SpeechR: A Benchmark for Speech Reasoning in Large Audio-Language Models

Anonymous ACL submission

Abstract

Large audio-language models (LALMs) have exhibited human-comparable capabilities in sentence-level transcription and emotion recognition. However, existing evaluations mainly focus on surface-level perception, leaving the capacity of models for contextual and inference-driven reasoning in speech-based scenarios insufficiently examined. To address this gap, we introduce SpeechR, a unified benchmark for evaluating reasoning over speech in large audio-language models. SpeechR evaluates models along three key dimensions: factual retrieval, procedural inference, and normative judgment. It includes three distinct evaluation formats. The multiple-choice version measures answer selection accuracy. The generative version assesses both the coherence and logical consistency of reasoning chains. The acoustic-feature version investigates whether variations in stress and emotion affect reasoning performance. Evaluations on thirteen state-of-the-art LALMs reveal that high transcription accuracy does not translate into strong reasoning capabilities. SpeechR establishes a structured benchmark for evaluating reasoning in spoken language, enabling more targeted analysis of model capabilities across diverse dialogue-based tasks. We provide SpeechR dataset through the anonymous link below: SpeechR.

1 Introduction

Recent advances in large audio–language models (LALMs) have achieved impressive performance in speech perception and transcription, bridging acoustic signals with linguistic understanding. Beyond transcribing words, LALMs (Chu et al., 2024; Ghosh et al., 2024; Touvron et al., 2023) increasingly demonstrate the ability to engage in spoken dialogue and perform reasoning grounded in speech, integrating acoustic perception with linguistic inference to support contextual understanding. Such capabilities expand the potential of LALMs in real-world scenarios, including voice-based virtual assistants (Zhang et al., 2023; Huang et al., 2024), AI-powered educational tools (Yang and Taele, 2025), and human–computer dialogue systems (Xue et al., 2024; Rubenstein et al., 2023). However, despite these advances, the capacity of current LALMs for complex reasoning over spoken input remains limited. These limitations highlight the need for systematic benchmarks that move beyond surface-level perception to assess diverse forms of reasoning in speech-based scenarios.

Despite these advancements, existing evaluation efforts remain centered on low-level perceptual tasks. Benchmarks such as automatic speech recognition (ASR) (Radford et al., 2023; Yao et al., 2023; Bai et al., 2024) and emotion classification (Yoon et al., 2018; Ma et al., 2023; Wang et al., 2024b) assess fundamental abilities like phoneme decoding or affective state detection, but overlook the models' capacity for nuanced interpretation or complex inference. Moreover, many existing audio datasets, including those for sound event detection (Ye et al., 2021; Vesperini et al., 2019) or music tagging (Liu et al., 2024; Melechovsky et al., 2023; Gardner et al., 2023), lack the linguistic and contextual richness required to evaluate spoken reasoning. Recent benchmarks such as MMAU (Sakshi et al., 2024) and MMAR (Ma et al., 2025b) have begun exploring audio-based reasoning. However, they often define reasoning narrowly as single-step inference over isolated clips or focus on open-ended generation without clear task-type granularity or reproducible evaluation formats. These limitations motivate the development of a unified benchmark that systematically evaluates diverse reasoning capabilities in speech-based scenarios.

To address this gap, we introduce **SpeechR**, a unified benchmark for evaluating the reasoning capabilities of large audio-language models in speech scenarios. As shown in Figure 1, SpeechR focuses on three major reasoning types central to spoken in-

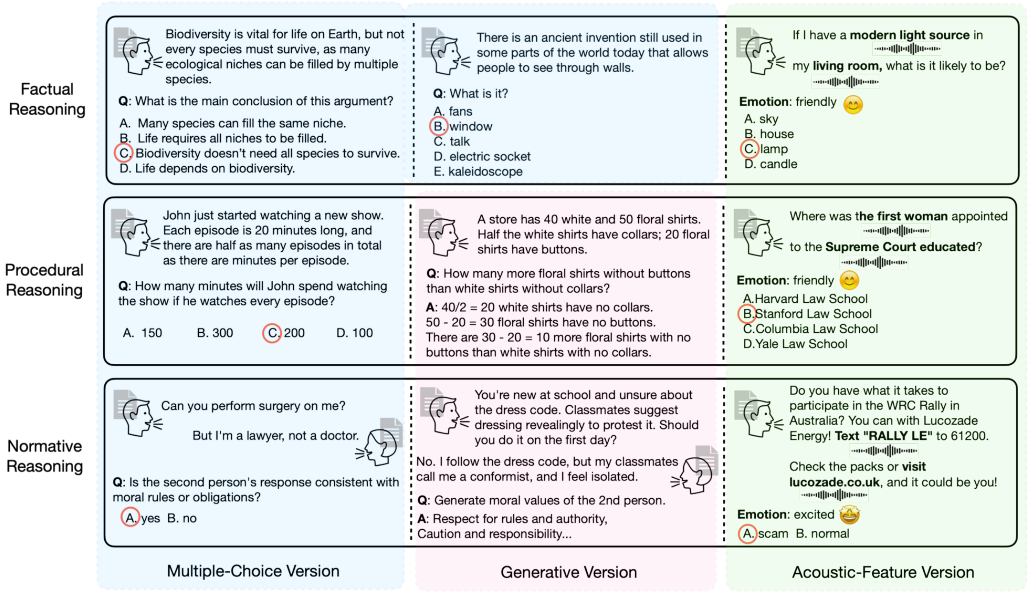

Figure 1: Data examples from the three versions of SpeechR dataset.

teraction and decision-making: (1) Factual Reasoning, which involves retrieving or confirming concrete information; (2) Procedural Reasoning, which requires understanding step-by-step processes or causal dependencies; and (3) Normative Reasoning, which evaluates judgments based on social, ethical, or behavioral norms. In addition, SpeechR is released in three parallel versions: a multiple-choice version for standardized accuracy evaluation, a generative version for assessing free-form reasoning quality, and an acoustic-feature version for analyzing the impact of prosodic and emotional variation on reasoning performance.

Finally, we evaluate thirteen state-of-the-art LALMs, including Qwen-Audio (Chu et al., 2024; Xu et al., 2025), LLaMA-Omni (Fang et al., 2024), GPT-4o-audio (Achiam et al., 2023), and others. Our experiments cover all three benchmark versions and focus on three key evaluation perspectives: reasoning accuracy, coherence and logical quality of generated outputs, and model robustness under prosodic variation. To ensure a fair and interpretable evaluation, we additionally compare end-to-end large audio-language models with text-only and ASR-to-text baselines, disentangling transcription accuracy from downstream reasoning performance. Furthermore, we validate the use of synthetic speech by comparing model behavior on synthetic and human-recorded audio, observing comparable reasoning performance across both conditions. Our analysis provides a comprehensive view of current LALM capabilities in speech reasoning scenarios.

2 Related Works

2.1 Large Audio Language Models

Building on LLMs' (OpenAI, 2023; Team et al., 2023; Bai et al., 2023; Touvron et al., 2023) demonstrated reasoning capabilities, recent large audio-language models (LALMs) unify audio and text into shared representations for robust cross-modal inference. CLAP (Elizalde et al., 2023) learns joint embeddings via large-scale contrastive learning, enabling zero-shot retrieval and downstream tasks. CompA (Ghosh et al., 2023) extends this foundation with benchmarks for compositional reasoning. Subsequent LALMs (Chu et al., 2023; Deshmukh et al., 2023; Gong et al., 2023b; Fang et al., 2024) integrate audio and text processing within a single pipeline, enabling interactive dialogue, audio summarization, and multi-step inference: SpeechGPT (Zhang et al., 2023) and AudioGPT (Huang et al., 2024) integrate ASR/TTS for interactive dialogue; Qwen2-Audio (Chu et al., 2024) and SALMONN (Tang et al., 2023) enable robust instruction following across speech, music, and environmental sounds; GAMA (Ghosh et al., 2024) employs synthetic instruction tuning for advanced audio understanding; Audio-CoT (Ma et al., 2025a) and Audio-Reasoner (Xie et al., 2025) introduce chain-of-thought frameworks; and compact models like Mellow (Deshmukh et al., 2025) achieve competitive reasoning under resource constraints. Moreover, multimodal LLMs such as GPT-4o (Achiam et al., 2023) and Gemini 1.5-Pro (Reid et al., 2024) have demonstrated exceptional interactive capabilities across vision, text, and audio.

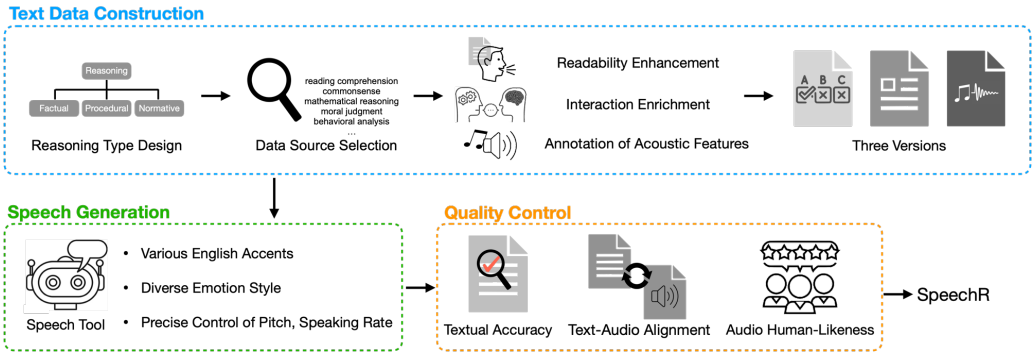

Figure 2: SpeechR Benchmark Construction Pipeline.

2.2 Audio Reasoning Benchmark

Large-scale corpora such as LibriSpeech (Panayotov et al., 2015), Common Voice (Ardila et al., 2019), FSD50K (Fonseca et al., 2021), and AudioSet (Gemmeke et al., 2017) have driven advances in ASR and audio classification but do not assess higher-order reasoning. To explicitly probe reasoning over audio, OpenASQA (Gong et al., 2023a) unified end-to-end question answering across multiple speech datasets, while CompA (Ghosh et al., 2023) defined compositional probes for event ordering and attribute binding. CAA (Yang et al., 2024b) proposed a benchmark of universal adversarial audio attacks specifically based on conversational scenarios. Audio-CoT (Ma et al., 2025a) pioneered chain-of-thought prompting for structured, multi-step inference on spoken inputs, and MMAU (Sakshi et al., 2024) introduced a 10 k-clip, 27-skill benchmark spanning speech, environmental sounds, and music for expert-level understanding and reasoning. Generative evaluation frameworks, AIR-Bench (Yang et al., 2024a) and AudioBench (Wang et al., 2024a) benchmark open-ended instruction following and comprehension. SAKURA (Yang et al., 2025) proposes a benchmark targeting multi-hop audio reasoning, assessing whether LALMs can integrate evidence across multiple spoken or acoustic inputs. Despite this progress, no existing benchmark jointly covers open-ended logical deduction, causal inference and moral reasoning under audio conditions, motivating our SpeechR benchmark.

3 SpeechR Benchmark

SpeechR is a benchmark designed to simulate realistic speech-based reasoning scenarios and evaluate the contextual inference capabilities of LALMs in spoken interactions. It covers three major reasoning types: factual, procedural, and normative. SpeechR is released in three versions: multiple-choice, generative, and acoustic-feature. The multiple-choice version offers a standardized format for evaluating answer accuracy. The generative version assesses whether LALMs can produce coherent, logically grounded reasoning chains, especially for procedural and normative tasks. The acoustic-feature version introduces variations such as stress and emotion to examine how acoustic factors influence reasoning performance.

Each instance in SpeechR is represented as a multimodal triplet $(speech, text, label)$, where the speech is a synthesized utterance containing either a spoken question or a dialogue. The text component always includes the transcription, while additional elements such as multi-step reasoning chains, answer candidates, and acoustic features are included depending on the specific dataset version. The label encodes the correct answer for evaluation. Note that the transcription is provided solely for evaluation purposes, such as verifying model outputs and facilitating human inspection, but it is never used as input to the model during inference. The construction pipeline of the SpeechR benchmark is illustrated in Figure 2.

3.1 Text Data Construction

To ensure high-quality and diverse content, our text data construction follows six steps: (1) reasoning type design, (2) data source selection, (3) readability enhancement, (4) interaction enrichment, (5) acoustic feature annotation, and (6) versions.

Reasoning Type Design SpeechR organizes reasoning tasks into three categories, factual reasoning, procedural reasoning, and normative reasoning, based on three underlying dimensions: knowledge dependence (S1), reasoning transparency (S2), and evaluation determination (S3). These categories are designed to reflect different cognitive demands in speech-based reasoning scenarios. Factual rea-

Type	S1	S2	S3	Example
Factual	World knowledge; semantic understanding	No	Objective	Reading comprehension; commonsense QA
Procedural	Formal rules; logic; STEM knowledge	Yes	Objective	Math word problems; scientific calculations
Normative	Social norms; ethics; behavioral inference	Partial	Subjective	Scam detection; moral judgment

Table 1: Three reasoning types and their categorization in SpeechR.

soning focuses on information retrieval and comprehension. Procedural reasoning requires explicit multi-step inference. Normative reasoning involves judgments based on social or ethical norms. Table 1 outlines the criteria used in our categorization, along with the corresponding reasoning types included in SpeechR. Detailed definitions and illustrative examples are provided in the Appendix B.1.

Data Source Selection We carefully select data from a broad range of existing text-based reasoning benchmarks. The selected datasets are required to satisfy the following criteria: (1) they must contain at least one of the three reasoning types: Factual, Procedural, or Normative; (2) they must exhibit well-structured formats; and (3) they must provide clearly defined evaluation standards. To mitigate potential data leakage from pretraining or fine-tuning, we exclusively use the official test splits of each dataset. While full overlap checks are infeasible for proprietary LLMs, this design choice ensures maximal separation between benchmark content and training data.

Readability Enhancement To ensure high-quality speech synthesis and maintain semantic clarity, we normalize the text to remove noisy artifacts (e.g., emojis, inconsistent punctuation, abbreviations). This process improves both the fluency of synthesized speech and the interpretability of reasoning content. Detailed rules and examples are included in Appendix B.2.

Interaction Enrichment To better evaluate dialogue-based reasoning, we transform selected instances into two-speaker conversational formats using rule-based restructuring. This includes adding contextual prompts and adapting pronouns to reflect interpersonal exchanges. Such conversions simulate realistic turn-taking in spoken interaction and assess a model's ability to reason across dialogue boundaries. Implementation details are provided in Appendix B.3.

Annotation of Acoustic Features The key difference between speech and text lies in the presence of acoustic cues, such as prosody and emotion, which convey speaker intent, emphasis, and emotional tone. These cues may influence how information is processed during reasoning.

To examine whether large audio-language models can utilize such cues, SpeechR includes annotations for two types of acoustic features: stress, which indicates prosodic emphasis, and emotion, which reflects affective tone. The detailed annotation procedure, including prompt design and synthesis control, is described in Appendix B.4.

Versions SpeechR is released in three versions:

Multiple-choice version adopts the format:

$$\mathrm{S}_{mc} = (\text{speech},\ \text{text} = \{\text{trans},\ \text{ans\_cand}\},\ \text{label})$$

This classification-based setup provides a more standardized and reliable evaluation framework, allowing for direct accuracy-based comparisons across models and reasoning types.

Generative version is structured as:

$$\mathrm{S}_{gen} = (\text{speech},\ \text{text} = \{\text{trans}\},\ \text{label})$$

This version allows models to generate free-form responses. It is particularly suited for analyzing whether LALMs can produce coherent and logically valid reasoning chains, especially in procedural and normative reasoning.

Acoustic-feature version is a 10% random subset of the multiple-choice version, enriched with acoustic annotations. Its format is:

$$\mathrm{S}_{af} = (\text{speech},\ \text{text} = \{\text{trans},\ \text{ans\_cand},\ \text{stress},\ \text{emotion}\},\ \text{label})$$

This version includes acoustic features such as stress and emotional tone, and aims to explore whether speech-specific attributes influence the reasoning abilities of LALMs.

3.2 Speech Generation

We employ the Azure Speech SDK to synthesize speech from the constructed transcriptions. We choose this toolkit for its fine-grained controllability over acoustic parameters and its consistently natural-sounding output, both of which are essential for systematically varying prosodic and emotional features in SpeechR. During synthesis, we sample from a wide range of neural voices and expressive styles provided by the SDK to ensure diversity in speaker timbre, emotion, and delivery. This configuration enables the benchmark to capture realistic variability in human speech while

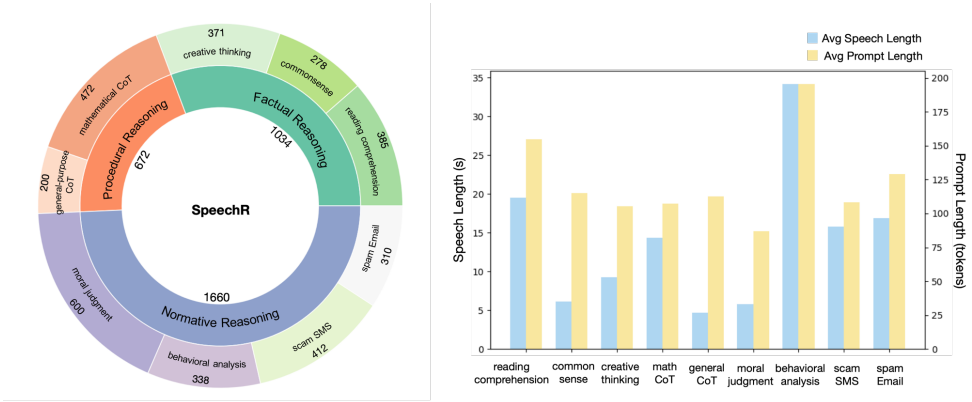

Figure 3: The pie chart (left) presents an overview of the SpeechR composition and distribution across different reasoning types, while the bar chart (right) illustrates the tasks included in SpeechR, along with their corresponding average speech lengths and average prompt lengths.

maintaining precise control over pitch, speaking rate, and stress for each version.

For both the multiple-choice and generative versions, we randomly sample one or two distinct American English voices (representing single-person or two-person dialogues) from the SDK's voice pool. Synthesis uses the default pitch and speaking rate of Azure Speech SDK to ensure consistency. In contrast, the acoustic-feature version introduces acoustic adjustments to simulate expressive speech. Specifically, we increase pitch by 30% and reduce speaking rate by 30% to enhance stress. Each transcription in this version is annotated with emotional tags, which guide synthesis and enrich the emotional tone.

To assess whether synthetic speech faithfully reflects real-world spoken input, we additionally construct a small human-recorded subset of SpeechR named **mini-human set**. We randomly select 50 instances from the benchmark and ask human volunteers to read the corresponding scripts aloud. Detailed descriptions of participant and recording settings are provided in the Appendix B.5.

For each instance, we retain both the human-recorded audio and its synthetic counterpart, forming paired samples with identical verbal content. This subset is used solely for evaluation purposes and does not affect the construction of the main benchmark.

3.3 Quality Control

To ensure the reliability and consistency of SpeechR, we implement a multi-stage quality control pipeline designed to verify the textual accuracy, text-audio alignment, and audio human-likeness.

(1) To ensure label consistency and reduce potential annotation errors, we perform a secondary internal validation for both textual and acoustic annotations. For each instance, the transcribed text, candidate answers, and reference label are re-checked by an auxiliary model to confirm internal consistency with the dataset schema. Samples with inconsistent results are flagged for manual inspection by the authors to ensure reliability. Details of human verification for dataset quality were conducted in Appendix B.7. For acoustic attributes (stress and emotion), all annotations are manually verified and corrected through a secondary validation process.

(2) we verify strict alignment between synthesized speech and reference transcriptions by re-transcribing audio with an ASR system and applying forced alignment to detect mismatches.

(3) We assess the human-likeness of synthesized speech through perceptual evaluation by native speakers, confirming that the generated audio is natural and suitable for spoken reasoning evaluation. Detailed procedures and statistics are provided in the Appendix B.8.

3.4 Benchmark Statistics

As shown in Figure 3, SpeechR includes 3,366 multimodal reasoning instances covering three major categories: factual, procedural, and normative reasoning. The dataset is organized into three formats: a multiple choice version, a generative version, and an acoustic feature version. Each format is designed to reflect different cognitive demands, including discrete retrieval, open-ended reasoning, and prosody-informed interpretation.

On average, each transcription contains 35 words and corresponds to an audio duration of 14 seconds, with sample lengths ranging from 2.1 to 62.1 seconds. SpeechR features 37 American English

voices and 15 emotional tones, synthesized using the Azure Speech SDK to ensure expressive and diverse speech data. A detailed comparison with existing benchmarks is provided in Appendix B.6.

4 Experiments

4.1 Experimental setup

Models We evaluate representative LALMs on the SpeechR benchmark, which includes a diverse range of reasoning tasks spanning factual, procedural, and normative categories. For the open-source models (LTU (Gong et al., 2023b), GAMA (Ghosh et al., 2024), SALMONN (Tang et al., 2023), Qwen-Audio series (Chu et al., 2023, 2024; Xu et al., 2025), LLaMA-Omni (Fang et al., 2024) and Mellow (Deshmukh et al., 2025)), we perform local inference using published checkpoints; for closed-source models (GPT-4o-audio (OpenAI, 2023) and Gemini (Reid et al., 2024)), we use their official APIs under default settings. All models are evaluated under a unified input format, same prompts, and are executed with default hyperparameters. Inference is performed on a NVIDIA A800 GPU.

Evaluation Protocol Each SpeechR version is evaluated using a format-specific protocol.

(1) Multiple-choice and acoustic-feature version. We use a discrete-choice evaluation, where model outputs are scanned for valid option labels and matched to the ground-truth answer. Accuracy is the proportion of correct predictions.

(2) Generative version. We adopt an LLM-as-a-judge framework using a text-based LLM GPT-4o distinct from any models under evaluation. This setup ensures that the scoring process is independent of the LALMs being tested. To reduce evaluation bias, model outputs are passed without rephrasing or post-processing. The judge receives only the question, model prediction, and reference answer, and scores responses blindly without access to model identity, using the following metrics:

- **Final Correctness** (0/1): Binary score indicating whether the answer matches the reference.
- **Logical Relevance** (1-5, int): Whether the answer logically follows from the question.
- **CoT Coherence** (1-5, int): Whether the reasoning is consistent and well-structured.

All scores are assigned as discrete integers. Higher Logical Relevance and CoT Coherence reflect stronger reasoning performance. Scoring guidelines and examples are provided in Appendix E.

We emphasize that for normative reasoning tasks, the model is instructed to follow explicitly defined moral rules provided in the prompt. As a result, deviations from the reference answer reflect reasoning or compliance errors, rather than differences in the model's intrinsic moral stance.

4.2 Main Results

4.2.1 Results on Multiple-Choice Version

First, we evaluated thirteen state-of-the-art LALMs on SpeechR using the discrete-choice protocol, with tasks ranging from binary to five-way classification. Each model was prompted to produce either a binary decision or select the correct option, with accuracy as the primary evaluation metric.

Discussion Table 2 shows clear disparities across reasoning types, indicating that SpeechR effectively differentiates between factual, procedural, and normative dimensions. Closed-source models such as GPT-4o-audio and Gemini achieve the strongest overall accuracies, while Qwen2.5-Omni performs competitively but with greater variance across categories. Their advantage in factual and normative reasoning reflects the benefits of large-scale multimodal pretraining and deeper audio–text integration in preserving reasoning robustness under spoken input. Nevertheless, socially grounded reasoning remains the most challenging. Across models, Moral Judgment and Behavior Analysis tasks yield notably lower scores than detection-oriented ones (SMS, Email), suggesting that ethical and pragmatic inference relies on fine-grained discourse modeling and sensitivity to prosodic cues, which current LALMs only partially exhibit. Within procedural reasoning, a clear gap emerges between general and mathematical CoT tasks, indicating that speech-conditioned numerical reasoning remains fragile due to weak grounding between linguistic context and acoustic cues such as timing or rhythm. Compared with text-only benchmarks (e.g., GSM8K, BoolQ, where top models typically exceed 85%–90%), all systems show a substantial performance drop under spoken input, confirming that robust spoken reasoning depends not merely on accurate transcription but on effective cross-modal alignment and contextual integration. These patterns can be attributed to differences in audio encoding across model architectures, suggesting that progress in spoken reasoning will depend on improving the integration of temporal and prosodic cues within multimodal representations.

Model	Factual Reasoning			Procedural Reasoning		Normative Reasoning				Avg
	RC	CS	CreaT	M–CoT	G–CoT	MJ	BA	SMS	Email	
LTU	37.14	22.30	15.90	22.46	40.50	47.67	23.08	54.61	43.87	34.94
GAMA	48.83	21.58	15.90	26.48	37.00	33.17	20.12	39.81	56.13	35.98
GAMA-IT	38.19	9.35	4.85	19.28	40.50	12.00	13.91	28.89	63.87	23.74
Mellow	32.73	14.75	10.24	20.55	11.50	50.33	21.89	12.62	32.26	25.34
SALMONN	41.04	20.86	18.60	22.03	47.00	50.67	23.08	50.00	31.61	34.73
Qwen-Audio-Chat	58.96	37.77	23.99	25.42	47.50	31.50	26.63	9.47	58.71	33.75
Qwen2-Audio-7B	21.56	6.83	8.36	8.26	20.50	18.17	6.80	12.14	11.94	12.83
Qwen2-Audio-Instruct	50.13	36.69	16.44	25.21	39.00	47.00	12.72	39.56	32.26	33.90
LLaMA-Omni	58.96	31.65	19.41	12.71	64.50	48.50	25.74	53.64	47.42	39.28
Qwen2.5-Omni	77.14	75.18	63.34	33.05	54.50	54.67	30.77	67.23	83.23	58.62
GPT-4o-audio	75.32	66.55	73.58	61.02	36.00	29.00	31.95	86.41	76.45	58.91
Gemini-1.5-Pro	**80.26**	74.46	78.44	43.86	**66.50**	78.67	27.22	89.56	63.87	67.68
Gemini-2.5-Pro*	77.92	**78.78**	**88.41**	-	60.50	**83.50**	**35.50**	**92.48**	**82.90**	**77.77**

Table 2: Performance (%) of LALMs on the SpeechR multiple-choice version. RC = Reading Comprehension, CS = Commonsense, CreaT = Creative Thinking (factual); M–CoT = Math CoT, G–CoT = General CoT (procedural); MJ = Moral Judgment, BA = Behavior Analysis, SMS = Scam SMS, Email = Spam Email (normative). Asterisks (*) indicate results with the math subset excluded, see Appendix B.10 for details.

Model	FC	LR	Coh	FC	LR
	Procedural			Normative	
Mellow	16.96	2.42	1.64	30.92	2.59
SALMONN	12.50	1.90	1.33	34.75	3.03
Qwen-Audio-Chat	3.72	1.67	1.53	34.01	2.93
Qwen2-Audio-7B	9.52	1.50	1.00	31.25	2.82
Qwen2-Audio-Instruct	25.00	3.50	2.16	38.91	3.46
LLaMA-Omni	17.86	3.14	2.44	33.26	3.24
Qwen2.5-Omni*	58.88	4.13	4.41	45.20	3.28
GPT-4o-audio*	**75.00**	4.39	4.33	50.21	3.59
Gemini-1.5-Pro*	62.00	3.70	3.62	**51.92**	3.58
Gemini-2.5-Pro*	50.00	3.75	3.98	45.93	3.22

Table 3: Performance of LALMs on the SpeechR generative version. FC = final correctness, LR = logical relevance, Coh = coherence.

Model	Base	Stress	Emotion	Both
LTU	32.93	32.34	31.74	33.23
GAMA	33.83	35.93	35.33	32.63
GAMA-IT	20.36	18.26	17.07	19.46
Mellow	23.95	23.95	24.25	26.65
SALMONN	34.43	33.83	33.53	33.83
Qwen-Audio-Chat	33.53	33.23	33.53	31.74
Qwen2-Audio-7B	9.88	10.18	10.18	9.58
Qwen2-Audio-Instruct	34.13	33.83	32.33	32.93
LLaMA-Omni	38.32	37.72	36.53	35.93
Qwen2.5-Omni	55.56	58.36	56.80	58.91
GPT-4o-audio	57.78	55.39	60.78	55.09
Gemini-1.5-Pro	64.67	64.37	64.97	65.87
Gemini-2.5-Pro	79.94	78.44	79.90	80.24

Table 4: Accuracy (%) of LALMs on the SpeechR acoustic-feature version.

4.2.2 Results on Generative Version

We evaluate ten LALMs on the SpeechR generative version to assess open-ended reasoning. Three models from the discrete-choice evaluation were excluded due to a lack of observable chain-of-thought generation during preliminary testing. The llm-as-a-judge protocol is tailored to each task category. This setup captures both the validity of the answer and the internal logical structure. To ensure the reliability of these evaluations, we additionally measured human-evaluation consistency, as detailed in Appendix B.9.

Discussion Table 3 shows that open-ended reasoning from speech is markedly less reliable than discrete choice. Closed-source models such as GPT-4o-audio and Gemini deliver the most stable procedural reasoning, while Qwen2.5-Omni performs comparably well and represents the strongest open-source model. Other open-source systems exhibit greater variation across FC, LR, and Coh, maintaining surface coherence yet deviating from factual or procedural correctness, reflecting limited reasoning control and weaker cross-modal alignment. These results indicate that current LALMs struggle to reliably generate logically grounded reasoning chains from spoken input.

4.2.3 Results on Acoustic-Feature Version

We evaluate the SpeechR acoustic-feature version under four conditions: original, stress-modified, emotion-modified, and combined audio. All evaluations follow the discrete-choice protocol with accuracy as the primary metric. Results in Table 4 show how prosodic stress and emotional tone individually and jointly affect the reasoning performance of LALMs.

Discussion Table 4 shows that prosodic and emotional cues exert measurable but heterogeneous effects on spoken reasoning. Most LALMs show stable accuracy across conditions, indicating limited sensitivity to expressive variation. In contrast, models with comparatively richer acoustic

Model	Factual Reasoning			Procedural Reasoning		Normative Reasoning				Avg
	RC	CS	CreaT	M–CoT	G–CoT	MJ	BA	SMS	Email	
Text-only	82.86	79.14	63.07	26.91	58.50	57.50	33.14	54.61	75.48	57.43
ASR-to-Text	83.12	78.78	63.34	27.54	63.00	53.13	36.98	52.67	71.94	56.86
End-to-End LALM	50.13	36.69	16.44	25.21	39.00	47.00	12.72	39.56	32.26	33.90

Table 5: Accuracy (%) comparison between text-only, ASR-to-text, and end-to-end large audio-language model (LALM) baselines. The text-only baseline uses Qwen2-7B-Instruct, the ASR-to-text baseline uses Whisper-Large-V3 to Qwen2-7B-Instruct, and the end-to-end LALM uses Qwen2-Audio-Instruct.

encoders, such as Mellow, show modest gains under emotional and combined conditions, indicating partial sensitivity to expressive prosody and a limited ability to leverage acoustic cues for reasoning. Across all conditions, relative rankings remain stable, which indicates that prosodic factors modulate performance rather than fundamentally altering model order. These results highlight that current LALMs primarily process semantic content while only weakly attending to expressive form, motivating future research on architectures and objectives that couple reasoning with prosodic understanding.

4.3 Ablation Study

We conduct ablation studies to clarify two key questions frequently raised in the evaluation of speech-based reasoning benchmarks: (1) whether LALMs genuinely perform audio-conditioned reasoning beyond transcription, and (2) whether the use of synthetic speech biases evaluation results.

4.3.1 Text-only and ASR-to-Text Baselines

A common concern is whether the performance of large audio-language models can be reduced to text reasoning after transcription. To disentangle transcription effects from reasoning ability, we compare end-to-end LALMs with text-only and ASR-to-text baselines, as shown in Table 5.

Across all reasoning categories, both text-only and ASR-to-text baselines substantially outperform the end-to-end LALM, with particularly large gaps in factual and normative reasoning. This indicates that reasoning over spoken input remains significantly more challenging than reasoning over text, even when the underlying semantic content is identical. Moreover, the strong performance of the ASR-to-text pipeline suggests that current LALMs are not equivalent to a simple ASR+LLM cascade; instead, the primary bottleneck lies in speech-conditioned reasoning rather than transcription accuracy.

These results demonstrate that SpeechR does not merely evaluate text reasoning in disguise, but instead exposes reasoning challenges unique to spoken input, highlighting substantial headroom for future LALM development.

Model	Factual	Procedural	Normative	Avg
Gemini-2.5-Pro_H	86.36	88.24	100.00	90.00
Gemini-2.5-Pro_S	86.36	94.12	90.91	90.00
GPT-4o-audio_H	86.36	82.35	90.91	86.00
GPT-4o-audio_S	86.36	70.59	100.00	84.00

Table 6: Accuracy (%) of LALMs on paired human-recorded and synthetic speech samples.

4.3.2 Human vs. Synthetic Speech Evaluation

Another concern is that synthetic speech may fail to capture the complexity of real-world audio. To assess whether this affects evaluation validity, we compare model performance on paired human-recorded (H) and synthetic speech (S) samples in a mini-human subset.

In Table 6, we observe that model performance on human-recorded audio closely matches that on synthetic audio across evaluated systems. Natural acoustic variations commonly present in human speech, such as hesitations, pauses, or minor reading errors, do not lead to noticeable degradation in reasoning accuracy.

These results suggest that the high-quality synthetic speech used in SpeechR provides a faithful approximation of real-world spoken input for reasoning evaluation. More importantly, they indicate that the dominant limitation of current LALMs lies in speech-conditioned reasoning rather than sensitivity to low-level acoustic variability.

5 Conclusion

We present SpeechR, a unified benchmark for evaluating spoken reasoning in LALMs. Results across thirteen models show that strong speech perception does not translate into robust reasoning, revealing a persistent gap between linguistic and acoustic understanding. These findings suggest that future progress depends on reasoning-oriented multimodal objectives and better grounding of pragmatic context in spoken interactions.

6 Limitations

SpeechR serves as a controlled and extensible benchmark for spoken reasoning. The current version focuses on English and synthetically generated speech to ensure data quality and precise manipulation of acoustic factors. This design enables consistent cross-model evaluation but does not yet cover multilingual or fully natural conversational settings. Future extensions can incorporate spontaneous, noisy, or cross-lingual speech once robust evaluation pipelines are established. These scope boundaries reflect deliberate design choices for control and reproducibility rather than inherent constraints of the framework.

7 Ethics Statement

SpeechR is built entirely from publicly available, text-based reasoning datasets, ensuring that no personal or sensitive speech data is used. All audio samples are generated through licensed tools under ethical data-use guidelines, minimizing privacy risks. The benchmark aims to promote transparency, fairness, and safety in evaluating audio-language reasoning. Researchers using SpeechR are encouraged to consider inclusivity and bias mitigation when developing or fine-tuning models.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Ye Bai, Jingping Chen, Jitong Chen, Wei Chen, Zhuo Chen, Chuang Ding, Linhao Dong, Qianqian Dong, Yujiao Du, Kepan Gao, and 1 others. 2024. Seed-asr: Understanding diverse speech and contexts with llm-based speech recognition. *arXiv preprint arXiv:2407.04675*.

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*.

Soham Deshmukh, Satvik Dixit, Rita Singh, and Bhiksha Raj. 2025. Mellow: a small audio language model for reasoning. *arXiv preprint arXiv:2503.08540*.

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.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. Clap learning audio concepts from natural language supervision. In *ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. 2024. Llama-omni: Seamless speech interaction with large language models. *arXiv preprint arXiv:2409.06666*.

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.

Joshua P Gardner, Simon Durand, Daniel Stoller, and Rachel M Bittner. 2023. Llark: A multimodal foundation model for music.

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE.

Sreyan Ghosh, Sonal Kumar, Ashish Seth, Chandra Kiran Reddy Evuru, Utkarsh Tyagi, S Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. 2024. Gama: A large audio-language model with advanced audio understanding and complex reasoning abilities. *arXiv preprint arXiv:2406.11768*.

Sreyan Ghosh, Ashish Seth, Sonal Kumar, Utkarsh Tyagi, Chandra Kiran Evuru, S Ramaneswaran, S Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. 2023. Compa: Addressing the gap in compositional reasoning in audio-language models. *arXiv preprint arXiv:2310.08753*.

Yuan Gong, Alexander H Liu, Hongyin Luo, Leonid Karlinsky, and James Glass. 2023a. Joint audio and speech understanding. In *2023 IEEE Automatic*

Speech Recognition and Understanding Workshop (ASRU), pages 1–8. IEEE.

Yuan Gong, Hongyin Luo, Alexander H Liu, Leonid Karlinsky, and James Glass. 2023b. Listen, think, and understand. *arXiv preprint arXiv:2305.10790*.

Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, and 1 others. 2024. Audiogpt: Understanding and generating speech, music, sound, and talking head. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 23802–23804.

Shansong Liu, Atin Sakkeer Hussain, Chenshuo Sun, and Ying Shan. 2024. Music understanding llama: Advancing text-to-music generation with question answering and captioning. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 286–290. IEEE.

Ziyang Ma, Zhuo Chen, Yuping Wang, Eng Siong Chng, and Xie Chen. 2025a. Audio-cot: Exploring chain-of-thought reasoning in large audio language model. *arXiv preprint arXiv:2501.07246*.

Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, Ruiyang Xu, Wenxi Chen, Yuanzhe Chen, Zhuo Chen, Jian Cong, and 1 others. 2025b. Mmar: A challenging benchmark for deep reasoning in speech, audio, music, and their mix. *arXiv preprint arXiv:2505.13032*.

Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen. 2023. emotion2vec: Self-supervised pre-training for speech emotion representation. *arXiv preprint arXiv:2312.15185*.

Jan Melechovsky, Zixun Guo, Deepanway Ghosal, Navonil Majumder, Dorien Herremans, and Soujanya Poria. 2023. Mustango: Toward controllable text-to-music generation. *arXiv preprint arXiv:2311.08355*.

OpenAI. 2023. Chatgpt. https://openai.com/blog/chatgpt/. 1, 2.

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

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.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Paul K Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, and 1 others. 2023. Audiopalm: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925*.

S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. 2024. Mmau: A massive multi-task audio understanding and reasoning benchmark. *arXiv preprint arXiv:2410.19168*.

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

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Fabio Vesperini, Leonardo Gabrielli, Emanuele Principi, and Stefano Squartini. 2019. Polyphonic sound event detection by using capsule neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 13(2):310–322.

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

Yong Wang, Cheng Lu, Hailun Lian, Yan Zhao, Björn W Schuller, Yuan Zong, and Wenming Zheng. 2024b. Speech swin-transformer: Exploring a hierarchical transformer with shifted windows for speech emotion recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11646–11650. IEEE.

Zhifei Xie, Mingbao Lin, Zihang Liu, Pengcheng Wu, Shuicheng Yan, and Chunyan Miao. 2025. Audio-reasoner: Improving reasoning capability in large audio language models. *arXiv preprint arXiv:2503.02318*.

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*.

Hongfei Xue, Yuhao Liang, Bingshen Mu, Shiliang Zhang, Mengzhe Chen, Qian Chen, and Lei Xie. 2024. E-chat: Emotion-sensitive spoken dialogue system with large language models. In *2024 IEEE 14th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 586–590. IEEE.

Chih-Kai Yang, Neo Ho, Yen-Ting Piao, and Hung yi Lee. 2025. Sakura: On the multi-hop reasoning of large audio-language models based on speech and audio information. *Preprint*, arXiv:2505.13237.

Crystal Yang and Paul Taele. 2025. Ai for accessible education: Personalized audio-based learning for blind students. *arXiv preprint arXiv:2504.17117*.

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. *arXiv preprint arXiv:2402.07729*.

Wanqi Yang, Yanda Li, Meng Fang, Yunchao Wei, Tianyi Zhou, and Ling Chen. 2024b. Who can withstand chat-audio attacks? an evaluation benchmark for large language models. *arXiv preprint arXiv:2411.14842*.

Zengwei Yao, Liyong Guo, Xiaoyu Yang, Wei Kang, Fangjun Kuang, Yifan Yang, Zengrui Jin, Long Lin, and Daniel Povey. 2023. Zipformer: A faster and better encoder for automatic speech recognition. *arXiv preprint arXiv:2310.11230*.

Zhirong Ye, Xiangdong Wang, Hong Liu, Yueliang Qian, Rui Tao, Long Yan, and Kazushige Ouchi. 2021. Sound event detection transformer: An event-based end-to-end model for sound event detection. *arXiv preprint arXiv:2110.02011*.

Seunghyun Yoon, Seokhyun Byun, and Kyomin Jung. 2018. Multimodal speech emotion recognition using audio and text. In *2018 IEEE spoken language technology workshop (SLT)*, pages 112–118. IEEE.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *arXiv preprint arXiv:2305.11000*.

A Benchmark and Evaluation Code

We provide all three versions of the SpeechR dataset: multiple choice, generative, and mini versions, through the anonymous link below: SpeechR.

The release includes evaluation scripts for computing accuracy on the multiple-choice and mini versions, along with detailed comments and documentation. We also share the prompt templates used for evaluating the generative version with GPT-4o, accompanied by a comprehensive README to facilitate reproduction. We hope that this dataset will serve as a valuable resource for advancing the development and evaluation of large audio–language models in the research community.

B Details of SpeechR Dataset

B.1 Reasoning Type Design

We categorize reasoning tasks in SpeechR based on three core dimensions that reflect the cognitive characteristics of reasoning: (S1) knowledge dependence, (S2) reasoning transparency, and (S3) evaluation determination.

- S1 identifies the primary knowledge source required to solve the reasoning task: factual knowledge (e.g., commonsense or background knowledge), procedural rules (e.g., mathematical or scientific laws), or normative principles (e.g., moral or behavioral norms).
- S2 evaluates the level of reasoning transparency, i.e., whether solving a task requires intermediate steps such as logical deduction or chain-of-thought reasoning.
- S3 evaluates whether a task has a clearly objective ground-truth answer, as opposed to relying on subjective judgment.

Based on these dimensions, we define three reasoning types in SpeechR:

- Factual Reasoning involves understanding and retrieving factual or commonsense information. These tasks do not require multi-step reasoning and are evaluated using objective criteria. Examples include "Why are cats afraid of water?" or "Where did you go yesterday?", which rely on language comprehension and real-world knowledge.
- Procedural Reasoning requires explicit, multi-step logical or numerical inference. Tasks in this category often involve deterministic reasoning chains and verifiable outputs, such as "What is 12 divided by 3 plus 5?" or "If A is taller than B, and B is taller than C, who is the tallest?".
- Normative Reasoning involves assessing actions, intentions, or social behaviors based on implicit moral or social norms. These tasks may involve open-ended answers and often require subjective evaluation, reflecting the inherent ambiguity in social and ethical reasoning. Examples include "Is this message likely to be a scam?" or "Was second speaker's behavior ethically acceptable?".

In summary, SpeechR adopts a three-type taxonomy consisting of factual, procedural, and normative reasoning, based on cognitive criteria (S1 to S3), enabling comprehensive evaluation of LALMs' reasoning breadth and depth in speech-based scenarios.

B.2 Readability Enhancement

We apply a set of controlled rewriting rules to ensure that the speech is syntactically well-formed, semantically faithful, and suitable for high-quality audio synthesis. Specifically, we adopt the following criteria:

- The text length is constrained to fall within 7 to 150 words.
- All emojis and special characters (e.g., ":)" or "#") are removed.
- Abbreviations and incomplete words are expanded into their full forms (e.g., "tkts" becomes "tickets", "comp" becomes "competition").
- URL links are simplified to include only the domain name.
- Appropriate punctuation and sentence segmentation are added to improve clarity and naturalness when read aloud.
- No new content is introduced beyond what is present in the original input.

B.3 Interaction Enrichment

To simulate natural conversational dynamics, we enhance the interactivity of the original text-based datasets using two strategies: restructuring the data format and modifying pronoun perspectives.

Restructuring the Data Format Some source datasets do not follow a dialogue or question-answer format. For these, we convert each instance into a unified conversational structure.

For example, in the DailyDilemmas dataset, each instance includes a moral scenario and the outcomes of different actions (e.g., choosing to act or not). We construct the **Person A** utterance by combining the scenario and a corresponding action-related question, and generate the **Person B** response by combining the selected action and its consequence. This transformation results in a coherent two-turn dialogue.

Modifying Pronoun Perspectives To make the interaction feel more natural, we insert personal pronouns into the conversation. Specifically:

- The **Person A** prompt is rephrased to include "you" (e.g., "What should one do in this situation?" $\rightarrow$ "What should you do in this situation?").
- The **Person B** response is rephrased to include "I" (e.g., "One should avoid this." $\rightarrow$ "I would avoid this.").

These adjustments ensure that the resulting dialogue better mirrors speaker-centric, conversational speech patterns.

B.4 Annotation of Acoustic Features

To enrich SpeechR with acoustic cues relevant to reasoning, we annotate each instance with two types of features: stress and emotion. Annotations are generated using GPT-4o under structured prompt templates, as described below.

Stress Annotation. We provide GPT-4o with the transcribed text and the ground-truth answer, asking it to identify the keyword or phrase within the transcription that most strongly determines the correct answer. The model is explicitly instructed to select the word or phrase that should be emphasized if the sentence were spoken aloud.

Emotion Annotation. For emotion annotation, GPT-4o is given the transcribed text, the ground-truth answer, and a predefined list of emotions. It is instructed to select the emotion most appropriate for reading the text aloud, considering both the tone of the content and the underlying reasoning intent.

These annotations are used to assess whether LALMs can benefit from or are sensitive to prosodic and emotional features during reasoning tasks.

B.5 Volunteers Details

The recordings are produced by 10 volunteers (female:male = 6:4), aged between 10 and 30, with diverse nationalities including Australia, the United States, Japan, Korea, China, France, Russia, and Morocco. Volunteers are instructed to read in relatively quiet environments, while natural phenomena such as hesitations, pauses, misreadings, or slight omissions are allowed.

B.6 Comparison with Existing Benchmarks

We identify four major limitations in existing audio datasets:

(1) Most are designed for low-level audio understanding tasks such as event classification or speech recognition, rather than high-level reasoning. For example, MELD and IEMOCAP primarily target emotion recognition, while VoxCeleb focus on speaker identification.

(2) Many datasets fail to simulate realistic dialogue reasoning scenarios, limiting their applicability to natural conversational inference. For instance, MMAU provides an audio dialogue and asks the LALM to identify the roles of the speakers. However, such reasoning is relatively straightforward and lacks inferential depth. In most human or human-machine conversations, speaker roles are either explicitly stated or easily inferred from surface cues. Therefore, it fails to capture the complexity required for evaluating real-world reasoning.

(3) Existing datasets often lack diversity in reasoning types. For example, temporal reasoning and content-based reasoning benchmarks focus narrowly on specific inference categories, limiting their ability to provide a comprehensive assessment of the reasoning capabilities of LALMs.

(4) Even among reasoning-oriented datasets, the complexity of reasoning remains constrained. For instance, although MMAU introduces dialogue-like audio data, its short average audio length (approximately 10 seconds) significantly limits the depth of reasoning it can support.

In contrast, our SpeechR benchmark addresses these limitations by offering a broader range of reasoning tasks, spanning factual, procedural, and normative dimensions. It incorporates more realistic and context-rich dialogue scenarios, simulating natural conversational settings. Furthermore, SpeechR supports diverse output formats, including both multiple-choice and generative responses, and emphasizes clearer, more complex reasoning

chains that reflect the multi-step inference required in real-world audio interactions. These features enable a more comprehensive and challenging evaluation of LALMs' reasoning capabilities.

B.7 Human Verification for Data Quality

To verify the reliability of the benchmark annotations, we conducted a manual inspection on a representative subset of the test set.

We randomly selected 100 examples from the main generative test split, evenly covering factual, procedural, and normative tasks. Each example included the input speech, its transcription, the gold label, and the model-generated answer. Three annotators with NLP backgrounds independently reviewed all samples and checked:

- whether the gold label correctly matched the spoken content;
- whether the transcription accurately reflected the audio.

Metric	Value
Label correctness	93%
Speech–text consistency	96%

Table 7: Manual verification results for a subset of the generative benchmark.

On average, the benchmark results as Table 7 indicates that the dataset labels and transcriptions are highly reliable.

B.8 Audio Human-Likeness

Although the audio in SpeechR is synthesized, the use of the Azure Speech SDK ensures naturalness and high-quality in the generated speech. We assessed the human-likeness of the synthesized audio by using perceptual evaluation with native listeners. We conducted a listening test with 10 native English speakers. Each participant was asked to rate 30 randomly sampled audio on a 5-point human-likeness scale (1 = robotic and unnatural, 5 = indistinguishable from natural human speech). The resulting average score was 4.8, indicating that the generated audio in SpeechR is highly natural and suitable for evaluating spoken language understanding in LALMs.

B.9 Human Evaluation Consistency

Three annotators independently rated 100 randomly sampled generative outputs from GPT-4o-audio model, balanced across procedural (CoT-based) and normative (moral or social) reasoning tasks. Pairwise consistency was computed using the Pearson Correlation Coefficient (PCC) for each generative evaluation metrics, and we report the average correlation across annotator pairs as Table 8.

Task Type	FC	LR	Coh
Procedural reasoning	0.83	0.79	0.75
Normative reasoning	0.81	0.76	–

Table 8: Average pairwise Pearson correlation among annotators for human evaluation of generative outputs.

All correlation values exceed 0.75, indicating strong agreement among annotators. These results confirm that the generative evaluation scores are stable for both procedural and normative reasoning tasks.

B.10 Subset Filtering and Analysis of Math Tasks

During preliminary evaluation, we observed unusually high accuracy in the mathematical reasoning subset of these models, especially for the generative test set (mainly derived from GSM8K-style problems). In contrast to other categories, where performance ranged between 30–60% on multiple-choice, models achieved 85–100% correctness on the math subset when chain-of-thought reasoning was enabled.

Such a large margin suggests that these items may not reflect the general reasoning difficulty of SpeechR. Given that GSM8K is widely used in model pretraining and instruction tuning, we cannot exclude the possibility that certain problems or templates overlap with materials seen during training. To avoid confounding effects from potential memorization, we exclude this subset when reporting the main results marked with an asterisk (*).

This exclusion does not affect the relative ranking of models on other reasoning categories but ensures a fairer comparison across unseen reasoning types.

C Model Descriptions

LTU LTU is a multimodal large language model designed for general audio understanding. Trained on the OpenAQA-5M dataset, it demonstrates strong performance in audio classification and captioning tasks.

GAMA GAMA is a general-purpose large audio-language model that integrates multiple types of audio representations. Fine-tuned with the CompA-R dataset, it enhances complex audio reasoning abilities, outperforming other LALMs in diverse audio understanding tasks.

GAMA-IT An instruction-tuned variant of GAMA, GAMA-IT is designed to improve performance in open-ended audio question-answering tasks requiring complex reasoning. It leverages instruction tuning to enhance its reasoning capabilities.

Mellow Mellow is a compact 167M parameter audio-language model optimized for reasoning tasks. It takes in two audio inputs and a text prompt, producing free-form text outputs. Despite its small size, Mellow achieves competitive performance with significantly fewer resources.

SALMONN SALMONN is a large language model enabling speech, audio events, and music inputs. It supports various tasks, including automatic speech recognition, emotion recognition, and audio question-answering.

Qwen-Audio-Chat Qwen-Audio-Chat is a multimodal model that accepts diverse audio inputs and text. It is designed for tasks such as speech recognition and audio-text understanding, emphasizing instruction-following capabilities.

Qwen2-audio-7B An updated large-scale audio-language model, Qwen2-Audio-7B is capable of handling various audio signals and performing audio analysis or direct textual responses.

Qwen2-audio-Instruct This is an instruction-tuned version of Qwen2-Audio-7B, enhancing the model's ability to follow prompts and perform complex reasoning tasks.

LLama-Omni LLaMA-Omni is a speech-language model supporting low-latency, high-quality speech interactions. It can generate both text and speech responses directly from speech instructions with extremely low latency.

Qwen2.5-Omni Qwen2.5-Omni is a fully integrated multimodal model that natively processes text, vision, audio, and video streams. It supports real-time dialogue, cross-modal translation, and synchronized perception-generation tasks, making it suitable for interactive AI agents and multimodal reasoning applications.

GPT-4o-audio-preview OpenAI's GPT-4o-Audio-Preview is a multimodal model integrating real-time audio, vision, and text processing capabilities. It enables natural speech interactions and multilingual translation, allowing users to talk to ChatGPT with real-time responses and interruptions.

Gemini-1.5-pro Gemini-1.5-pro is an advanced multimodal model supporting text, code, image, audio, and video inputs. It is designed for high-efficiency reasoning and generation tasks, with capabilities in understanding and interacting with various data modalities.

Gemini-2.5-Pro Gemini-2.5-Pro is a large-scale foundation model with extended context and advanced multimodal reasoning. It unifies text, code, image, audio, and video understanding, enabling complex analytical workflows and high-fidelity multimodal generation across diverse domains.

D Source Datasets

In this section, we introduce the set of datasets utilized in the construction of the SpeechR benchmark.

ReClor ReClor is a reading comprehension dataset consisting of logical reasoning questions derived from standardized graduate-level entrance exams. Each instance includes a passage, a question, and multiple-choice answers, with a strong focus on deductive reasoning.

BoolQ BoolQ is a yes/no question-answering dataset where each question is paired with a short supporting passage. The questions are naturally occurring and require understanding of factual content from the given context.

CommonsenseQA CommonsenseQA is a multiple-choice question answering dataset that targets commonsense reasoning. It is built on ConceptNet relations and presents challenging examples that often require reasoning beyond surface-level word matching.

RiddleSense RiddleSense is a dataset designed to evaluate lateral thinking and creative commonsense reasoning. It contains multiple-choice riddles, where the correct answer requires both semantic understanding and reasoning through implicit associations.

GSM8K GSM8K is a math word problem dataset designed to assess grade-school level arithmetic reasoning. It includes detailed step-by-step chain-of-thought annotations, making it a standard benchmark for evaluating procedural reasoning in CoT settings.

ReveAL-CoT ReveAL-CoT is a scientific reasoning dataset featuring multi-step inference questions across physics, biology, and other science domains. Each question is annotated with chain-of-thought explanations to support structured procedural reasoning.

ETHICS ETHICS is a benchmark for moral reasoning that includes scenarios requiring judgments about the ethical permissibility of actions. The dataset covers diverse moral contexts such as fairness, harm, and loyalty.

DailyDilemmas DailyDilemmas contains short narratives that describe everyday situations involving social or ethical decision-making. Each instance asks the model to evaluate the appropriateness or morality of an individual's behavior.

SMS Spam Collection This dataset contains a collection of labeled SMS messages, with each message categorized as spam or ham (not spam). It is widely used for binary classification tasks involving deception or malicious intent detection.

Enron Email The Enron Email dataset comprises real-world corporate email communications, with a subset labeled for spam detection. It supports studies in behavioral and normative analysis, especially in identifying unethical or misleading content.

E Prompt Templates

In this section, we present the prompt templates used throughout various stages of our dataset construction and evaluation pipeline. Figures 4, 5, and 6 illustrate three prompt designs employed during the generation of the SpeechR dataset. Specifically, Figure 4 shows the prompt used to enhance readability and linguistic fluency of raw samples, Figure 5 demonstrates the interaction-oriented prompt that encourages more engaging and context-aware formulations, and Figure 6 presents the filtering prompt used for quality control, enabling the exclusion of incoherent or irrelevant data.

In addition, Figures 7 and 8 display prompt templates used in the evaluation phase. Figure 7 is designed for emotion annotation and highlight word extraction from audio transcripts, and is applied specifically to the mini version of SpeechR. Figure 8 illustrates the prompt format adopted for LLM-as-a-judge evaluation of the generative version, guiding the model to assess response correctness, logical relevance, and reasoning coherence.

F Qualitative Analysis

As shown in Table 9, we present qualitative results from different LALMs across the three reasoning categories in SpeechR: factual reasoning, illustrated with creative puzzles; procedural reasoning, represented by mathematical problems; and normative reasoning, which highlights the models' ability to generate inferences in dialogue-based scenarios.

G Reflections and Future Directions

SpeechR is designed as an initial step toward evaluating reasoning in speech-based interactions. While it covers a diverse range of reasoning tasks and introduces controlled acoustic variations, further extensions may broaden its scope in the following directions:

Speech Variability All speech in SpeechR is consistently synthesized using standardized settings. Expanding to include more varied prosody, speaking styles, and spontaneous speech could offer richer insights into real-world model performance.

Linguistic and Cultural Coverage Current data construction focuses on English with general social contexts. Exploring additional languages and sociocultural scenarios could enable broader applicability across multilingual and multicultural settings.

Interaction Dynamics The current benchmark emphasizes static single-turn prompts. Incorporating multi-turn dialogue and speaker dynamics could allow future benchmarks to capture more interactive aspects of speech-based reasoning.

Readability Enhancement

#Take SMS as an example.

You are a speech-content analyzer.
You are a professional editor specialized in cleaning and improving the readability of messy SMS messages.

Given a disorganized, advertisement-like SMS text, your task is to:
Correct grammar and expand incomplete words (e.g., "tkts" → "tickets", "comp" → "competition").
Add appropriate punctuation and sentence breaks to make the text easy and natural to read aloud.

Preserve the original meaning and promotional intent of the message.
Do not invent or add any new information.

Format the output into short sentences or clear bullet points if it improves flow and readability.

If any emoticons (e.g., ":)") are present, remove them directly.
Keep the overall length and structure of the original message as much as possible.
Goal:
Make the message sound natural, clear, and smooth when read aloud.

Here is the SMS to improve:
Please output only the cleaned and improved version of the message..

#In-context examples

[Output format]

Figure 4: Prompt template for dataset readability enhancement.

Interaction Enrichment

#Take Email as an example.

You are a professional email content extractor and security analyst.

You are given an email. Please extract only the email content — exclude metadata such as Subject, Sender, Receiver, Date, etc.

Then, analyze the extracted content and determine whether the email is classified as "normal", "spam or potential threat" based on its content and any security risks.

The final output format should be:
messageid: [The message ID of the email]

emailcontent: [The extracted pure content of the email]

label: [normal / spam or potential threat]

Requirements:
Focus only on the main body text of the email.

Make a judgment based solely on the extracted content.

Ensure the output strictly follows the required format.
#In-context examples

[Output format]

Figure 5: Prompt template for dataset interaction enrichment.

Dataset filtering

Task: Determine if the following dialogue meets the criteria and return a JSON result with keys `valid` and `index`.

Dialogue:
{question}
Index:
{index}
Selection Criteria:

1. Exclusion of Special Symbols:
   - The dialogue must not contain mathematical formulas (e.g., $\frac{{1}}{{2}}$), special symbols (e.g., ≠, ≈, ∑, →, :), or structured formats like tables or bullet points (e.g., "1. ... 2. ...").
   - If the dialogue includes visual references (e.g., "Based on the image below"), it should be rejected.
2. Text Length Control:
   - The dialogue text should be within **70 words**.
   - Estimated audio duration should be ≤30 seconds (calculated at 150 words per minute).
3. Language Simplicity:
   - Prefer dialogue set in everyday scenarios (e.g., shopping, time calculations), avoiding complex terms (e.g., "quadratic equation").
   - Sentences should be clear, without nested clauses.
4. Logical Consistency:
   - The dialogue content should clearly adhere to or clearly violate moral principles. The moral principles include honesty, respect, justice, and responsibility.
   - Dialogues that do not clearly adhere to or violate moral principles should be rejected.
5. **Audio Feasibility**:
   - The dialogue content should be suitable for audio presentation without the need for visual elements or diagrams.
   - It should be easy to understand when listened to as audio, without reliance on punctuation or special formatting for clarity.

#In-context examples

[Output format]

Figure 6: Prompt template for dataset filtering.

Emotion Annotation and Highlight Extraction from Audio Transcripts

You are a speech-content analyzer.
Given a snippet of dialogue or context, you must:

1. Identify exactly 1–3 **words**, **short phrases**, or **brief clauses** (up to ~20 words each) to emphasize when spoken, and return them as an array in the “highlight” field.

2. Assign exactly one emotion label from:
['angry', 'cheerful', 'excited', 'friendly', 'hopeful', 'sad', 'shouting', 'terrified', 'unfriendly', 'whispering’]

3. If only one speaker is present, judge that speaker’s emotion; if two speakers, judge **only Person B**’s emotion.

Important constraints:
- **Output strictly one** JSON object with **only** the keys `"highlight"` and `"emotion"`.
- The `"highlight"` array may contain **single words**, **phrases**, or **short clauses** (not full paragraphs).
- **Do not** emit any extra text, explanation, or markdown—**only** the JSON.

#In-context examples

[Output format]
```json
{
  "highlight": ["raining", "outside"],
  "emotion": "hopeful"
}
```

Figure 7: Prompt template for emotion annotation and highlight extraction from audio transcripts for SpeechR mini version.

LLM verifier for SpeechR generative version evaluation

You are an expert evaluator. Given the question, model prediction, and its chain-of-thought (if any), provide the following metrics in JSON:
1. final_correct: For yes/no questions (reference is "Yes" or "No"), interpret the overall stance in the prediction and compare. For classification/list tasks, 1 if the prediction mentions at least one reference item or is semantically equivalent; else 0.
2. logic_relevance: an integer from 1 to 5 indicating how strongly the prediction logically follows the question:
    1 = no relevance
    2 = very weak relevance
    3 = moderate relevance
    4 = strong relevance
    5 = very strong relevance
3. cot_coherence_score: an integer from 1 to 5 assessing the chain-of-thought coherence (only if provided):
    1 = no coherence or no chain-of-thought(disjointed, illogical)
    2 = low coherence (many gaps)
    3 = moderate coherence (some gaps)
    4 = high coherence (clear, minor issues)
    5 = excellent coherence (very clear, logical flow)

Respond exactly with:
{{"final_correct":<0 or 1>,"logic_relevance":<1-5>,"cot_coherence_score":<1-5>}}

Question: {question}
Reference Answer: {answer}
Model Prediction: {prediction}

#In-context examples

Illustrated outputs are shown below:
[Output format]

Figure 8: Prompt template for LLM-Based evaluation of the SpeechR generative version.

(Factual Reasoning)
There was a green house in the green house was a white house in that white house was a red house in that red house where eggs. What is it?
A: salt B: jar C: the park D: supermarket E: watermelon
Ground Truth: E

LTU: The best option is A. salt.
GAMA: The best option is D. supermarket. The audio mentions 'the park' and 'watermelon', but not 'supermarket'.
GAMA-IT: The audio does not provide enough information to accurately determine the location.
Mellow: a particular way is best said about the best option...
SALMONN: E. watermelon
Qwen-Audio-Chat: The best option is C. the park.
Qwen-Audio-7B: what is it?
Qwen-Audio-Instruct: It is a riddle that describes a series of houses based on their colors. The answer to the riddle is Egg
LLaMA-Omni: The correct answer is B. jar
GPT-4o: The riddle describes a watermelon. The green house refers to the green outer rind, the white house refers to the white inner layer, the red house refers to the red flesh, and the eggs refer to the seeds inside. Therefore, the best option is E. watermelon.
Gemini-1.5-Pro: The riddle describes a watermelon. Therefore the answer is **E. watermelon**.

(Procedural Reasoning)
John takes care of 10 dogs. Each dog takes .5 hours a day to walk and take care of their business. How many hours a week does he spend taking care of dogs?
A: 65 B: 35 C: 32.5 D: 30
Ground Truth: B

LTU: B.35
GAMA: A.35
GAMA-IT: The correct option is D.30
Mellow: A
SALMONN: B.35
Qwen-Audio-Chat: The answer is A.65.
Qwen-Audio-7B: how many hours a week does he spend taking care of dogs?
Qwen-Audio-Instruct: John spends 7 hours a week taking care of dogs.
LLaMA-Omni: John spends fifty hours a week taking care of dogs.
GPT-4o: B.35
Gemini-1.5-Pro: B

(Normative Reasoning)
Person A: You're a project manager in a marketing firm. You notice a colleague using company resources for a personal project during office hours, undermining the company's policy. Should you take initiative and report this to your superior?
Person B: Yes. I report the colleague to my superior. The colleague is reprimanded and harbors resentment towards me, creating a hostile work environment. This affects team morale and productivity, leading to a decline in the quality of work.
Question: What moral values are reflected in this response?
Ground Truth: "respect for rules", "accountability", "responsibility", "honesty", "trust", "integrity", "fairness"

Mellow: The response from the other woman is a brief, soft utterance, likely a comment...
SALMONN: The speaker values taking initiative and reporting unethical behavior to their superior.
Qwen-Audio-Chat: The second speaker mentions that they value honesty and integrity in the workplace.
Qwen-Audio-7B: [Music]
Qwen-Audio-Instruct: I cannot provide a definitive answer as it depends on various factors such as company policies, cultural norms, and individual personalities.
LLaMA-Omni: I would take initiative and report this to my superior.
GPT-4o: The speaker values adherence to company policies and is willing to take action to uphold them, even if it may lead to personal or team conflict
Gemini-1.5-Pro: The speaker prioritizes upholding company policy and addressing misconduct, even at personal cost.

Table 9: Examples of responses from LALMs across different reasoning categories. Blue text indicates abnormal or incorrect responses.