AUDIOGENIE-REASONER: A TRAINING-FREE MULTI-AGENT FRAMEWORK FOR COARSE-TO-FINE AUDIO DEEP REASONING

Yan Rong[1], *Chenxing Li*[2], *Dong Yu*[2], *Li Liu*[1]*

[1] The Hong Kong University of Science and Technology (Guangzhou)
[2] Tencent AI Lab

ABSTRACT

Audio deep reasoning is a challenging task that requires expert-level perception, multi-step logical inference, and the integration of contextual knowledge. However, existing models suffer from a gap between audio perception and reasoning abilities due to the lack of training data with explicit reasoning chains and the absence of mechanisms for active exploration and iterative refinement. To address these challenges, we propose **AudioGenie-Reasoner (AGR)**, the **first** unified training-free multi-agent system that coordinates perception and reasoning over an evolving chain of textual evidence. Our key idea is a paradigm shift that transforms audio deep reasoning into complex text understanding task from a new perspective, thereby unlocking the full potential of large language models. Specifically, the design of AGR mimics the human coarse-to-fine cognitive process. It first transforms the input audio into a coarse text-based document. Then, we design a novel proactive iterative document refinement loop, featuring tool-augmented routes and specialized agents, to continuously search for missing information and augment the evidence chain in a coarse-to-fine manner until sufficient question-related information is gathered for making final predictions. Experimental results show that AGR achieves state-of-the-art (SOTA) performance over existing open-source audio deep reasoning models across various benchmarks. The code will be made publicly available.

Index Terms— Audio Deep Reasoning, Multi-Agent, Training-Free, Large Language Models, Iterative Refinement

1. INTRODUCTION

Audio deep reasoning [1, 2, 3] is a challenge task in audio understanding, requiring expert-level perception, multi-step logical inference, and the integration of contextual knowledge to interpret complex acoustic scenes. This technology has many applications in our daily life, such as embodied intelligence [4] and autonomous systems [5].

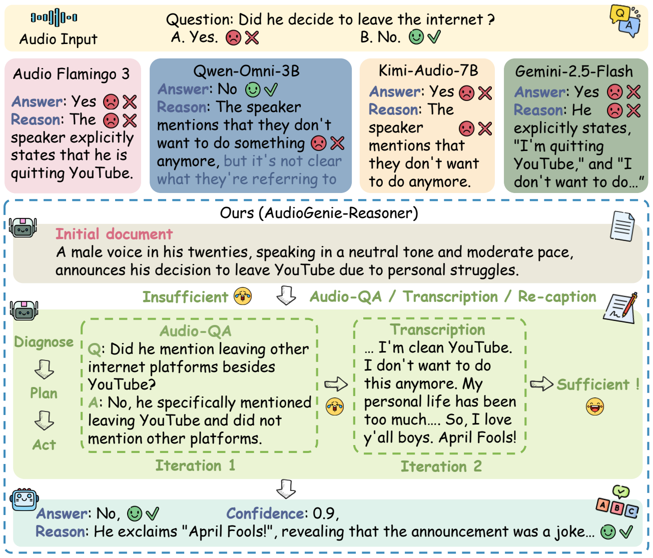
Audio Input
Question: Did he decide to leave the internet ?
A. Yes. ☹ ✗
B. No. ☺ ✓
Audio Flamingo 3
Answer: Yes ☹ ✗
Reason: The ☹ ✗ speaker explicitly states that he is quitting YouTube.
Qwen-Omni-3B
Answer: No ☺ ✓
Reason: The speaker mentions that they don't want to do something ☹ ✗ anymore, but it's not clear what they're referring to
Kimi-Audio-7B
Answer: Yes ☹ ✗
Reason: The speaker ☹ ✗ mentions that they don't want to do anymore.
Gemini-2.5-Flash
Answer: Yes ☹ ✗
Reason: He ☹ ✗ explicitly states, "I'm quitting YouTube," and "I don't want to do..."
Ours (AudioGenie-Reasoner)
Initial document
A male voice in his twenties, speaking in a neutral tone and moderate pace, announces his decision to leave YouTube due to personal struggles.
Insufficient
Audio-QA / Transcription / Re-caption
Diagnose
Plan
Act
Audio-QA
Q: Did he mention leaving other internet platforms besides YouTube?
A: No, he specifically mentioned leaving YouTube and did not mention other platforms.
Iteration 1
Transcription
... I'm clean YouTube. I don't want to do this anymore. My personal life has been too much.... So, I love y'all boys. April Fools!
Iteration 2
Sufficient !
Answer: No, ☺ ✓
Confidence: 0.9,
Reason: He exclaims "April Fools!", revealing that the announcement was a joke... ☺ ✓

Fig. 1. Performance comparison of AudioGenie-Reasoner with other audio reasoning models. Our framework excels in providing correct answers and valid reasoning.

In the literature, great progress in audio reasoning has been achieved by prior works [6, 7, 8, 9]. However, audio deep reasoning remains a significant challenge. The recently proposed audio deep reasoning benchmark, MMAR [2], reveals the poor performance of existing audio models. On this challenging benchmark, many open-source models fail to achieve an accuracy better than random guessing, reflecting a gap between the abilities of audio perception and the cognitive reasoning. This gap stems from two fundamental challenges: **Firstly**, existing models are hindered by the lack of training data with explicit reasoning chains. Constructing high-quality, step-by-step reasoning annotations for audio is resource-intensive. Lacking this fine-grained supervision, most Audio Large Language Models (ALLMs) are trained on simpler objectives like audio-text alignment [10] or direct question-answering [11]. In particular, when it comes to complex scenarios with mixed audio sources (*e.g.*, speech, music, and sound effects), their reasoning capabilities degrade sharply. **Secondly**, current methods lack a mechanism for active exploration and iterative refinement. Models typically function as passive information receivers, generating answers based on a single pass of perceptual results. This

*Corresponding Author: avrillliu@hkust-gz.edu.cn

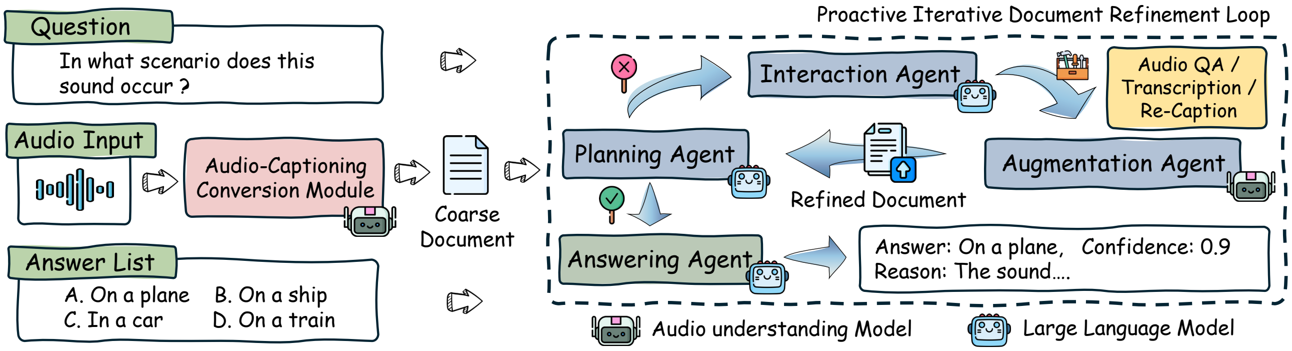
Question
In what scenario does this sound occur ?
Audio Input
Audio-Captioning Conversion Module
Coarse Document
Answer List
A. On a plane
B. On a ship
C. In a car
D. On a train
Proactive Iterative Document Refinement Loop
Interaction Agent
Audio QA / Transcription / Re-Caption
Planning Agent
Refined Document
Augmentation Agent
Answering Agent
Answer: On a plane, Confidence: 0.9
Reason: The sound....
Audio understanding Model
Large Language Model

Fig. 2. The multi-agent architecture of AudioGenie-Reasoner. Specialized agents for planning, interaction, and augmentation collaborate within an iterative loop to refine a coarse initial caption into an evolving textual evidence chain.

static, single-pass process prevents them from diagnosing evidence gaps, planning to acquire missing information, or progressively deepening their understanding. As a result, they are ill-equipped to handle complex problems that require multi-step and in-depth analysis.

To address above two challenges, we propose **AudioGenie-Reasoner (AGR)**, a novel unified training-free multi-agent system (MAS) that coordinates perception and reasoning over an evolving chain of textual evidence. Our core design mimics the human coarse-to-fine cognitive process: forming an initial general understanding, conducting a detailed examination of relevant cues based on the specific query, and finally drawing a conclusion from sufficient evidence. Specifically, **for the first challenge**, instead of directly training heavy audio-reasoning models, we introduce a paradigm shift that transforms audio deep reasoning into a complex text understanding task. This transformation decouples perception from cognition, elegantly bypassing the need for vast audio-specific reasoning data and unlocking the full potential of Large Language Models (LLMs). **For the second challenge**, instead of a conventional single-pass pipeline, we introduce a proactive iterative document refinement loop, driven by tool-augmented routes and specialized agents. This process empowers the model to dynamically find the potential missing information and augment these information within language space. Through this "diagnose-plan-act" loop, the model is transformed from a passive information receiver into an active, self-improving investigator.

In summary, the main contributions of this work are as follows: **(1)** A unified training-free MAS, named AudioGenie-Reasoner, which coordinates perception and reasoning over an evolving chain of textual evidence, is proposed. To the best of our knowledge, this is the first exploration of MAS in audio deep reasoning. **(2)** We establish a coarse-to-fine cognitive framework that transforms audio reasoning into a text understanding task, featuring a novel proactive iterative document refinement loop to dynamically search for missing information and augment the evidence chain. **(3)** Experimental results show that AGR achieves SOTA performance over existing open-source models across various audio deep reasoning benchmarks.

2. OUR METHOD

Our framework's design is founded on two core innovations. First, we introduce a paradigm shift that transforms the audio reasoning problem into a text-based understanding task, thereby decoupling perception from cognition. Second, we design a proactive multi-agent loop for iterative evidence refinement, turning the system into an active investigator. The overall architecture is illustrated in Figure 2.

2.1. Paradigm Shift: From Audio Reasoning to Text Understanding

Instead of attempting to build a data-hungry audio reasoning model, we transform the audio reasoning task into a complex text understanding problem. This is achieved by initially converting the raw input audio A into a coarse-grained textual document D_0:

$$D_0 = \mathcal{F}_{\text{caption}}(A), \tag{1}$$

where $\mathcal{F}_{\text{caption}}(\cdot)$ is an audio-captioning module implemented with a powerful ALLM.

This initial transformation is the foundation of our paradigm shift. It decouples the system's perceptual abilities, which are handled by the ALLM, from its cognitive reasoning, which is governed by LLM-based agents in the subsequent steps. By doing so, we elegantly bypass the need for specialized audio-reasoning datasets and instead unlock the vast, pre-existing reasoning capabilities of LLMs. The resulting document, D_0, serves as the initial state of an evolving evidence chain, forming the textual foundation upon which all subsequent reasoning and refinement will be performed.

2.2. Proactive Iterative Document Refinement Loop

To bridge the gap between the coarse initial description and the fine-grained details required for complex queries, we introduce a proactive iterative refinement loop. This loop is coordinated by a team of specialized agents that collaborate

Table 1. Comparison with SOTA methods on MMAU-mini. The '/' separates results from raw outputs *vs.* those after GPT-4o post-processing (see Implementation Details). Best performances are highlighted in bold, while second-best are underlined.

Methods	Sound	Music	Speech	Easy	Medium	Hard	Avg
Random Guess	49.25 / 49.25	30.24 / 30.24	39.94 / 39.94	33.93 / 33.93	43.70 / 43.70	36.44 / 36.44	39.80 / 39.80
Gemini-2.5-Flash [12]	74.77 / 76.58	65.27 / 65.57	72.97 / 75.58	64.29 / 65.62	75.93 / 76.66	66.10 / 70.19	71.00 / 71.90
Gemini-2.0-Flash [13]	73.27 / 73.27	64.97 / 64.97	**78.38 / 78.38**	62.95 / 62.95	**77.22 / 77.22**	**69.49 / 69.49**	72.20 / 72.20
Gemini-2.0-Flash-Lite [14]	69.97 / 69.97	65.27 / 65.27	74.17 / 74.47	60.27 / 60.27	74.07 / 74.07	69.07 / 69.48	69.80 / 69.90
MiDashengLM-7B [15]	66.37 / 69.67	58.98 / 58.98	61.56 / 62.16	53.12 / 53.12	68.89 / 70.37	55.93 / 58.05	62.30 / 63.60
Audio Flamingo 3 [7]	74.76 / **76.88**	60.18 / 61.08	60.96 / 63.06	58.04 / 59.82	70.19 / 71.30	61.02 / 63.98	65.30 / 67.00
Audio Flamingo 3 (T) [7]	69.97 / 74.47	59.28 / **67.37**	44.74 / 61.26	56.70 / 63.84	61.85 / 74.07	50.42 / 56.78	58.00 / 67.70
Audio-Reasoner [6]	32.13 / 66.97	41.02 / 63.77	34.23 / 57.06	43.75 / 61.61	30.51 / 64.81	34.81 / 58.47	35.80 / 62.60
Qwen2.5-Omni-3B [16]	73.57 / 73.87	60.78 / 60.78	63.66 / 64.56	57.14 / 57.14	70.93 / 71.30	63.14 / 63.98	66.00 / 66.40
Audio Flamingo 2-0.5B [17]	26.43 / 47.15	17.96 / 35.93	15.32 / 27.93	24.11 / 36.61	16.10 / 38.33	19.81 / 34.32	19.90 / 37.00
Audio Flamingo 2-1.5B [17]	42.34 / 50.15	35.63 / 46.71	34.53 / 37.54	36.16 / 41.52	34.75 / 48.33	39.26 / 39.83	37.50 / 44.80
Audio Flamingo 2-3B [17]	62.46 / 63.96	50.60 / 55.09	39.34 / 47.15	49.11 / 52.23	53.52 / 58.70	46.19 / 50.85	50.80 / 55.40
Kimi-Audio-7B-Instruct [8]	59.46 / 74.17	42.51 / 58.38	61.56 / 66.07	44.64 / 56.70	59.26 / 71.48	52.97 / 63.14	54.50 / 66.20
AudioGenie-Reasoner	**75.08** / 75.08 (+8.7) / (+5.4)	**66.17** / 66.17 (+7.2) / (+7.2)	76.58 / 76.58 (+15.0) / (+14.4)	**69.20 / 69.20** (+16.1) / (+16.1)	76.67 / 76.67 (+7.8) / (+6.3)	66.53 / 66.53 (+10.6) / (+8.5)	**72.60 / 72.60** (+10.3) / (+9.0)

to progressively enrich the initial document into a comprehensive evidence chain. At its core, the loop operates iteratively: it first assesses the current evidence, then plans and executes actions to augment it with missing information via tool-augmented routes. This process repeats until the evidence is deemed sufficient for a confident answer.

Planning Agent. Each iteration begins with the planning agent $\mathcal{F}_{\text{plan}}(\cdot)$, which assesses if the current document contains sufficient evidence to confidently answer the question:

$$(s, H_{i+1}) = \mathcal{F}_{\text{plan}}(Q, L, D_i, H_i), \tag{2}$$

where Q is the question, L is the answer list, D_i is the document at iteration i, and H_i is the analysis history. The agent returns a status flag $s \in \{\text{Sufficient}, \text{Insufficient}\}$. If the evidence is insufficient, the history is updated to H_{i+1} with an analysis of the information gap.

Interaction Agent. If the status s is insufficient, the interaction agent $\mathcal{F}_{\text{interact}}(\cdot)$ formulates a plan to acquire the missing information:

$$P = \mathcal{F}_{\text{interact}}(D_i, H_{i+1}), \tag{3}$$

where P is a structured augmentation plan. The plan outlines one of three tool-based actions: audio question-answering, guided re-captioning, or automatic speech recognition.

Augmentation Agent. The augmentation agent $\mathcal{F}_{\text{Aug}}(\cdot)$ executes the plan P by invoking the specified tool to generate new evidence $E_{\text{new}} = \mathcal{F}_{\text{Aug}}(P)$ and integrate it into the document:

$$D_{i+1} = D_i \oplus E_{\text{new}}, \tag{4}$$

where the $\oplus$ operator denotes the integration of E_{new} into the existing document D_i. The enriched document D_{i+1} is then passed back to the planning agent for the next iteration.

Answering Agent. Once the iterative refinement loop concludes (*i.e.*, when $s =$ Sufficient) or the maximum number of iterations is reached, the answer agent $\mathcal{F}_{\text{answer}}(\cdot)$ generates the final output from the enriched document D_f:

$$(A^*, S_c, R) = \mathcal{F}_{\text{answer}}(D_f, Q, L), \tag{5}$$

where A^*, S_c, and R represent the final selected answer, the associated confidence score, and a detailed textual rationale explaining the reasoning process, respectively.

3. EXPERIMENTAL RESULTS

3.1. Experimental Setup

Datasets. We evaluate our framework on two well-known audio deep reasoning benchmarks: MMAU-mini [3] and MMAR [2]. MMAU-mini consists of 1,000 closed-form questions covering three audio types: sound, music, and speech. MMAR is a more challenging benchmark that includes not only single audio types but also various mixtures of them. Since the audio data for MMAR is not directly provided, we successfully collected 905 samples after filtering for inaccessible data due to issues like expired links.

Implementation Details. We select MiDashengLM-7B [15] and GPT-4o-2024-08-06 [18] as the ALLM and LLM in our framework, respectively. Whisper-Turbo [19] is employed as the transcription model in our tool-based actions. The max number of iterations is set to three. All experiments are conducted on a single NVIDIA A800 GPU. For the evaluation metric, we follow the methodology of MMAU and MMAR, comparing the model's prediction with the ground truth using regular expressions and string matching. To handle cases where some ALLMs produce semantically correct but improperly formatted answers, **we use GPT-4o-2024-08-06 to post-process the raw outputs**. This step normalizes the generated text by mapping it to the corresponding answer in the predefined list (*e.g.*, mapping a free-form response like "The final answer is C" to the third item in the choice list), ensuring a fair and accurate evaluation.

Table 2. Comparison with SOTA methods on MMAR. The '/' separates results from raw outputs *vs.* those after GPT-4o post-processing (see Implementation Details). So, Mu, and Sp denote Sound, Music, and Speech, respectively. Best performances are highlighted in bold, while second-best are underlined.

Methods	Sound	Music	Speech	So-Mu	So-Sp	Mu-Sp	Sn-Mu-Sp	Avg
Random Guess	27.74 / 27.74	24.58 / 24.58	35.38 / 35.38	18.18 / 18.18	24.63 / 24.63	28.00 / 28.00	13.64 / 13.64	28.18 / 28.18
Gemini-2.5-Flash [12]	**56.13** / **57.42**	39.11 / 48.04	**76.92** / **79.23**	45.45 / 45.45	73.40 / **75.37**	**68.00** / **74.67**	54.55 / 54.55	63.43 / **67.07**
Gemini-2.0-Flash [13]	52.90 / 52.90	**53.07** / **53.07**	71.15 / 71.15	**100** / **100**	**73.89** / 73.89	66.67 / 68.00	**63.64** / **63.64**	**64.86** / 64.97
Gemini-2.0-Flash-Lite [14]	52.26 / 52.26	45.25 / 45.25	66.54 / 66.92	72.73 / 72.73	66.01 / 66.01	66.67 / 66.67	54.55 / 54.55	59.56 / 59.67
MiDashengLM-7B [15]	43.23 / 43.87	40.22 / 40.22	51.15 / 51.15	18.18 / 18.18	45.81 / 45.81	57.33 / 57.33	36.36 / 36.36	46.19 / 46.30
Audio Flamingo 3 [7]	45.81 / 47.10	31.84 / 32.40	53.85 / 54.23	27.27 / 27.27	46.31 / 47.29	54.67 / 56.00	45.45 / 45.45	45.97 / 46.74
Audio Flamingo 3 (T) [7]	41.94 / 52.26	25.70 / 32.40	39.23 / 49.62	18.18 / 27.27	44.83 / 54.19	42.67 / 52.00	18.18 / 31.82	37.79 / 47.18
Audio-Reasoner [6]	26.45 / 39.35	20.67 / 35.75	24.23 / 39.62	36.36 / 54.55	28.57 / 44.33	38.67 / 48.00	27.27 / 31.82	26.30 / 40.55
Qwen2.5-Omni-3B [16]	50.97 / 50.97	46.37 / 46.37	48.85 / 48.85	27.27 / 27.27	51.72 / 51.72	61.33 / 61.33	45.45 / 45.45	50.06 / 50.06
Audio Flamingo 2-0.5B [17]	11.61 / 21.94	6.70 / 16.20	13.85 / 22.69	9.09 / 9.09	15.27 / 26.11	17.33 / 22.67	18.18 / 22.73	12.71 / 21.88
Audio Flamingo 2-1.5B [17]	21.29 / 25.81	20.11 / 29.05	24.62 / 28.85	9.09 / 9.09	18.23 / 21.67	26.67 / 30.67	27.27 / 31.82	21.77 / 26.74
Audio Flamingo 2-3B [17]	39.35 / 43.23	27.37 / 31.84	36.15 / 38.85	36.36 / 36.36	28.57 / 30.05	29.33 / 30.67	31.82 / 36.36	32.60 / 35.47
Kimi-Audio-7B-Instruct [8]	49.03 / 50.32	32.96 / 37.99	52.69 / 56.15	18.18 / 36.36	56.65 / 61.58	52.00 / 60.00	36.36 / 45.45	48.18 / 52.60
AudioGenie-Reasoner	49.68 / 49.68 (+6.5) / (+5.8)	43.26 / 43.26 (+3.0) / (+3.0)	69.23 / 69.23 (+18.1) / (+18.1)	45.45 / 45.45 (+27.3) / (+27.3)	64.53 / 64.53 (+18.7) / (+18.7)	65.33 / 65.33 (+8.0) / (+8.0)	59.09 / 59.09 (+22.7) / (+22.7)	58.85 / 58.85 (+12.7) / (+12.6)

Table 3. Results of ablation studies on different model components. Best performances are highlighted in bold, while second-best are underlined.

ALLM	LLM	Whisper	MMAU	MMAR
Our Framework(w/ Proactive Iterative Document Refinement Loop)				
MiDashengLM-7B [15]	GPT-3.5-turbo [20]	Turbo	67.30	49.72
MiDashengLM-7B [15]	GPT-4o [18]	Turbo	72.60	**58.85**
Audio Flamingo 3 [7]	GPT-4o [18]	-	69.40	55.36
Audio Flamingo 3 [7]	GPT-4o [18]	Turbo	**74.10**	55.80
Audio Flamingo 3 [7]	GPT-4o [18]	Large	71.80	57.24
Qwen2.5-Omni-3B [16]	GPT-4o [18]	Turbo	70.64	56.35
Our Framework (w/o Proactive Iterative Document Refinement Loop)				
MiDashengLM-7B [15]	GPT-4o [18]	/	63.40	41.88
Audio Flamingo 3 [7]	GPT-4o [18]	/	68.90	44.09
Qwen2.5-Omni-3B [16]	GPT-4o [18]	/	66.70	45.41

3.2. Main Results

Comparison with SOTA Methods. Table 1 presents a comparison of AGR with SOTA audio reasoning methods on MMAU-mini. AGR not only surpasses open-source models but also outperforms the proprietary Gemini model, achieving the best performance. On MMAR (see Table 2), AGR significantly outperforms all open-source models and achieves results comparable to Gemini-2.0-Flash-Lite [14]. Besides, our multi-agent framework yields substantial performance gains over direct inference with MiDahengLM, particularly on reasoning tasks involving speech and mixed audio types.

Ablation Studies. The results of ablation studies are shown in Table 3. A significant performance drop is observed when replacing GPT-4o [18] with GPT-3.5-turbo [20] in our iterative document refinement loop, particularly on the MMAR dataset. Since the LLM serves as the planning, interaction, and answering agent, its reasoning capability is a decisive factor in the final performance. We also replace our ALLM with Audio Flamingo 3 [7] (in configurations with and without the Whisper) and Qwen2.5-Omni-3B [16], which results in only slight performance variations. We infer the reason is that current ALLMs have comparable perceptual abilities, but their reasoning capabilities still differ significantly. Furthermore, the removal of our iterative document refinement loop causes a consistent performance drop for all tested ALLMs, most notably on the MMAR dataset. This confirms the effectiveness of our loop, which allows the model to continuously reflect on existing information, complete any missing evidence, and build a comprehensive evidence chain to support the final reasoning result.

Table 4. Performance of different rounds on MMAU-mini and MMAR. Best results are highlighted in bold.

Dataset	Iteration 0	Iteration 1	Iteration 2	Iteration 3	Iteration 4
MMAU-mini	68.90	72.90	**73.80**	71.80	71.90
MMAR	44.09	54.59	56.35	**57.24**	57.02

Effects of Iterative Rounds. We analyze the impact of the number of iterative rounds on model performance in Table 4. The initial iteration yields the most significant performance gain on both datasets, as it recovers the most critical missing information, validating our framework's effectiveness. Performance peaks at two rounds on MMAU-mini and at three rounds on MMAR, consistent with MMAR's higher complexity and need for deeper exploration. With four rounds, performance drops on both datasets, likely because extra rounds introduce noise and irrelevant cues.

4. CONCLUSION AND DISCUSSION

In this work, we proposed AGR, a unified, training-free MAS that transforms audio deep reasoning into a text-based task. By decoupling perception from reasoning and employing a proactive iterative refinement loop, our framework synergizes the perceptual strengths of ALLMs with the advanced reasoning capabilities of LLMs. Experiments validate the effectiveness of this "diagnose-plan-act" strategy, showing significant performance gains, particularly on high-level semantic tasks like speaker and content analysis. Future work will focus on enhancing signal-level reasoning by developing more specialized evidence generators for low-level acoustic cues.

5. ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (No. 62471420), GuangDong Basic and Applied Basic Research Foundation (2025A1515012296), and CCF-Tencent Rhino-Bird Open Research Fund.

6. REFERENCES

[1] Yi Su, Jisheng Bai, Qisheng Xu, Kele Xu, and Yong Dou, "Audio-language models for audio-centric tasks: A survey," *arXiv preprint arXiv:2501.15177*, 2025.

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

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

[4] Xiulong Liu, Sudipta Paul, Moitreya Chatterjee, and Anoop Cherian, "Caven: An embodied conversational agent for efficient audio-visual navigation in noisy environments," in *Proceedings of the AAAI conference on artificial intelligence*, 2024, vol. 38, pp. 3765–3773.

[5] Ming Nie, Renyuan Peng, Chunwei Wang, Xinyue Cai, Jianhua Han, Hang Xu, and Li Zhang, "Reason2drive: Towards interpretable and chain-based reasoning for autonomous driving," in *European Conference on Computer Vision*, 2024, pp. 292–308.

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

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

[8] Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, et al., "Kimi-audio technical report," *arXiv preprint arXiv:2504.18425*, 2025.

[9] Sreyan Ghosh, Sonal Kumar, Ashish Seth, Chandra Kiran Reddy Evuru, Utkarsh Tyagi, S Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha, "Gama: A large audio-language model with advanced audio understanding and complex reasoning abilities," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 6288–6313.

[10] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang, "Clap learning audio concepts from natural language supervision," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023, pp. 1–5.

[11] Samuel Lipping, Parthasaarathy Sudarsanam, Konstantinos Drossos, and Tuomas Virtanen, "Clotho-aqa: A crowdsourced dataset for audio question answering," in *IEEE European Signal Processing Conference*, 2022, pp. 1140–1144.

[12] Google Cloud, "Gemini 2.5 flash — generative ai on vertex ai," 2025.

[13] Google Cloud, "Gemini 2.0 flash — generative ai on vertex ai," 2025.

[14] Google Cloud, "Gemini 2.0 flash-lite — generative ai on vertex ai," 2025.

[15] Heinrich Dinkel, Gang Li, Jizhong Liu, Jian Luan, Yadong Niu, Xingwei Sun, Tianzi Wang, Qiyang Xiao, Junbo Zhang, and Jiahao Zhou, "Midashenglm: Efficient audio understanding with general audio captions," *arXiv preprint arXiv:2508.03983*, 2025.

[16] Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al., "Qwen2. 5-omni technical report," *arXiv preprint arXiv:2503.20215*, 2025.

[17] Sreyan Ghosh, Zhifeng Kong, Sonal Kumar, S Sakshi, Jaehyeon Kim, Wei Ping, Rafael Valle, Dinesh Manocha, and Bryan Catanzaro, "Audio flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities," in *Forty-second International Conference on Machine Learning*, 2025.

[18] OpenAI, "Gpt-4o system card," https://openai.com/index/gpt-4o-system-card, 2024, Model snapshot used: gpt-4o-2024-08-06.

[19] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*, 2023, pp. 28492–28518.

[20] OpenAI, "Gpt-3.5 turbo fine-tuning and api updates," https://openai.com/index/gpt-3-5-turbo-fine-tuning-and-api-updates, Aug. 2023.