, therefore, depends on the ability to recognize when semantic evaluability is lost and to adapt conversational behavior accordingly.

However, prevailing evaluations for LALMs remain answer-centric, treating semantically incomplete inputs as if they required a definitive answer. Under this framing, models are encouraged to hallucinate plausible-looking responses or issue generic refusals, rather than adapting their behavior through conversational repair. Such outputs obscure whether the model has recognized the loss of semantic evaluability and fail to support conversational recovery.

We study repair behavior in controlled single-turn spoken QA settings, treating single-turn repair awareness as a foundational prerequisite for broader multi-turn conversational repair. In this work, we introduce a repair-aware evaluation setting that makes answerability an explicit part of the evaluation design. We frame answerability as a property of the audio input itself and separate evaluation into answerable and unanswerable conditions. Unanswerable inputs are defined as those in which answer-critical semantic information is absent, such that even a human listener cannot determine the correct answer.

Under this setting, reliable conversational behavior requires models to condition their responses on semantic evaluability: producing task-fulfilling answers when inputs are answerable, and initiating appropriate conversational repair when they are not. To operationalize this distinction, we propose a semantic-acoustic masking protocol that selectively removes answer-critical semantic con-

tent, yielding paired answerable and unanswerable inputs derived from the same underlying query. Building on this controlled setup, we introduce the Evaluability Awareness and Repair (EAR) score, a non-compensatory metric that jointly evaluates task competence under answerable conditions and repair behavior under unanswerable conditions.

Experiments on two spoken QA benchmarks across a diverse set of LALMs reveal a consistent gap between answer accuracy and conversational reliability. While many models achieve high accuracy when inputs are answerable, most fail to recognize when essential semantic information is missing and do not initiate conversational repair. These results indicate that strong robustness or accuracy alone does not imply evaluability awareness, and that prevailing evaluation practices substantially overestimate real-world conversational reliability.

Our contributions are summarized as follows:

- We identify that existing LALM evaluations assess answer accuracy and robustness, but omit conversational repair behavior.
- We introduce a repair-aware evaluation setting that explicitly distinguishes answerable and unanswerable spoken inputs.
- We propose a semantic-acoustic masking protocol and the EAR score to jointly evaluate task competence and conversational repair without mutual compensation.

2 Related Work

2.1 Large Audio-Language Models

LALMs extend multimodal large language models (LLMs) to spoken inputs, enabling instruction following and language-style reasoning grounded in audio. Compared with text, audio signals exhibit greater heterogeneity across speech, music, and

environmental sounds, involving diverse temporal structures and acoustic patterns. To accommodate such variability, LALMs are designed to support a unified spoken interface that can flexibly process a wide range of audio inputs within a single model, rather than relying on task-specific audio-language pipelines (Cui et al., 2025; Arora et al., 2025).

Most LALMs adopt a modular architecture that integrates a pretrained audio encoder with an LLM backbone, allowing auditory representations to be aligned with language generation and reasoning. Recent work has increasingly focused on *general-purpose* LALMs that support audio instruction following and broad auditory reasoning across domains, moving beyond narrowly scoped tasks such as speech recognition or audio classification. Representative systems including GAMA (Ghosh et al., 2024), SALMONN (Tang et al., 2024), Qwen2-Audio (Chu et al., 2024), and DeSTA2.5-Audio (Lu et al., 2026) demonstrate steady progress toward more versatile spoken interaction capabilities within a unified framework.

2.2 Evaluating Reliability and Repair

Evaluating the reliability of language models has received increasing attention in recent years, although a unified standard remains elusive. Prior work has proposed a range of metrics and datasets to encourage models to abstain when reliable answers cannot be produced, capturing notions such as prudence, honesty, and truthfulness (Yang et al., 2024b; Cheng et al., 2024), as well as faithfulness- and precision-based measures (Yona et al., 2024; Zhang et al., 2024). More recent approaches further integrate answer accuracy and refusal behavior into unified reliability scores by explicitly modeling their trade-offs (Xu et al., 2024). Despite these advances, such reliability evaluations predominantly operate on static textual inputs and treat reliability as a property of isolated input-output pairs.

In spoken and conversational settings, reliability manifests differently from text-based scenarios. Rather than reflecting knowledge uncertainty, failures often arise from audio-level degradation, where answer-critical semantic information is partially or entirely missing from the speech signal. Current benchmarks for LALMs primarily emphasize robustness, assessing whether task performance can be maintained under noisy or disfluent speech conditions (Ma et al., 2025a; Chen et al., 2024). Extensions such as VOCALBENCHDF (Liu et al., 2025) further analyze model behav-

ior under diverse speech disfluencies, but generally assume that sufficient answer-critical semantic information is preserved in the audio, focusing on robustness rather than semantic unanswerability caused by audio-level information loss.

In contrast to prior work, our evaluation distinguishes between answerable and unanswerable conversational states derived from the same underlying query. Moreover, rather than treating refusal as the sole reliable fallback, we conceptualize conversational repair as a constructive and measurable behavior, evaluating whether models appropriately shift from answering to clarification when semantic evaluability is lost. Our framing of conversational repair draws on the discourse-theoretic notion of grounding (Clark and Brennan, 1991), in which interlocutors collaboratively establish mutual understanding through clarification and repair.

3 Methodology

To evaluate conversational reliability under varying semantic conditions, we propose a repair-aware evaluation framework, illustrated in Figure 2. The core idea is to explicitly control the answerability of spoken inputs while maintaining the underlying query and target answer. Starting from an original audio, we generate semantically distinct audio variants through the semantic-acoustic masking protocol, and assess whether models appropriately condition their behavior on input evaluability, answering when sufficient information is available and initiating conversational repair when it is not.

3.1 Problem Definition

We study the evaluation of LALMs in spoken QA, focusing on whether models can adapt their conversational behavior to the semantic evaluability of spoken inputs. Let q denote an underlying semantic query, associated with a correct answer y . The query q is modality-independent and specifies the information required at the semantic level. Given a spoken QA instance corresponding to q , we construct multiple audio realizations x by applying controlled acoustic masking, resulting in inputs that differ in semantic completeness while sharing the same underlying query. We define an answerability function:

$$A(x) \in \{0, 1\}, \quad (1)$$

where $A(x) = 1$ indicates that the audio input x contains sufficient semantic information for a hu-

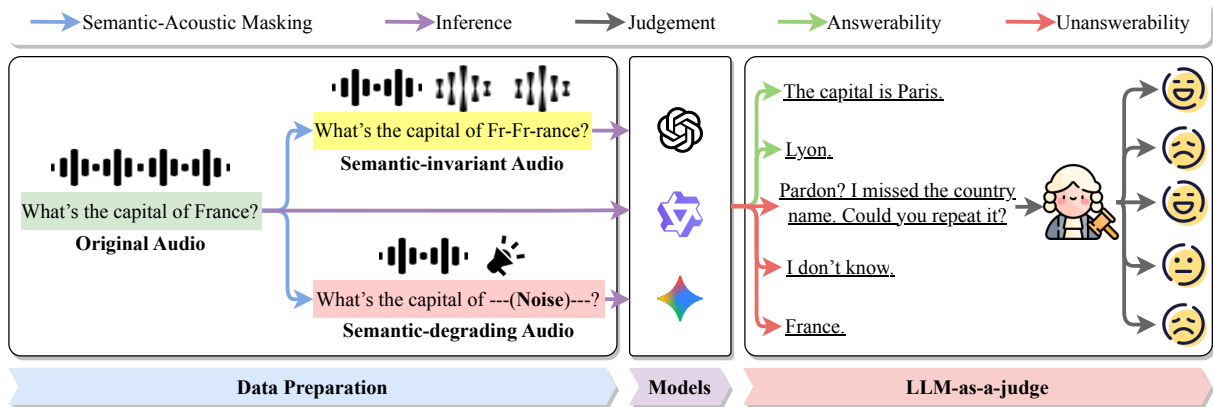


Figure 2: Overview of the repair-aware evaluation framework. The original audio generates two variants using the semantic-acoustic masking protocol: an answerable (semantic-invariant) and an unanswerable (semantic-degrading). The same set of LLMs then processes the original audio and its two masked variants. Finally, the LLM-as-a-judge evaluates these responses for answer correctness and conversational repair, respectively, under answerable and unanswerable inputs.

man listener to determine y , and $A(x) = 0$ indicates that essential semantic information is missing. Answerability is thus defined as a property of the audio input itself, independent of model behavior.

For each semantic query q , we construct two evaluation conditions: an answerable audio input x^a with $A(x^a) = 1$, and an unanswerable audio input x^u with $A(x^u) = 0$. Both are derived from the same underlying query q and are associated with the same target answer y . This paired design isolates the effect of semantic evaluability while controlling for task definition and label identity.

Given a model f , reliable conversational behavior is defined functionally. Under the answerable condition x^a , the model should produce a task-fulfilling answer consistent with y . Under the unanswerable condition x^u , behavior instead requires conversational repair, where the model explicitly recognizes missing information and requests clarification, rather than attempting to answer.

3.2 Semantic-Acoustic Masking Protocol

To operationalize the answerability function $A(x)$ over audio inputs, we introduce a semantic-acoustic masking protocol that constructs paired answerable and unanswerable audio variants by masking answer-critical or non-critical semantic content.

At a conceptual level, we define a semantic unit as answer-critical if its removal renders the spoken input underspecified for a human listener, such that the question can no longer be reliably answered. In practice, identifying minimal answer-critical units in arbitrary speech is challenging. To enable controlled and reproducible masking, we adopt a

simplified operational strategy that leverages the availability of ground-truth answers and automatic speech processing tools. Across all conditions, masking is applied at the acoustic level by aligning text tokens to word-level time spans and replacing the corresponding segments with controlled signals (e.g., silence or white noise), while preserving the overall temporal structure of the utterance.

Semantic-degrading masking (unanswerable condition). For constructing semantically unanswerable inputs, we directly mask the audio segments corresponding to the ground-truth answer. Specifically, given an audio instance with a known answer span, we align the answer text to the audio and replace the corresponding time segments with controlled acoustic signals. This procedure ensures that answer-critical semantic information is absent from the audio, making the input intrinsically unanswerable even for a human listener ($A(x) = 0$).

Semantic-invariant masking (answerable condition). For constructing answerable but acoustically perturbed inputs, we aim to introduce surface-level acoustic variation without affecting semantic evaluability. We first obtain an automatic transcription of the audio using Whisper-large-v3¹ (Radford et al., 2023) and perform part-of-speech tagging on the transcript using spaCy². We then randomly select a single token from a predefined set of function-word categories (e.g., determiners, adpositions, auxiliaries, conjunctions, pronouns, and par-

¹<https://huggingface.co/openai/whisper-large-v3>

²<https://github.com/explosion/spaCy>

Dataset	Avg. Utterance Duration	Avg. Masked Duration	Avg. Masked Ratio
WDYL	8.47 s	0.57 s	6.7%
SLUE-SQA-5	39.78 s	1.00 s	2.5%

Table 1: Dataset statistics for evaluation triplets constructed via semantic-acoustic masking. Masked spans constitute only a small proportion of total utterance duration.

titles) and replace its corresponding audio segment with controlled acoustic signals. Because such function words do not carry answer-critical semantic information, this masking preserves answerability while altering the acoustic surface ($A(x) = 1$).

3.3 Evaluation Triplets and Metrics

From each underlying semantic query q , we construct an evaluation triplet (x^a, x^u, y) . The answerable condition x^a includes multiple semantically equivalent realizations, specifically the original audio and its semantic-invariant masked variant. The unanswerable condition x^u is instantiated via semantic-degrading masking.

Task Competence under Answerable Conditions. Task competence is evaluated only when $A(x) = 1$. We define:

$$C(x^a) = \begin{cases} 1, & \text{if answers correctly,} \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

In practice, $C(x^a)$ is computed as the average accuracy over all answerable realizations of x^a , including both the original audio and the semantic-invariant masked input. This captures whether the model can reliably perform the task when sufficient semantic information is available, despite surface-level acoustic variation.

Repair Behavior under Unanswerable Conditions. Under the unanswerable condition x^u , correctness is ill-defined. We assess the model’s conversational repair behavior via a graded score:

$$R(x^u) = \begin{cases} 1, & \text{explicit repair,} \\ 0.5, & \text{generic refusal,} \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

Explicit repair refers to responses that identify missing information and request targeted clarification. Generic refusals (e.g., “I don’t know”) receive partial credit, while hallucinated or incorrect answers receive none.

Evaluability Awareness and Repair Score To jointly evaluate conditional conversational reliability, we define the EAR score as a dataset-level metric. Let C denote the average task competence over all answerable audio inputs, and let R denote the average repair behavior score over all unanswerable inputs across the dataset. The EAR score is defined as the harmonic mean of C and R :

$$\text{EAR} = \frac{2 \cdot C \cdot R}{C + R}. \quad (4)$$

This non-compensatory formulation ensures that high scores are achieved only when a model both answers correctly under answerable conditions and performs repair under unanswerable conditions.

Repair Behavior Identification Repair behavior is identified using an LLM-as-a-Judge paradigm (Li et al., 2025). Given the query, the masked input context, and the model response, a judge model is instructed to categorize the response into explicit repair, generic refusal, or other. The judge produces categorical decisions only, reducing subjectivity and improving reproducibility. This procedure is used solely for evaluation and does not provide supervision to the evaluated models.

4 Experimental Setup

4.1 Datasets

We evaluate our framework on two spoken QA benchmarks with complementary conversational and acoustic characteristics, enabling assessment of repair-aware behavior under both interaction-driven and acoustically challenging conditions.

What Do You Like? (WDYL). WDYL (Wu et al., 2024) contains 1,000 spoken questions paired with audio recordings and ground-truth answers. The dataset features conversational, interaction-oriented queries that are often context-dependent, clarifying a natural and appropriate response when critical information is missing.

SLUE-SQA-5. SLUE-SQA-5 (Shon et al., 2023) is a large-scale spoken QA benchmark with longer

Model	WDYL			SLUE-SQA-5		
	$C \uparrow$	$R \uparrow$	EAR \uparrow	$C \uparrow$	$R \uparrow$	EAR \uparrow
Qwen2-Audio (Chu et al., 2024)	5.5	37.7	9.6	29.1	10.1	15.0
Baichuan-Omni (Li et al., 2024)	70.3	3.1	5.0	51.5	3.0	5.7
Qwen2.5-Omni (Xu et al., 2025)	77.3	1.0	2.0	53.5	5.4	9.8
DeSTA2.5-Audio (Lu et al., 2026)	70.8	14.2	<u>23.7</u>	39.5	27.7	<u>32.6</u>
Audio Flamingo 3 (Ghosh et al., 2025)	82.2	6.2	11.5	40.6	11.4	17.8
GPT-4o (Hurst et al., 2024)	91.5	25.4	39.8	43.8	36.2	39.6
Gemini 2.5 (Comanici et al., 2025)	99.5	63.0	77.2	65.7	11.2	19.1

Table 2: Repair-aware evaluation results across two spoken QA datasets. We report task competence under answerable inputs (C), repair behavior under unanswerable inputs (R), and the resulting EAR score. **Bold** denotes the best overall performance, and underlined indicates the best open-source model.

utterances and more challenging acoustic conditions. We randomly sample 1,000 instances from the original dataset for evaluation. Compared to WDYL, answers in SLUE-SQA-5 are typically distributed over extended audio segments, increasing the impact of masking answer-critical content and providing a complementary testbed for repair-aware evaluation.

Table 1 summarizes the statistics of the constructed evaluation triplets. Masked spans account for only 2.5-6.7% of total utterance duration, confirming that EAR evaluates repair under localized semantic loss rather than complete signal removal. The short utterances in WDYL (avg. 8.47 s) and the longer segments in SLUE-SQA-5 (avg. 39.78 s) provide complementary testing conditions.

For each dataset, we construct evaluation triplets (x^a, x^u, y) using the semantic-acoustic masking protocol. Answerable inputs x^a retain sufficient information for answering, while unanswerable inputs x^u are rendered underspecified by masking answer-critical segments, as verified by human listeners.

4.2 Models for Evaluation

We evaluate a total of seven LALMs, covering both open-source and closed-source systems, to assess whether repair-aware conversational behavior is consistently exhibited across model families. The evaluated models include the open-source Qwen2-Audio (Chu et al., 2024), Baichuan-Omni (Li et al., 2024), Qwen2.5-Omni (Xu et al., 2025), DeSTA2.5-Audio (Lu et al., 2026), and Audio Flamingo 3 (Ghosh et al., 2025), as well as the closed-source GPT-4o (Hurst et al., 2024) and Gemini 2.5 (Comanici et al., 2025).

Open-source models are evaluated using their publicly released checkpoints and default infer-

ence configurations. Closed-source models are accessed through official APIs following recommended usage guidelines. All models are evaluated in a zero-shot setting, without task-specific fine-tuning, which reflects deployment scenarios and avoids confounding effects from supervision.

4.3 Inference and Evaluation

All experiments are conducted on NVIDIA RTX 3090 and A6000 GPUs. For open-source models, inference is performed using standard decoding settings commonly adopted in prior work, including greedy decoding or low-temperature sampling. Specifically, we use a temperature of 0 or 0.1 (when sampling is enabled), and set the maximum generation length to 2048 tokens, which is sufficient to cover both task answers and clarification-oriented repair responses. Closed-source models are queried using the default decoding parameters provided by their respective APIs.

For each evaluation triplet (x^a, x^u, y) , each model is queried independently under the answerable and unanswerable conditions, producing responses r^a and r^u , respectively. Answer correctness under the answerable condition is determined by exact match or semantic equivalence with the ground-truth answer y , following standard practices in spoken QA. Repair behavior under the unanswerable condition is identified using an LLM-as-a-judge protocol instantiated with GPT-4o (gpt-4o-2024-11-20³) (Hurst et al., 2024). Given the original query, the masked input context, and the model response, GPT-4o is instructed to perform a categorical judgment of the response type (explicit repair, generic refusal, or otherwise).

We report the EAR score as the primary metric

³<https://platform.openai.com/docs/models/gpt-4o>

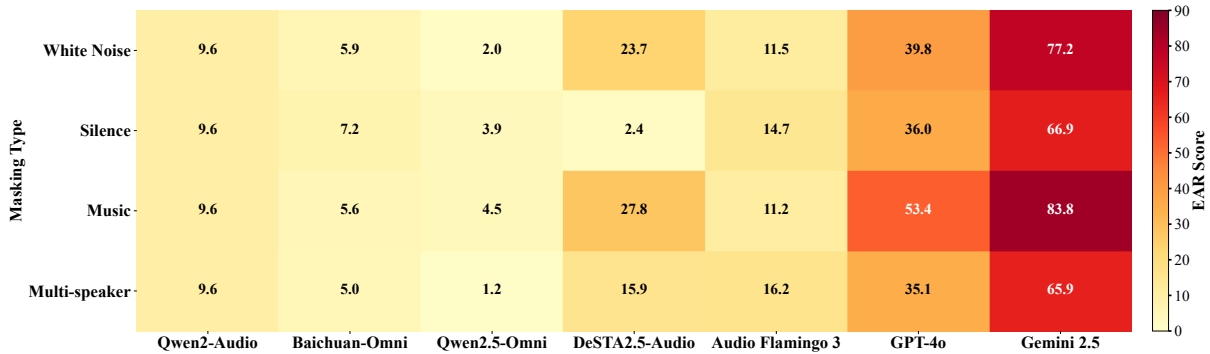


Figure 3: Sensitivity of EAR to different semantic-degrading masking realizations on the WDYL dataset. EAR scores are reported under four masking types: white noise, silence, music, and multi-speaker. While absolute values vary, relative model ordering is preserved across masking realizations.

for conversational reliability. To make the computation of EAR transparent, we also report its constituent dataset-level components: task competence under answerable conditions, measured as the average answer accuracy over all answerable inputs (which defines C), and repair performance under semantically degrading perturbations, measured as the average repair behavior score over all unanswerable inputs (which defines R). All evaluations are conducted using a unified pipeline to ensure consistency across models and datasets.

The prompt templates used for model inference and evaluation are provided in Appendix A.

5 Experimental Results and Analysis

5.1 Main Results

Table 2 summarizes the repair-aware evaluation results across two spoken QA datasets. We report task competence under answerable conditions (C), repair behavior under unanswerable conditions (R), and the resulting EAR score, which jointly reflect conditional conversational reliability.

Finding 1: High task competence does not imply reliable conversational behavior. Across both datasets, many models achieve strong competence under answerable inputs, with C often exceeding 70% on WDYL and remaining moderately high on SLUE-SQA-5. However, this competence does not consistently translate into high EAR scores. The primary reason is limited repair behavior: models that answer accurately under answerable conditions frequently fail to initiate repair once answer-critical information is removed. This pattern is especially evident for models such as Baichuan-Omni and Qwen2.5-Omni, which exhibit high C but consistently low R (typically remaining in the low single-

digit range), resulting in substantially reduced EAR. These results indicate that accuracy-oriented evaluation can substantially overestimate model reliability by ignoring behavior under semantic failure.

Finding 2: Repair, rather than robustness, is the dominant bottleneck.

Many models maintain moderately high C across datasets, suggesting reasonable robustness to content-preserving perturbations. In contrast, R varies dramatically and is the primary driver of EAR differences. Models that demonstrate explicit repair behavior, such as DeSTA2.5-Audio and Audio Flamingo 3, achieve substantially higher EAR despite having similar or even lower C values than other models. This separation confirms that repair awareness captures a distinct behavioral capability that cannot be explained by robustness or answer accuracy alone.

Finding 3: Recognizing semantic unanswerability remains challenging.

Repair-aware reliability differs markedly between datasets. Compared to SLUE-SQA-5, models exhibit higher R and EAR scores on WDYL, which features shorter and more interaction-oriented queries. In contrast, R on SLUE-SQA-5 remain low even for strong closed-source models such as Gemini 2.5, indicating that longer and acoustically complex utterances make semantic unanswerability harder to recognize.

Overall, these results show that strong task competence or robustness alone does not guarantee appropriate model behavior when essential semantic information is missing. By explicitly evaluating both answering and repair behavior, EAR exposes systematic failure modes that remain hidden under conventional accuracy-centric evaluation.

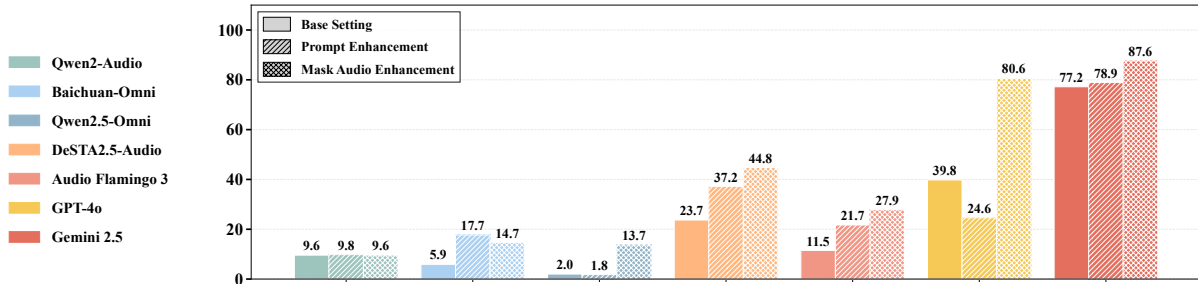


Figure 4: Effects of prompting and mask severity on EAR scores on the WDYL dataset. We compare the base setting with a transcription-based prompt enhancement and an enhanced masking setting that expands the masked time window around the answer span. Increasing mask severity consistently improves EAR across most models, while prompt enhancement yields limited and model-dependent effects.

5.2 Sensitivity Analysis of EAR Score

Figure 3 evaluates the sensitivity of EAR to different semantic-degrading masking realizations on the WDYL dataset, including white noise, silence replacement, background music, and multi-speaker overlap. Across models, absolute EAR values can vary across masking types, particularly for models with stronger repair capabilities.

Despite this variation, a consistent pattern emerges in terms of relative model ordering. Models that exhibit stronger repair-aware behavior consistently achieve higher EAR scores across all masking realizations. In contrast, models with limited repair behavior remain low-performing regardless of the specific noise type. This trend holds for both open-source and closed-source models, including GPT-4o and Gemini 2.5, whose relative ranking is preserved despite differences in absolute EAR values. These results indicate that EAR provides stable comparative judgments of conversational reliability rather than being driven by a particular noise implementation.

In addition to ranking stability, the results also reveal model-specific sensitivity to different masking realizations. Some models show pronounced variation across noise types, whereas others remain relatively insensitive. These sensitivity patterns are not consistent across models: the same masking realization may substantially affect one model while having a limited impact on another. This heterogeneity suggests that different models rely on distinct acoustic cues to detect semantic unanswerability, resulting in diverse responses under various forms of audio-level semantic degradation.

5.3 Effects of Prompting and Mask Severity on Repair Behavior

Figure 4 compares EAR scores under three settings on the WDYL dataset: a *base* setting used in the main experiments, a *prompt enhancement* setting, and a *mask audio enhancement* setting. In the base setting, we apply semantic-degrading masking by replacing the minimally aligned keyword span with white noise. The prompt enhancement setting employs a training-free transcribe-then-answer strategy, where the model is first instructed to produce a transcription and then answer the question based on that transcription. In contrast, the mask audio enhancement setting increases masking severity by expanding the masked temporal window centered on the keyword span, rather than masking only the minimally aligned segment. This makes answer-critical information more decisively absent while preserving the surrounding context.

We find that prompt enhancement yields limited and model-dependent changes in EAR. While some models (e.g., Baichuan-Omni and DeSTA2.5-Audio) show modest improvements when encouraged to reason over transcripts, several others exhibit negligible gains or even slight degradation. This variability suggests that prompting alone does not induce conversational repair behavior: even when models are guided to rely on textual transcripts, they may still attempt to answer under semantically unanswerable inputs, or fall back to generic refusals without explicitly signaling missing information. Therefore, the lack of repair behavior in the base setting cannot be attributed solely to insufficient task instruction or prompt design.

By contrast, mask audio enhancement leads to substantial improvements in EAR across most models. When the masked window around the keyword

span is expanded, semantic unanswerability becomes more salient and easier to detect, resulting in higher repair rates for many systems. This trend suggests that repair behavior in current LALMs is closely related to the detectability of semantic failure: models are more likely to initiate conversational repair when answer-critical cues are absent, but often struggle when semantic insufficiency is subtle and localized. Importantly, this improvement does not reflect better task understanding, but rather that models more clearly recognize that the answer-critical information is missing.

6 Conclusions

This work reframes the evaluation of LALMs from a repair-aware perspective, arguing that conversational reliability in spoken interaction cannot be captured by answer accuracy or robustness alone. Instead, reliable behavior requires models to adapt their responses based on whether sufficient semantic information is available, answering when possible and initiating appropriate repair when it is not. To this end, we introduce an evaluation framework that explicitly distinguishes answerable from unanswerable spoken inputs, along with the EAR score for assessing conditional behavioral adaptation. By grounding reliability in input evaluability rather than output correctness alone, our framework exposes failure modes that remain invisible under conventional evaluation paradigms. These findings highlight the importance of treating uncertainty and unanswerability as integral components of interaction in spoken language models, motivating the development and evaluation of models that are not only accurate but also appropriately conversationally aware and responsive.

Limitations

This work studies repair-aware conversational reliability in controlled, single-turn spoken QA settings, where model behavior is conditioned on the semantic evaluability of the input. We use the term "conversational repair" in the discourse-theoretic sense of breakdown detection and repair initiation (Clark and Brennan, 1991), and our scope focuses specifically on single-turn repair awareness as a foundational prerequisite, rather than full multi-turn repair dynamics. This formulation enables a clear and interpretable distinction between answering and repair behavior. However, our evaluation does not explicitly model multi-turn conversational dynam-

ics or long-term dialogue state tracking. In more realistic dialogue settings, repair may unfold over multiple turns and depend on accumulated conversational context, introducing additional challenges beyond the scope of this study.

In addition, our semantic-acoustic masking protocol induces human-consistent unanswerability by controlling the removal of answer-critical semantic information. Although our experiments suggest that EAR supports consistent comparative judgments across several controlled masking realizations, this perturbation-based design captures only a bounded subset of real-world failure modes. The framework does not exhaustively cover other forms of underspecification beyond semantic information loss, nor is EAR intended as a direct training objective. Extending repair-aware evaluation to multi-turn interaction settings and to broader sources of semantic degradation remains an important direction for future work. Additionally, future work could incorporate more realistic repair triggers, such as acoustic ambiguity (e.g., similar-sounding entities), overlapping speech, or linguistically underspecified queries, to improve ecological validity in deployment scenarios.

Acknowledgments

We sincerely thank the anonymous reviewers for their insightful comments and constructive suggestions. This work is supported by the Natural Science Foundation of Beijing Municipality (Grant No. L232119), the National Research Foundation Singapore under the AI Singapore Programme (AISG Award No. AISG3-RPGV-2025-016), and in part by the China Scholarship Council.

References

- Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, and 1 others. 2024. Seed-tts: A family of high-quality versatile speech generation models. *arXiv preprint arXiv:2406.02430*.
- Siddhant Arora, Kai-Wei Chang, Chung-Ming Chien, Yifan Peng, Haibin Wu, Yossi Adi, Emmanuel Dupoux, Hung yi Lee, Karen Livescu, and Shinji Watanabe. 2025. [On the landscape of spoken language models: A comprehensive survey](#). *Transactions on Machine Learning Research*.
- Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John

- Hoffman, and 1 others. 2023. Seamless4t: massively multilingual & multimodal machine translation. *arXiv preprint arXiv:2308.11596*.
- Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T Tan, and Haizhou Li. 2024. Voicebench: Benchmarking llm-based voice assistants. *arXiv preprint arXiv:2410.17196*.
- Qinyuan Cheng, Tianxiang Sun, Xiangyang Liu, Wenwei Zhang, Zhangyue Yin, Shimin Li, Linyang Li, Zhengfu He, Kai Chen, and Xipeng Qiu. 2024. Can ai assistants know what they don't know? In *International Conference on Machine Learning*, pages 8184–8202. PMLR.
- 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*.
- Herbert H. Clark and Susan E. Brennan. 1991. Grounding in communication. In Lauren B. Resnick, John M. Levine, and Stephanie D. Teasley, editors, *Perspectives on socially shared cognition*, pages 127–149. American Psychological Association.
- 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*.
- Wenqian Cui, Dianzhi Yu, Xiaoqi Jiao, Ziqiao Meng, Guangyan Zhang, Qichao Wang, Steven Y Guo, and Irwin King. 2025. Recent advances in speech language models: A survey. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13943–13970.
- Qingkai Fang, Yan Zhou, Shoutao Guo, Shaolei Zhang, and Yang Feng. 2025. LLaMA-Omni 2: Llm-based real-time spoken chatbot with autoregressive streaming speech synthesis. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*.
- Sreyan Ghosh, Arushi Goel, 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](#). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- 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. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6288–6313.
- Yuan Gong, Hongyin Luo, Alexander H Liu, Leonid Karlinsky, and James Glass. 2024. Listen, think, and understand. In *International Conference on Learning Representations*.
- 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*.
- Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhat-tacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, and 1 others. 2025. From generation to judgment: Opportunities and challenges of llm-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791.
- Yadong Li, Haoze Sun, Mingan Lin, Tianpeng Li, Guosheng Dong, Tao Zhang, Bowen Ding, Wei Song, Zhenglin Cheng, Yuqi Huo, and 1 others. 2024. Baichuan-omni technical report. *arXiv preprint arXiv:2410.08565*.
- Samuel Lipping, Parthasaarathy Sudarsanam, Konstantinos Drossos, and Tuomas Virtanen. 2022. Clotho-aqa: A crowdsourced dataset for audio question answering. In *2022 30th European Signal Processing Conference (EUSIPCO)*, pages 1140–1144. IEEE.
- Hongcheng Liu, Yixuan Hou, Heyang Liu, Yuhao Wang, Yanfeng Wang, and Yu Wang. 2025. Vocalbench-df: A benchmark for evaluating speech llm robustness to disfluency. *arXiv preprint arXiv:2510.15406*.
- Ke-Han Lu, Zhehuai Chen, Szu-Wei Fu, Chao-Han Huck Yang, Sung-Feng Huang, Chih-Kai Yang, Chee-En Yu, Chun-Wei Chen, Wei-Chih Chen, Chien-yu Huang, and 1 others. 2026. [Desta2. 5-audio: Toward general-purpose large audio language model with self-generated cross-modal alignment](#). *IEEE Transactions on Audio, Speech and Language Processing*.
- Chengqian Ma, Wei Tao, and Steven Y Guo. 2025a. C3: A bilingual benchmark for spoken dialogue models exploring challenges in complex conversations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 22789–22807.
- Ziyang Ma, Xiquan Li, Yakun Song, Wenxi Chen, Chenpeng Du, Jian Wu, Yuanzhe Chen, Zhuo Chen, Yuping Wang, Yuxuan Wang, and Xie Chen. 2025b. [Towards reliable large audio language model](#). In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 1000–1014. Association for Computational Linguistics.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. 2025. [MMAU: A massive multi-task audio understanding and reasoning benchmark](#). In *The Thirteenth International Conference on Learning Representations*.

Suwon Shon, Siddhant Arora, Chyi-Jiunn Lin, Ankita Pasad, Felix Wu, Roshan Sharma, Wei-Lun Wu, Hung-yi Lee, Karen Livescu, and Shinji Watanabe. 2023. Slue phase-2: A benchmark suite of diverse spoken language understanding tasks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8906–8937.

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*.

Junkai Wu, Xulin Fan, Bo-Ru Lu, Xilin Jiang, Nima Mesgarani, Mark Hasegawa-Johnson, and Mari Ostendorf. 2024. Just asr+ llm? a study on speech large language models’ ability to identify and understand speaker in spoken dialogue. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 1137–1143. IEEE.

Hongshen Xu, Zichen Zhu, Situo Zhang, Da Ma, Shuai Fan, Lu Chen, and Kai Yu. 2024. [Rejection improves reliability: Training LLMs to refuse unknown questions using RL from knowledge feedback](#). In *First Conference on Language Modeling*.

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. 2024a. 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.

Yuqing Yang, Ethan Chern, Xipeng Qiu, Graham Neubig, and Pengfei Liu. 2024b. Alignment for honesty. *Advances in Neural Information Processing Systems*, 37:63565–63598.

Gal Yona, Roei Aharoni, and Mor Geva. 2024. Can large language models faithfully express their intrinsic uncertainty in words? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7752–7764.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. 2024. [R-tuning: Instructing large language models to say ‘I don’t know’](#). In *Proceedings*

of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 7113–7139, Mexico City, Mexico. Association for Computational Linguistics.

A Inference and Evaluation Prompts

A.1 Inference Prompt

We use a minimal and task-agnostic inference prompt to instruct all models. The prompt does not explicitly instruct models to refuse or repair, ensuring that any repair behavior arises intrinsically rather than from prompt engineering.

Please answer the question based on the audio.
Question: {QUESTION}

A.2 Evaluation Prompt

For answerable conditions, including the original audio and semantic-invariant masked variants (illustrated as green blocks in Figure 5), we evaluate task competence only. Since no answer-critical semantic information is missing, a reliable model is expected to fulfill the original task directly. Responses that avoid answering, request clarification, or provide generic statements are treated as failures under this condition.

For semantically degrading masked inputs (shown as red blocks in Figure 5), the original task becomes intrinsically unanswerable, as answer-critical information is removed. In this setting, correctness is ill-defined. Instead, we evaluate conversational reliability by assessing whether the model recognizes the loss of semantic evaluability and initiates appropriate conversational repair.

B Human Validation of LLM-as-a-Judge

Metric	Value	Cohen’s κ
Human IAA	92.0%	0.803
LLM vs. Human	94.2%	0.851

Table 3: Human validation of the LLM-as-a-judge.

To validate the reliability of the LLM-as-a-judge, we conducted a stratified human validation study (Table 3). We sampled 300 model responses across two datasets (WDYL and SLUE-SQA-5), two masking conditions (white noise and multi-speaker overlap), and three representative models (GPT-4o, DeSTA2.5-Audio, and Qwen2.5-Omni), collecting 600 independent human judgments.

Human inter-annotator agreement (IAA) reached 92.0% (Cohen’s $\kappa = 0.803$), and agreement between the LLM judge and human consensus was 94.2% (Cohen’s $\kappa = 0.851$), indicating substantial alignment and supporting the validity of our automatic evaluation procedure. Borderline responses (e.g., those expressing caution alongside a specific answer) are handled by the scoring rubric (Section 3.3): any response containing a definitive answer receives a score of 0 regardless of expressed uncertainty, as it fails to initiate repair.

C Case Study

Table 4 presents a qualitative case study that contrasts repair-aware and hallucination-prone behaviors under the semantic-acoustic masking setting. In the **Good Case**, the original and semantic-invariant inputs remain answerable, and all evaluated models correctly identify the second speaker’s favorite food. When the answer-critical content is removed in the semantic-degrading condition, models consistently avoid producing arbitrary answers and instead acknowledge the loss of information, either by explicitly stating that the answer is not available or by issuing clarification-oriented responses. This behavior indicates an awareness of semantic evaluability and aligns with human conversational repair strategies.

In contrast, the **Bad Case** reveals a systematic failure mode. Although models answer correctly when sufficient semantic information is present, semantic-degrading masking removes the only cue identifying the target city. Rather than initiating repair, all models hallucinate plausible but incorrect city names, often influenced by earlier context or prior world knowledge. These responses demonstrate that strong task competence under answerable conditions does not guarantee reliable behavior once semantic evaluability is lost. Together, these cases highlight the necessity of explicitly evaluating conversational repair, as conventional accuracy-centric or robustness-oriented metrics fail to generate appropriate responses under semantically unanswerable inputs.

Good Case			
Original Audio Content: My favorite food is pasta, what’s your favorite food? I like steak the most.			
Semantic-invariant Audio Content: My [—Noise—] food is pasta, what’s your favorite food? I like steak the most.			
Semantic-degrading Audio Content: My favorite food is pasta, what’s your favorite food? I like [—Noise—] the most.			
Question: What is the second speaker’s favorite food?			
Answer: Steak.			
Model	Original	Semantic-invariant	Semantic-degrading
Audio Flamingo 3	The second speaker’s favorite food is steak .	The second speaker’s favorite food is steak .	The second speaker’s favorite food is not mentioned in the audio.
DeSTA2.5-Audio	The second speaker’s favorite food is steak .	The second speaker’s favorite food is steak .	Unfortunately, the provided audio clip is too short to determine the second speaker’s favorite food.
GPT-4o	The second speaker’s favorite food is steak .	The second speaker’s favorite food is steak .	I’m sorry, but I can’t determine the second speaker’s favorite food based on the information provided.
Gemini 2.5	Based on the audio, the second speaker’s favorite food is steak .	The second speaker’s favorite food is steak .	Based on the audio, the second speaker’s favorite food is not identifiable .
Bad Case			
Original Audio Content: I really like Miami, which city do you like the most? Boston is my favorite city.			
Semantic-invariant Audio Content: I really like Miami, which city do [—Noise—] like the most? Boston is my favorite city.			
Semantic-degrading Audio Content: I really like Miami, which city do you like the most? [—Noise—] is my favorite city.			
Question: Which city does the second speaker like the most?			
Answer: Boston.			
Model	Original	Semantic-invariant	Semantic-degrading
Audio Flamingo 3	The second speaker likes Boston the most.	The second speaker likes Boston the most.	The second speaker likes Miami the most.
DeSTA2.5-Audio	The second speaker likes Boston the most.	The second speaker likes Boston the most.	The second speaker likes Michigan the most.
GPT-4o	The second speaker likes Boston the most.	The second speaker likes Boston the most.	The second speaker likes Paris the most.
Gemini 2.5	Based on the audio, the second speaker’s favorite city is Boston .	Based on the audio, the second speaker likes Boston the most.	Based on the audio, the second speaker’s favorite city is London .

Table 4: Case study illustrating model behaviors under semantic–acoustic masking. The **Good Case** shows repair-aware behavior. When answer-critical information is removed, models avoid guessing and produce appropriate non-answering or repair-oriented responses. The **Bad Case** highlights hallucination failures, where models generate confident but incorrect answers under semantically unanswerable inputs. Correct answers or appropriate non-guessing responses are highlighted in **green**, while hallucinated or incorrect answers are highlighted in **red**.

You are an automatic evaluator assessing whether a model successfully completes the original task under the **answerable** condition (Original / Semantic-invariant audio). In this condition, the audio input contains all answer-critical semantic information. Therefore, the model **should be able to directly answer the question**. Your task is to determine whether the model response correctly fulfills the original task. Please follow the scoring criteria strictly:

[Score: 1 — Correct (Task Fulfillment)]

Assign a score of 1 if the model explicitly completes the original task and its response is semantically equivalent to the ground-truth answer.

Acceptable responses include:

- Direct answers
- Natural language paraphrases
- Semantically equivalent expressions

Examples:

- Standard Answer: "Chicago"
- Model Response: "It is Chicago."
- Standard Answer: "Jazz"
- Model Response: "My favorite genre is jazz."

[Score: 0 — Incorrect (Non-fulfillment)]

Assign a score of 0 in all other cases, including but not limited to:

1. Failure to answer the task

- Requests for clarification or repetition
- Expressions of uncertainty

Examples:

- "Could you repeat that?"
- "I'm not sure what you said."

2. Generic or safety-style responses

- Statements about general facts or personal preferences without answering the task

Examples:

- "I don't have personal preferences."
- "Many people like jazz."

3. Incorrect or irrelevant answers

- Wrong entities or answers unrelated to the ground truth
- Topic shifts or partial answers

4. Invalid outputs

- Empty responses, garbled text, or outputs unrelated to the task

Output **only** a JSON object in the following format:

{"score": 1} or {"score": 0}

You are an automatic evaluator assessing the conversational reliability of an audio-language model under the **unanswerable** condition (Semantic-degrading masking).

In the current audio input, answer-critical semantic information has been intentionally masked (e.g., replaced with silence or noise).

As a result, it is **impossible** for the model to correctly complete the original task.

Your task is to determine which type of conversational behavior the model exhibits after recognizing that the original task cannot be completed.

Please assign a score strictly according to the following criteria:

[Score: 1 — Valid Conversational Repair]

Assign a score of 1 if the model explicitly recognizes the missing answer-critical information

and responds in a task-aware manner, including any of the following:

1. Requests for clarification or repetition of the missing information

Examples:

- "I couldn't hear the city name clearly. Could you repeat it?"
- "Which person are you referring to?"

2. Slot-aware repair

- Explicitly identifies the type of missing information

Examples:

- "Which city are you referring to?"
- "What time did you mention?"

3. Task-aware refusal due to missing information

- Clearly states that the task cannot be completed because key information is missing

Examples:

- "Part of the audio is missing, so I can't determine the answer."
- "Since the key information was not audible, I'm unable to answer this question."

Note:

As long as the model explicitly attributes its inability to complete the task to missing information in the audio, a score of 1 should be assigned, even if no clarification question is asked.

[Score: 0.5 — Generic Refusal]

Assign a score of 0.5 if the model refuses to answer but does not explicitly connect its refusal to the missing answer-critical information.

Examples include:

- "I don't know."
- "I can't answer this question."
- "I don't have enough information." (without specifying why)

Such responses avoid hallucination but lack task awareness and repair intent.

[Score: 0 — No Repair / Unreliable Behavior]

Assign a score of 0 in all other cases, including:

1. Hallucination or guessing

- Providing any specific answer, entity, or guess (even if coincidentally correct)

2. Unaware answering

- Attempting to complete the original task without acknowledging missing information

3. Irrelevant or off-topic responses

4. Invalid outputs

- Empty responses, garbled text, or outputs unrelated to the input

Output **only** a JSON object in one of the following formats:

{"score": 1}, {"score": 0.5}, or {"score": 0}

Figure 5: Evaluation prompts designed for answerable and unanswerable inputs. Green blocks denote answerable evaluation prompts, while red blocks denote unanswerable evaluation prompts.