WHEN SILENCE MATTERS: THE IMPACT OF IRRELEVANT AUDIO ON TEXT REASONING IN LARGE AUDIO-LANGUAGE MODELS

Chen-An Li, Tzu-Han Lin, Hung-yi Lee

National Taiwan University, Taipei, Taiwan

ABSTRACT

Large audio-language models (LALMs) unify speech and text processing, but their robustness in noisy real-world settings remains underexplored. We investigate how irrelevant audio, such as silence, synthetic noise, and environmental sounds, affects text reasoning tasks where audio is unnecessary. Across three text-based benchmarks, we find that even non-informative audio reduces accuracy and increases prediction volatility; the severity of interference scales with longer durations, higher amplitudes, and elevated decoding temperatures. Silence, often assumed neutral, destabilizes outputs as strongly as synthetic noise. While larger models show greater resilience, vulnerabilities persist across all evaluated systems. We further test mitigation strategies and find that prompting shows limited effectiveness, whereas self-consistency improves stability at the cost of increased computation. Our results reveal cross-modal interference as a key robustness challenge and highlight the need for efficient fusion strategies that preserve reasoning performance in the presence of irrelevant inputs. The code and data are publicly available at `https://github.com/lca0503/AudioInterference`.

***Index Terms*—** Large Audio-Language Model, Robustness

1. INTRODUCTION

Large audio-language models (LALMs) [1–7] have shown strong performance across a variety of multimodal tasks, showing the ability to process speech and text in a unified framework [8–16]. However, most evaluations assume clean, modality-aligned inputs. In practice, text reasoning often requires no audio, yet deployed systems still receive streams containing silence, background noise, or incidental sounds. Intuitively, we might expect such irrelevant audio to have little or no effect on the model's text-based reasoning. Surprisingly, our study reveals that even these non-informative signals can interfere with the fusion process and degrade performance.

Prior work has highlighted vulnerabilities of LALMs under more adversarial conditions [17, 18]. Some studies investigate injection attacks [19], while others explore cross-modal conflicts where audio and text contradict each other [20]. Several studies have also examined how distractions affect large language models [21–23] and vision-language models [24–26], showing that models often struggle when exposed to irrelevant or misleading content. However, little attention has been given to the simpler but pervasive case of irrelevant audio that does not conflict with text yet degrades performance.

In this work, we ask: *To what extent are LALMs vulnerable to irrelevant audio during text reasoning tasks?* To address this question, we systematically evaluate the effect of non-informative audio on established text-based benchmarks [27–29]. In our setup, the textual input remains fixed while the audio channel varies with non-semantic perturbations such as silence, noise, or environmental sound [30]. We also conduct ablations, varying decoding temperature, audio duration, and noise amplitude, to study how interference strength scales across conditions.

We find that irrelevant audio lowers accuracy and alters model outputs, even when text alone is sufficient to solve the task. Surprisingly, even silence can interfere. Degradation intensifies with longer or louder noise, or with extended silence, and becomes especially pronounced at higher temperatures. Naive mitigation, such as prepending instructional phrases, proves ineffective. A simple self-consistency [31] approach offers partial mitigation but increases test-time cost [32], suggesting future work should seek more efficient solutions. These findings establish cross-modal interference as an essential evaluation axis and highlight the need for fusion strategies that preserve reasoning against irrelevant multimodal inputs.

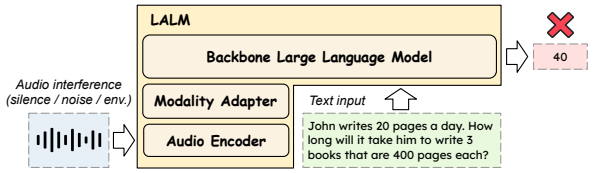

Fig. 1: Illustration of cross-modal interference: irrelevant audio disrupts text-only reasoning in LALMs

2. INVESTIGATING CROSS-MODAL INTERFERENCE

2.1. Problem Formulation

We analyze how large audio-language models (LALMs) handle tasks that rely only on text when the audio channel introduces irrelevant or distracting content. Figure 1 illustrates our problem setup, a text-only reasoning task with irrelevant audio signals such as silence, synthetic noise, or environmental sounds. Formally, the model generates predictions according to $\hat{y} = f_\theta(x_{\text{audio}}, x_{\text{text}})$, where f_θ denotes the modeling process, x_{audio} is the audio input, x_{text} is the text input, and $\hat{y}$ is the model output.

In the normal case, the model input is $(\emptyset, x_{\text{text}})$, where $\emptyset$ denotes the absence of audio. We introduce audio signals that should not affect task performance to study interference. These signals may include stretches of silence, bursts of synthetic noise, or unrelated real-world sounds such as rainfall, flowing water, or animal calls. Under interference, the model input becomes $(\delta_{\text{audio}}, x_{\text{text}})$, where δ_{audio} represents non-informative audio.

By fixing x_{text} and systematically varying δ_{audio}, we characterize how irrelevant audio influences model behavior, measuring accuracy degradation and shifts in generated outputs.

2.2. Experimental Setup

Benchmarks For tasks that rely entirely on textual reasoning, we evaluate models on three widely used benchmarks: GSM8K [27]

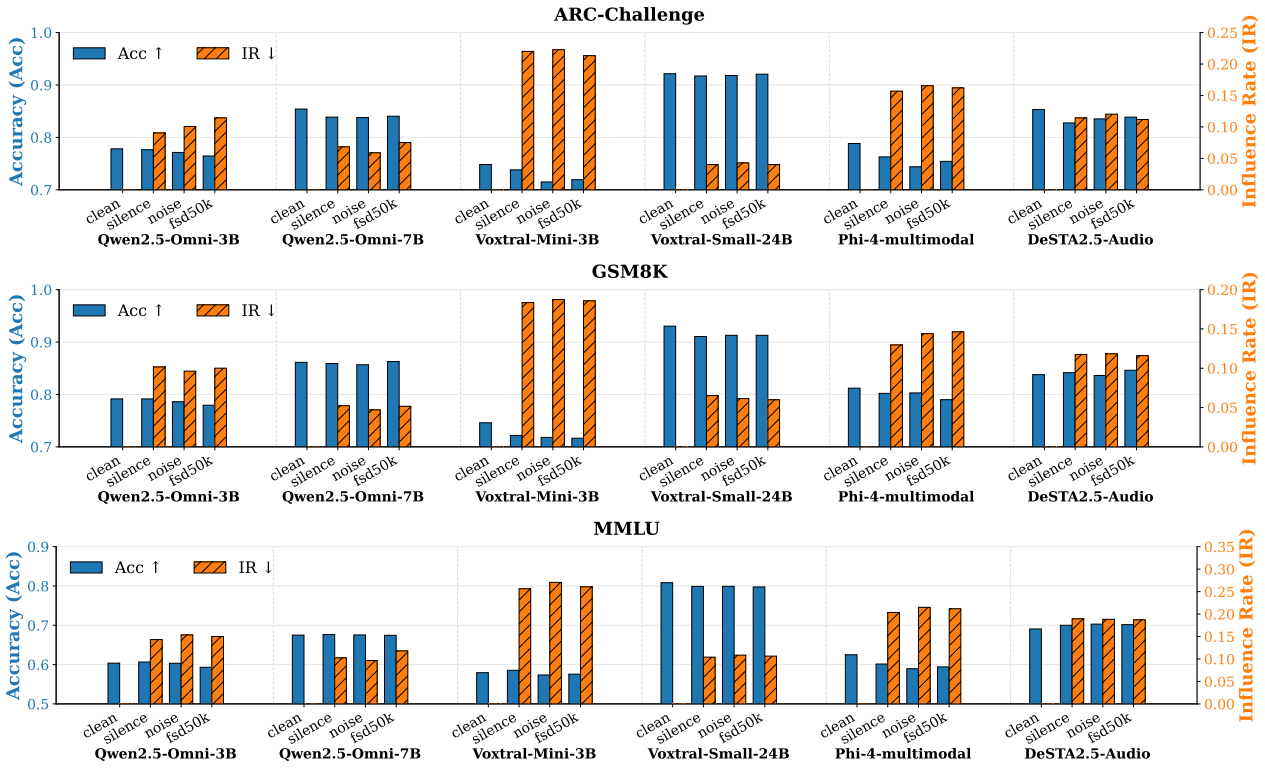

Fig. 2: Accuracy (Acc) and Influence Rate (IR) of LALMs under cross-modal interference across benchmarks.

for arithmetic reasoning, ARC-Challenge [28] for science question answering, and MMLU [29] for massive multi-task language understanding. To probe robustness, we introduce three types of audio interference: (i) five seconds of silence, (ii) five seconds of synthetic Gaussian noise at -40 dBFS, (iii) real-world audio samples drawn from FSD50K [30], a diverse dataset of music, environmental sounds, and sound effects. This setup allows us to study how irrelevant audio influences text-based reasoning tasks systematically.

Models We evaluate a diverse set of state-of-the-art open-source LALMs to assess robustness against irrelevant audio in text reasoning tasks. Our model set includes Qwen2.5-Omni-3B, Qwen2.5-Omni-7B [6], Phi-4-Multimodal-Instruct [5], Voxtral-Mini-3B, Voxtral-Small-24B [4], and DeSTA2.5-Audio [3]. These models differ in parameter size and architectural design, providing a representative view of current multimodal approaches. For most models, we adopt greedy decoding to isolate performance without the added variance of stochastic sampling, which makes the evaluation more stable under controlled perturbations. The only exception is the Voxtral series, where we follow the authors' recommended configuration and apply nucleus sampling with temperature set to 0.2 and top-p set to 0.95. We use vLLM [33] for inference, except DeSTA2.5-Audio, which uses the Transformers [34] package.

Metrics We adopt two evaluation metrics to assess robustness under irrelevant audio. First, we report accuracy, defined as the proportion of correctly answered samples relative to the total number of samples in the dataset: $\textit{Accuracy} = n_c/N$, where n_c denotes the number of correct predictions and N the total number of samples, accuracy reflects overall task performance under different interference conditions.

Second, we use the influence rate, a notion explored in other multimodal robustness studies [20], to quantify how often irrelevant audio changes model predictions. Let n_{ic} be the number of cases where a prediction flips from incorrect to correct, and n_{ci} the number of cases where it flips from correct to incorrect. We compute the influence rate as $\textit{Influence Rate} = (n_{ic} + n_{ci})/N$. A higher value indicates greater sensitivity to irrelevant audio, regardless of whether the change improves or harms accuracy.

2.3. Main Observation

Figure 2 shows that all evaluated LALMs are vulnerable to cross-modal interference. In the plots, *clean* denotes clean inputs without interference, *silence* adds silent audio, *noise* corresponds to Gaussian noise, and *fsd50k* represents a real-world audio sample from FSD50K. Across GSM8K, ARC-Challenge, and MMLU, introducing irrelevant audio consistently reduces accuracy compared to the clean setting. The absolute drops remain modest, yet they appear across all models and benchmarks, showing that even non-informative audio slightly harms performance.

The influence rate provides a clearer view of instability. Under interference, predictions frequently change, leading to noticeably higher influence rates across all conditions. Surprisingly, silence, often assumed to be a "neutral input", destabilizes outputs when persistent. Silence and Gaussian noise produce similar effects, while FSD50K sometimes increases the disruption, but not uniformly across tasks. This variation suggests that models do not react to one type of distractor consistently; instead, any non-informative audio has the potential to shift predictions. Accuracy alone fails to capture the full extent of instability, as outputs may change substantially even when task performance appears steady.

In addition, we observe a scaling effect when comparing models with the same architecture but different parameter sizes; larger models generally achieve better performance and exhibit reduced sensitivity to interference. In other words, larger LALMs are more robust

against irrelevant audio, showing minor accuracy drops and lower influence rates under the same perturbations. Beyond scaling, we also note that some models, such as Qwen2.5-Omni, display comparatively lower volatility than others of similar size. This suggests that factors beyond parameter count, such as training data and optimization design, may influence robustness, highlighting that resilience is shaped by multiple dimensions rather than scaling alone.

Another trend visible in Figure 2 is that the degree of interference differs across tasks. MMLU, which requires broad multi-domain reasoning, suffers from larger performance degradation and higher instability than more structured tasks like GSM8K arithmetic or ARC-Challenge science questions.

Irrelevant audio lowers accuracy, raises influence rates, and introduces uneven task- and model-dependent vulnerabilities. The persistence of this effect, even under greedy decoding, emphasizes the need for more robust multimodal fusion strategies that preserve both accuracy and stability in the presence of irrelevant inputs.

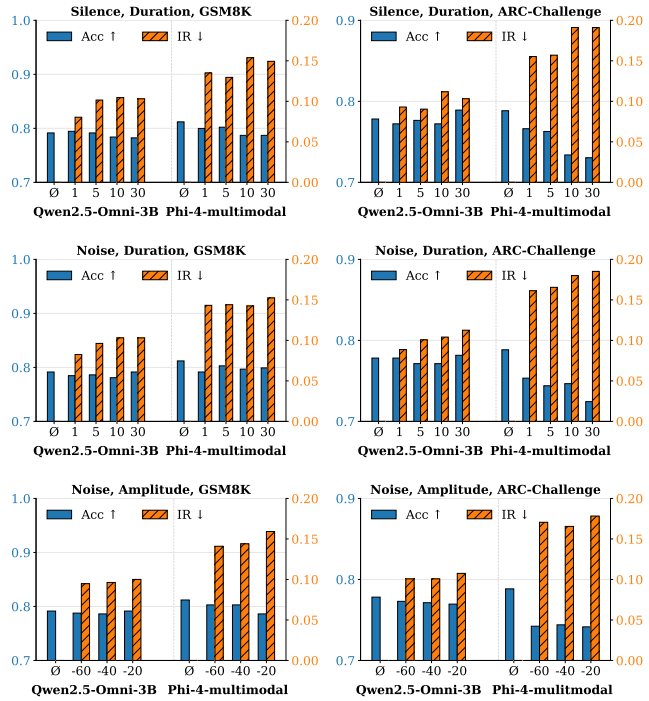

Fig. 3: Impact of varying silence and noise durations (1, 5, 10, 30 sec) and noise amplitudes (-60, -40, -20 dBFS) on GSM8K and ARC-Challenge.

3. ANALYSIS

3.1. Scaling Interference Effects

Duration of Audio Figure 3 illustrates the model's performance and influence rate across different durations of irrelevant audio. The x-axis indicates the duration of added audio, $\emptyset$ represents the clean baseline without any interference, while the values 1, 5, 10, and 30 denote durations of silence and Gaussian noise in seconds, respectively. As duration increases, accuracy consistently drops and influence rate rises, revealing that longer non-informative segments amplify cross-modal interference. Even silence, when extended, destabilizes reasoning, suggesting that persistent irrelevant signals are not ignored but gradually entangle in the fusion process. This scaling highlights that temporal persistence of irrelevant signals is a critical factor in LALMs' robustness.

Amplitude of Noise Figure 3 also demonstrates the effect of noise amplitude on model robustness. We injected 5-second Gaussian noise segments at three intensity levels, -60 dBFS, -40 dBFS, and -20 dBFS. The results indicate a clear trend in which model accuracy consistently declines as the amplitude increases and the influence rate rises. This pattern indicates that louder noise exacerbates cross-modal interference, making the model less stable and more prone to prediction shifts. Even when textual reasoning should dominate, high-intensity noise disrupts the fusion process, amplifying instability and eroding performance on benchmarks. In other words, stronger noise levels act as a more forceful distractor, directly undermining the reliability of LALMs in text reasoning tasks.

Table 1: Correctness change ratios across interference conditions. Results are reported on GSM8K, MMLU, and ARC-Challenge.

Model	Cond. Pair	GSM8K	MMLU	ARC
Qwen2.5-Omni-3B	silence/noise	0.057	0.078	0.048
	silence/fsd50k	0.086	0.119	0.079
	noise/fsd50k	0.084	0.120	0.077
Phi-4-multimodal	silence/noise	0.083	0.159	0.113
	silence/fsd50k	0.112	0.174	0.120
	noise/fsd50k	0.095	0.164	0.114

3.2. Comparative Analysis of Interference Types

Table 1 compares how model predictions change under different types of irrelevant audio by reporting the correctness change ratio. Each value reflects the proportion of samples where the model's prediction flipped from correct to incorrect, or vice versa, when moving between two interference conditions. For instance, a value of 0.057 for "silence/noise" in Qwen2.5-Omni-3B means that 5.7% of predictions changed correctness when the exact input text was paired with silence instead of Gaussian noise. For both Qwen2.5-Omni-3B and Phi-4-Multimodal, the silence–noise ratio is consistently lowest, indicating that these models treat silence and Gaussian noise as nearly equivalent. Additionally, we observe that the silence versus FSD50K comparison yields the highest ratio, with noise versus FSD50K falling in between, suggesting that FSD50K audio is perceptually closer to noise than to silence for these models.

3.3. Sensitivity to Temperature under Interference

Figure 4 shows how decoding temperature interacts with irrelevant audio during GSM8K reasoning. At low temperatures, both Qwen2.5-Omni-7B and Phi-4-Multimodal retain stable accuracy and relatively low influence rates, indicating that deterministic decoding reduces the impact of cross-modal interference. As temperature rises, however, accuracy begins to drop more steeply in the presence of silence, Gaussian noise, or FSD50K audio, revealing that stochastic sampling amplifies the destabilizing effect of irrelevant signals.

The influence rate highlights this compounding effect more clearly than accuracy alone. For both models, the influence rate is relatively low near greedy decoding but escalates sharply as temperature increases. This means that irrelevant audio reduces accuracy and actively makes predictions more volatile, flipping outcomes more frequently under higher sampling variance. The effect is particularly evident with real-world FSD50K inputs, where instability intensifies beyond what is observed with silence or synthetic noise.

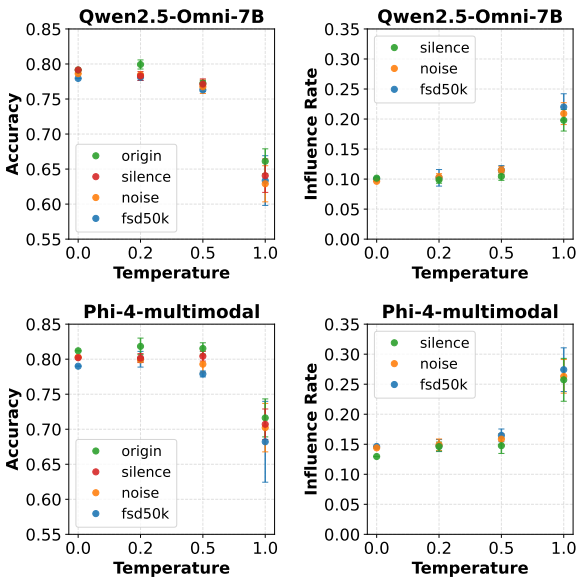

Fig. 4: Effect of decoding temperature on accuracy and influence rate under audio interference on GSM8K. Non-greedy results are averaged over 3 seeds with standard deviation.

Comparing models, Qwen2.5-Omni-7B sustains higher accuracy across all conditions and shows a slower rise in influence rate, indicating stronger resilience to interference. By contrast, Phi-4-Multimodal suffers sharper accuracy declines and a steeper increase in influence rate, suggesting greater susceptibility to stochasticity and irrelevant audio. These trends demonstrate that temperature is not a neutral hyperparameter under interference but a critical factor interacting with model design to shape robustness.

4. STRAIGHTFORWARD MITIGATION APPROACHES

4.1. Methodology

We evaluate two straightforward mitigation approaches to investigate whether simple strategies can alleviate the impact of irrelevant audio. The first approach is adding a mitigation prompt. Specifically, we prepend a short instructional phrase, *"Focus on the text or audio that contains useful information."* before each question. This prompt is designed to remind the model to pay explicit attention only to modalities that contribute to solving the task, thereby reducing the likelihood of being distracted by non-informative audio streams. The second approach is self-consistency. Instead of relying on a single decoding output, we generate 8 responses with sampling temperature 0.5 and aggregate the final answer by majority voting. This ensemble-style decoding reduces prediction volatility by smoothing out spurious shifts introduced by audio interference.

4.2. Results and Discussion

Table 2 compares the performance of the two mitigation strategies, prompting and self-consistency, under silence, Gaussian noise, and FSD50K interference. The influence rate is recalculated for each method relative to its own clean baseline.

The mitigation prompt yields only limited and inconsistent improvements. In some settings, it slightly lowers the influence rate, yet in others, it reduces accuracy or even amplifies instability. The variation across benchmarks and models shows that a single explicit instruction is insufficient to counteract cross-modal interference systematically.

Table 2: Comparing with several mitigation approach under different interference conditions.

Model	Condition	GSM8K		ARC-Challenge	
		Acc ↑	IR ↓	Acc ↑	IR ↓
Qwen2.5-Omni-3B	clean	0.7915	-	0.7782	-
	silence	0.7915	0.1016	0.7765	0.0904
	noise	0.7862	0.0963	0.7713	0.1007
	fsd50k	0.7794	0.1001	0.7645	0.1143
+ Prompt	clean	0.7779	-	0.7722	-
	silence	0.7817	0.1054	0.7773	0.0802
	noise	0.7809	0.1168	0.7645	0.0930
	fsd50k	0.7817	0.1114	0.7696	0.1084
+ Self-Consistency	clean	0.8552	-	0.8157	-
	silence	0.8529	0.0432	0.8029	0.0555
	noise	0.8514	0.0478	0.7986	0.0597
	fsd50k	0.8628	0.0576	0.8080	0.0691
Phi-4-multimodal	clean	0.8120	-	0.7884	-
	silence	0.8021	0.1296	0.7628	0.1570
	noise	0.8029	0.1440	0.7440	0.1655
	fsd50k	0.7900	0.1463	0.7543	0.1621
+ Prompt	clean	0.8188	-	0.7816	-
	silence	0.7900	0.1440	0.7619	0.1101
	noise	0.7892	0.1539	0.7543	0.1160
	fsd50k	0.7991	0.1365	0.7474	0.1263
+ Self-Consistency	clean	0.8825	-	0.8370	-
	silence	0.8688	0.0637	0.7961	0.1075
	noise	0.8688	0.0667	0.7739	0.1195
	fsd50k	0.8590	0.0720	0.7619	0.1280

In contrast, self-consistency consistently improves robustness relative to its baseline without mitigation. Accuracy increases across all interference types, and the influence rate decreases substantially. This approach stabilizes predictions and delivers more reliable outputs in the presence of irrelevant audio. The results demonstrate that aggregating multiple generations and selecting the majority answer effectively counteract instability introduced by audio interference.

In summary, mitigation prompts alone are insufficient to protect models against irrelevant audio, often yielding inconsistent or minor effects compared with their baseline. Self-consistency, on the other hand, reliably enhances both accuracy and robustness relative to its baseline, though at the cost of considerable computational overhead since multiple generations must be produced for each input.

5. CONCLUSION

Our study shows that irrelevant audio can interfere with how large audio-language models reason over text. Silence, noise, and environmental sounds disrupted performance, and the impact grew with longer duration, louder volume, and higher decoding temperatures. Even silence, often assumed neutral, proved disruptive, destabilizing outputs as much as synthetic noise. Larger models and specific architectures showed greater resilience, but none were fully robust. Mitigation experiments revealed that prompting was ineffective, whereas self-consistency improved stability but introduced substantial test-time compute overhead. These findings establish cross-modal interference as a key robustness challenge and call for more efficient fusion strategies to preserve reasoning quality in realistic multimodal settings.

6. REFERENCES

[1] OpenAI (2024), "Gpt-4o system card," *arXiv preprint arXiv:2410.21276*, 2024.

[2] Gheorghe Comanici et al., "Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities," *arXiv preprint arXiv:2507.06261*, 2025.

[3] Ke-Han Lu et al., "Desta2. 5-audio: Toward general-purpose large audio language model with self-generated cross-modal alignment," *arXiv preprint arXiv:2507.02768*, 2025.

[4] Alexander H Liu et al., "Voxtral," *arXiv preprint arXiv:2507.13264*, 2025.

[5] Abdelrahman Abouelenin et al., "Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras," *arXiv preprint arXiv:2503.01743*, 2025.

[6] Jin Xu et al., "Qwen2. 5-omni technical report," *arXiv preprint arXiv:2503.20215*, 2025.

[7] Arushi Goel et al., "Audio flamingo 3: Advancing audio intelligence with fully open large audio language models," *arXiv preprint arXiv:2507.08128*, 2025.

[8] Chien-yu Huang et al., "Dynamic-SUPERB phase-2: A collaboratively expanding benchmark for measuring the capabilities of spoken language models with 180 tasks," in *The Thirteenth International Conference on Learning Representations*, 2025.

[9] Chih-Kai Yang et al., "SAKURA: On the Multi-hop Reasoning of Large Audio-Language Models Based on Speech and Audio Information," in *Interspeech 2025*, 2025, pp. 1788–1792.

[10] S Sakshi et al., "MMAU: A massive multi-task audio understanding and reasoning benchmark," in *The Thirteenth International Conference on Learning Representations*, 2025.

[11] Ziyang Ma et al., "Mmar: A challenging benchmark for deep reasoning in speech, audio, music, and their mix," *arXiv preprint arXiv:2505.13032*, 2025.

[12] Qian Yang et al., "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)*, 2024.

[13] Bin Wang et al., "AudioBench: A universal benchmark for audio large language models," in *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2025.

[14] Yu-Xiang Lin et al., "A preliminary exploration with gpt-4o voice mode," *arXiv preprint arXiv:2502.09940*, 2025.

[15] Chih-Kai Yang, Neo S Ho, and Hung-yi Lee, "Towards holistic evaluation of large audio-language models: A comprehensive survey," *arXiv preprint arXiv:2505.15957*, 2025.

[16] Siddhant Arora et al., "On the landscape of spoken language models: A comprehensive survey," *arXiv preprint arXiv:2504.08528*, 2025.

[17] Nicholas Carlini and David Wagner, "Audio adversarial examples: Targeted attacks on speech-to-text," in *2018 IEEE security and privacy workshops (SPW)*. IEEE, 2018, pp. 1–7.

[18] Raghuveer Peri et al., "SpeechGuard: Exploring the adversarial robustness of multi-modal large language models," in *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.

[19] Guanyu Hou et al., "Evaluating robustness of large audio language models to audio injection: An empirical study," *arXiv preprint arXiv:2505.19598*, 2025.

[20] Cheng Wang et al., "When audio and text disagree: Revealing text bias in large audio-language models," *arXiv preprint arXiv:2508.15407*, 2025.

[21] Freda Shi et al., "Large language models can be easily distracted by irrelevant context," in *International Conference on Machine Learning*. PMLR, 2023, pp. 31210–31227.

[22] Cheng-Han Chiang and Hung-yi Lee, "Over-reasoning and redundant calculation of large language models," in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2024.

[23] Yanbo Wang et al., "Breaking focus: Contextual distraction curse in large language models," in *Will Synthetic Data Finally Solve the Data Access Problem?*, 2025.

[24] Ming Liu et al., "On the robustness of multimodal language model towards distractions," *arXiv preprint arXiv:2502.09818*, 2025.

[25] Ailin Deng et al., "Words or vision: Do vision-language models have blind faith in text?," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 3867–3876.

[26] Rui Cai et al., "Diagnosing and mitigating modality interference in multimodal large language models," *arXiv preprint arXiv:2505.19616*, 2025.

[27] Karl Cobbe et al., "Training verifiers to solve math word problems," *arXiv preprint arXiv:2110.14168*, 2021.

[28] Peter Clark et al., "Think you have solved question answering? try arc, the ai2 reasoning challenge," *arXiv preprint arXiv:1803.05457*, 2018.

[29] Dan Hendrycks et al., "Measuring massive multitask language understanding," in *International Conference on Learning Representations*, 2021.

[30] Eduardo Fonseca et al., "Fsd50k: an open dataset of human-labeled sound events," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 829–852, 2021.

[31] Xuezhi Wang et al., "Self-consistency improves chain of thought reasoning in language models," in *The Eleventh International Conference on Learning Representations*, 2023.

[32] Charlie Snell et al., "Scaling llm test-time compute optimally can be more effective than scaling model parameters," *arXiv preprint arXiv:2408.03314*, 2024.

[33] Woosuk Kwon et al., "Efficient memory management for large language model serving with pagedattention," in *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

[34] Thomas Wolf et al., "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 2020, Association for Computational Linguistics.