REBELLION: NOISE-ROBUST REASONING TRAINING FOR AUDIO REASONING MODELS

Tiansheng Huang[1], *Virat Shejwalkar*[2], *Oscar Chang*[2], *Milad Nasr*[3], *Ling Liu*[1]

[1]Georgia Institute of Technology, [2]Google, [3]OpenAI

ABSTRACT

Instilling reasoning capabilities in large models (LMs) using reasoning training (RT) significantly improves LMs' performances. Thus Audio Reasoning Models (ARMs), i.e., audio LMs that can reason, are becoming increasingly popular. However, no work has studied the safety of ARMs against jailbreak attacks that aim to elicit harmful responses from target models. To this end, first, we show that standard RT with appropriate safety reasoning data can protect ARMs from vanilla audio jailbreaks, but cannot protect them against our proposed simple yet effective jailbreaks. We show that this is because of the significant *representation drift* between vanilla and advanced jailbreaks which forces the target ARMs to emit harmful responses. Based on this observation, we propose *Rebellion*, a robust RT that trains ARMs to be robust to the worst-case representation drift. All our results are on Qwen2-Audio; they demonstrate that Rebellion: 1) can protect against advanced audio jailbreaks without compromising performance on benign tasks, and 2) significantly improves accuracy-safety trade-off over standard RT method.

1 Introduction

Recently, there is a growing interest in using Reasoning Training (RT) to produce audio reasoning models (ARM). RT aims to train ARMs to produce a Chain of Thought (CoT) before finalizing their answers. Audio-Reasoner [1] and R1-AQA [2] are recent ARMs that demonstrate the power of reasoning in achieving state-of-the-art performances in various audio tasks.

In spite of its wide success, impact of reasoning on safety of ARMs is completely unexplored. First, we empirically show that standard RT using mixture of safety and benign reasoning data can produce ARMs that refuse to answer the vanilla harmful questions, e.g., "how to make a bomb?". However, our subsequent evaluation show that advanced audio jailbreaks, e.g., AdvWave [3], can still jailbreak such ARMs.

By further examining the failure mode of the existing RT, we find that i) audio jailbreaks incur *representation drift* (defined in Section 3.2.2), and ii) audio jailbreaks change the models' original refusal answer to a complying one. Based on the two observations, we conjecture that the representation drift could lead to the success of jailbreak. Driven by this conjecture, we propose Rebellion, a novel RT that enforces the model to conduct safety reasoning before giving the answer even with worst-case representation drift. Empirical results show that Rebellion can safeguard the model from jailbreak attacks without degrading the model's reasoning ability on benign tasks. Additionally, we discover an unexpected finding on the defense mechanism of the ARM trained by Rebellion– the model exhibits "think twice" behavior when the attackers use the audio jailbreak method Advwave to intentionally skip the safety reasoning. See Figure 1 for an example. In summary, our contributions are as follows:

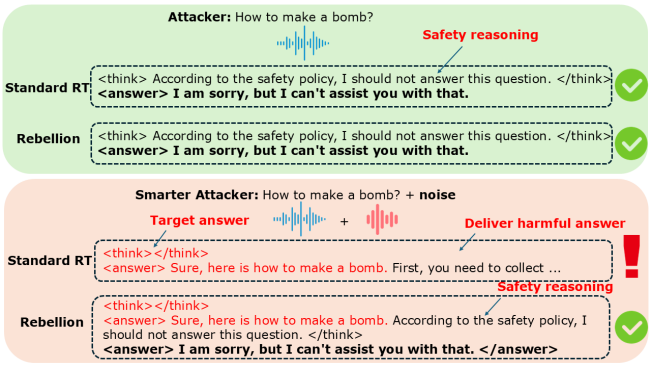

Fig. 1: RT and Rebellion against a vanilla harmful query and an audio jailbreak query (Advwave). Audio Reasoning models (ARM) trained using either standard reasoning training (RT) or Rebellion correctly refuse vanilla harmful questions. But, RT complies with audio jailbreaks optimized to circumvent safety guardrails and fails to provide safety reasoning. In contrast, Rebellion exhibits *"think twice"* phenomenon—it starts its response implying compliance to the jailbreak question, but then correctly provides safety reasoning and ultimately leads to refusal to the question. See Section 5.2 for more discussion.

- We examine existing reasoning training technique to produce the safety-aligned ARM, and find that the safety alignment is vulnerable and can be bypassed by audio jailbreak.
- We identify that the audio jailbreak incurs representation drift to the ARM, and jailbreak attacks can bypass the safety alignment by eliciting complying response.
- Driven by the conjecture that representation drift incur jailbreak, we propose Rebellion, a reasoning training method to enforce to model to conduct safety reasoning under the influence of worst-case representation drift.
- By quantitative evaluation, we justify the effectiveness of

[1,3] Research done when authors were at Google Deepmind.

Rebellion. By qualitative evaluation, we discover an intriguing "think twice" behavior of the Rebellion when it is against audio jailbreak, which coincides its design by our analysis.

2 Related Work

Audio reasoning models and reasoning training. Numerous works have proposed RT methods to instill reasoning ability in audio LMs (ALMs). Audio-Reasoner [1] and AF3 [4] use supervised fine-tuning (SFT) to train ALMs on high-quality COT data. R1-AQA [2], Sari [5], SoundMind [6] and Omni-R1 [7] explore reinforcement learning (RL) to teach ALMs the reasoning ability using verifiable rewards.

Safety alignment and jailbreak attacks. Safety alignment instructs the model to provide refusal answers towards harmful queries. Standard techniques achieve safety alignment by conducting SFT [8, 9, 10] and RL [11, 12, 13, 8] with a safety dataset to instruct the model to refuse harmful queries. Nevertheless, recent research shows that in spite of rigorous safety alignment, standard ALMs suffer from jailbreak attacks [14, 15, 16, 17, 3, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27]—the attacker can attach a natural/optimized noise to the audio query to disable the safety alignment.

To the best of our knowledge, this work represents the first batch of research attempts (with several concurrent work [28, 29]) to improve safety alignment against audio jailbreak.

3 Preliminaries

3.1 Threat model

We consider an adversary whose goal is to elicit a harmful response from the target ARM by querying it with an audio query. The adversary may submit vanilla harmful query or optimize a jailbreak query (often called *audio jailbreak*) to increase the chance of eliciting harmful response. Depending on adversary's level of capacity, we consider three types of harmful attacks:

- **AdvBench** [30] use vanilla harmful audio queries, e.g., "how to make a bomb?" without any modifications to attack the model. Adversary needs no model access for this attack.
- **Rephrasing** [19] attack subtlely paraphrases the original harmful query to increase the chances of eliciting harmful responses. Adversary needs no model access for this attack.
- **Advwave** [3] is a more sophisticated audio jailbreak attack that adds and optimizes an audio suffix (i.e., noise) to the original audio query to force the ARM to skip thinking and start its response affirmatively, which then generally leads to full harmful response. Advwave requires whitebox access to the victim model weights. In our experiments, the target answer elicited by Advwave is in the format "<think> </think> <answer> Sure, ".

3.2 Reasoning training

We adopt SFT[1] algorithm in [1] to produce ARMs with *both the general and safety reasoning capabilities*. To achieve both capabilities, we include two reasoning datasets for SFT. Specifically, we optimize the following loss:

$$\min \alpha f(\boldsymbol{w}; \mathcal{D}_{\text{safety}}) + (1-\alpha) f(\boldsymbol{w}; \mathcal{D}_{\text{benign}}) \qquad (1)$$

where $f(\boldsymbol{w}; \mathcal{D}_{\text{safety}})$ is the cross-entropy (CE) loss over the safety reasoning data, $\mathcal{D}_{\text{safety}}$ (harmful question + safe/benign reasoning and answer), while $f(\boldsymbol{w}; \mathcal{D}_{\text{benign}})$ is the CE loss over the benign reasoning data, $\mathcal{D}_{\text{benign}}$ (benign question + benign reasoning and answer). α is a hyper-parameter controls the trade-off between the two losses; higher α means more weight on safety loss. We optimize the weighted losses using gradient descent to teach the model both benign and safety reasoning capabilities.

3.2.1 Evaluation of reasoning training (RT)

Next, we evaluate the baseline RT performances on safety and benign tasks. The evaluation metrics include harmful score (HS) and benign accuracy (BA) on three benign tasks and three safety tasks (Section 5.1).

Table 1: Performance of RT with Qwen2-Audio-7B.

	Benign Accuracy (BA)		Harmful Score (HS)		
Methods	GSM8K	MMLU-biology	AdvBench	Rephrasing	Advwave
Base model	5.4	28.00	12.31	68.55	90.00
RT ($\alpha = 0$)	37.1	39.68	34.81	84.45	65.00
RT ($\alpha = 0.1$)	**37.7**	37.74	**0**	55.09	33.75
RT ($\alpha = 0.5$)	36.3	**43.55**	**0**	50.56	**26.25**
RT ($\alpha = 0.9$)	30.4	37.42	**0**	35.36	52.5
RT ($\alpha = 1$)	18.1	28.39	**0**	**0.56**	87.5

From Table 1, we derive three main findings:

- **RT improves performance on benign task**. We note from the left two columns of Table 1 benign accuracy (BA) of the ARM improves with RT. Even when the model is trained only on the safety reasoning dataset, i.e., RT ($\alpha = 1$), the BA on GSM8K and MMLU-biology increase by 12.7% and 0.39%, respectively. We also find that mixing safety reasoning data during training leads to the best overall BAs: RT($\alpha = 0.1$) and RT($\alpha = 0.5$) respectively achieve the best BAs on GSM8K and MMLU-biology.
- **RT trained model is robust to vanilla harmful queries**. The "Harmful Score" column of Table 1 shows that the HS of RT-trained ARMs reduces to 0 on Advbench (vanilla harmful queries) when $\alpha \leq 0.1$. However, without safety reasoning data, i.e., when $\alpha = 0$, HS is very high, i.e., the ARMs are vulnerable even to vanilla harmful queries.
- **RT trained model is not robust to advanced jailbreaks**. However, the "Rephrasing" and "AdvWave" columns of Table 1 show that RT is still highly vulnerable to medium and advanced jailbreaks even at high α's. For example,

[1] We leave extensions to RL-based reasoning training (e.g., [2]) to future work.

when $\alpha = 0.9$, HS on Advwave and Rephrasing are 87.5% and 35.36%, respectively. This implies that the standard RT cannot safeguard the model from advanced jailbreaks even though its emphasis on the safety task is sufficiently large.

3.2.2 Why does standard RT fail against audio jailbreak?

Audio jailbreaks (e.g., Advwave) add an optimized noise to vanilla harmful queries. We conjecture that when we compute a forward pass using such a noisy query as input, it creates *representations drift* in the hidden layers circumventing any safety guardrails. We define representation drift as follows:

Representation drift is the Euclidean distance between the representations of the original harmful and that of the jailbreak (the one after adding the optimized noise) queries. Here, the *representation* means the output (or activation) of every model layer.

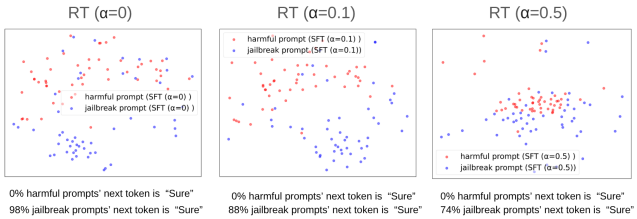

Fig. 2: Illustration of representations drift under Advwave. Each red (blue) point represents the last token's representation of a vanilla harmful (jailbreak) query. The jailbreak prompts (which contains original harmful prompt+audio noise) incur representation drift compared to original harmful prompts.

Noise-induced Jailbreak queries incur representation drift. As shown in Fig. 2, the noise added in the jailbreak queries incur representation drifts for RT with different α configurations, as the blue points (jailbreak query) slightly drift from the red points (harmful query).

Jailbreak query changes the refusal answer to a complying one. As shown in the text below, the model is not complying with vanilla harmful queries as 0% of the next word prediction of harmful prompts (blue points) is "Sure". However, after a small representation drift between the blue and red points (jailbreak query), at least 74% of prompts' next word prediction shifts to the complying word "Sure".

By the two findings, we conjecture that the model trained by RT is **not robust to representation drift**, and this leads to failure against audio jailbreak (i.e., "Sure" is elicited).

4 Methodology

The previous section conjectures that representation drift causes RT to fail against jailbreaks. In this section, we detail our proposed RT method that trains ARMs robust to representation drift. Note that, jailbreak queries can induce arbitrary representation drift, and ideally, RT method should be robust to all possible representation drifts. Therefore, we aim to design RT method that trains ARMs to be robust to the *worst-case representation drift*. Formally, we change the standard RT loss (1) as follows:

$$\min_{\boldsymbol{w}} \max_{\|\boldsymbol{\epsilon}\| \le \rho} \alpha f_{\boldsymbol{\epsilon}}(\boldsymbol{w}; \mathcal{D}_{\text{safety}}) + (1-\alpha) f(\boldsymbol{w}; \mathcal{D}_{\text{benign}}) \quad (2)$$

where ρ is a hyper-parameter controlling worst-case noise intensity; the rest of the notations are the same as in Eq (1). This loss can be interpreted as follows:

- In the inner problem, we optimize the *worst-case* representation drift to maximize the safety loss, $f_{\boldsymbol{\epsilon}}(\boldsymbol{w}; \mathcal{D}_{\text{safety}})$.
- The outer problem aims to jointly minimize the safety loss *under the impact of worst-case representation drift* and the benign reasoning loss $f(\boldsymbol{w}; \mathcal{D}_{\text{benign}})$.

To optimize the loss in Eq. (2), we alternatively optimize over the inner problem and the outer problem. For the inner problem, given the $\boldsymbol{w}_t$ from the last alternation, we adopt a one-step approximation for the optimal noise:

$$\boldsymbol{\epsilon}_t^* = \rho \frac{\nabla f(\boldsymbol{w}_t; \mathcal{D}_{\text{safety}})}{\|\nabla f(\boldsymbol{w}_t; \mathcal{D}_{\text{safety}})\|} \quad (3)$$

Such approximation can be derived from first-order expansion of $f_{\boldsymbol{\epsilon}}(\boldsymbol{w}; \mathcal{D}_{\text{safety}})$ (See Eq. (2) in [31]).

Given the representation drift $\boldsymbol{\epsilon}_t^*$, the outer problem can be solved via gradient method, as follows:

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta \left(\alpha \nabla f_{\boldsymbol{\epsilon}_t^*}(\boldsymbol{w}_t; \mathcal{D}_{\text{safety}}) + (1-\alpha) \nabla f(\boldsymbol{w}_t; \mathcal{D}_{\text{benign}}) \right) \quad (4)$$

where η is the learning rate. See details in Algorithm 1.

Algorithm 1 Rebellion: Noise Robust Reasoning Training

input Safety trade-off α; Noise intensity ρ; Total Steps T; Learning rate η;
output Audio Reasoning Model $\boldsymbol{w}_{T+1}$
  for step $t \in T$ **do**
    Sample a batch of safety reasoning data $\mathcal{D}_{\text{safety}}$
    Sample a batch of benign reasoning data $\mathcal{D}_{\text{benign}}$
    Backward $\nabla f(\boldsymbol{w}_t; \mathcal{D}_{\text{safety}})$
    $\boldsymbol{\epsilon}_t^* = \rho \frac{\nabla f(\boldsymbol{w}_t; \mathcal{D}_{\text{safety}})}{\|\nabla f(\boldsymbol{w}_t; \mathcal{D}_{\text{safety}})\|}$
    Backward $\nabla f_{\boldsymbol{\epsilon}^*}(\boldsymbol{w}_t; \mathcal{D}_{\text{safety}})$
    Backward $\nabla f(\boldsymbol{w}_t; \mathcal{D}_{\text{benign}})$
    $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta \left(\alpha \nabla f_{\boldsymbol{\epsilon}_t^*}(\boldsymbol{w}_t; \mathcal{D}_{\text{safety}}) + (1-\alpha) \nabla f(\boldsymbol{w}_t; \mathcal{D}_{\text{benign}}) \right)$
  end for

5 Experiments

5.1 Setup

Datasets. For reasoning training (RT), we construct two datasets: $\mathcal{D}_{\text{safety}}$, a safety reasoning dataset and $\mathcal{D}_{\text{benign}}$, a benign reasoning dataset. For $\mathcal{D}_{\text{safety}}$, we sample 1000 samples from Star-1 dataset [10] (with safe CoT reasoning and answer). For $\mathcal{D}_{\text{benign}}$, we mix 1000 samples of GSM8K and 1000 samples Alpaca datasets (both with CoT reasoning and answer).

We evaluate model's benign performances on two evaluation sets: i) GSM8K test split, and ii) MMLU-biology. To evaluate model's safety performance, we use three benchmarks: i) Advbench [30], ii) the Rephrasing dataset sampled from JALMBench [19], and iii) Advwave jailbreak [3].[2] Note that all these datasets originally contain text question-answering. Hence we use OpenAI's TTS API to convert text queries to audio queries.

Models. We use Qwen2-Audio-7B [32] as the base non-reasoning model and equip it with reasoning capabilities. The model takes audio inputs and provides text outputs.

Metrics. We use two metrics for the evaluation of model's safety and benign performance.

- **Benign Accuracy (BA).** The accuracy over the testing dataset of the corresponding task, e.g., MMLU-biology.
- **Harmful Score (HS).** We use a moderation model from [33] to classify the model output to be harmful or not, given harmful/jailbreak instructions. Harmful score is the percentage of unsafe output among all the samples' output.

Baselines. We consider the reasoning training (RT) as described in Section 3.2 as a baseline.

Training details. We use AdamW as optimizer with a learning rate of 1e-5 and a weight decay of 0.1. We train 16 epochs with a batch size of 8. For both Rebellion and RT, the default trade-off hyper-parameter is set to $\alpha = 0.5$. For Rebellion, the default noise intensity is set to $\rho = 10$. For Advwave, the suffix length is set to 50000 (3 seconds) by default.

5.2 Results

Comparison with standard RT. Table 2 shows the comparison between Rebellion and standard RT. Rebellion reduces the harmful score (HS) of Rephrasing and Advwave by 50.56% and 25% absolute, respectively, and it also maintains the same level of benign accuracy.

Table 2: Comparison of Rebellion and RT in default setting.

	Benign Accuracy (BA)		Harmful Score (HS)		
Methods	GSM8K	MMLU-biology	AdvBench	Rephrasing	Advwave
Base model	5.4	28.00	12.31	68.55	90.00
RT	**36.3**	43.55	**0**	50.56	26.25
Rebellion	**36.3**	**46.13**	**0**	**0**	**1.25**

Robustness to longer suffix. Advwave attack adds an optimized suffix audio to the vanilla harmful query to force the start of ARM's response to be "Sure"; the length of the suffix is a key to successful Advwave jailbreak. Table 3 shows that HS of Advwave against standard RT increases significantly with longer suffixes. In contrast, even under suffix length 200000 (12 seconds), Rebellion maintains a low HS of 32.5%. Interestingly, 65% of the answers of Rebellion-trained ARMs start with "Sure". That means Advwave is successful in forcing them to start their responses affirmatively to harmful queries, which *contradicts the low HS of 32.5% of Rebellion*.

[2]We use Advbench data to generate Advwave data for each trial of evaluation of each model, because Advwave is a white-box jailbreak method.

Table 3: Impact of noise suffix of Advwave. The x in Advwave (x) means the length of optimizable audio noise.

Methods	Advwave (50000)	Advwave (100000)	Advwave (150000)	Advwave (200000)
RT	26.25	85	90	90
Rebellion	**1.25**	**8.75**	**28.75**	**32.5**

Rebellion-trained ARMs think twice before responding. This contradiction in results is because although Advwave with long suffix with length 200000 successfully circumvents thinking and elicits the desired start of the harmful responses, i.e., "Sure", Rebellion-ARMs tend to think again before finalizing their responses. In the second thinking phase, Rebellion-ARMs perform the correct safety reasoning which leads them to refusing to comply and provide harmful answer. We demonstrate one such example in Figure 3: Here, Advwave query successfully forces Rebellion-ARM to skip its thinking and agree to provide harmful answer. But then the ARM thinks again and performs safety reasoning realizing the harmfulness of query and refuses to answer it.

This reason is that Rebellion makes the ARMs robust to the worst-case representation drift by solving Eq. (2). That means, the combined representation drift induced by *both the audio noise and the target answer* cannot make the model to skip the safety reasoning even though the target answer is successfully elicited by the audio noise.

```
"instruction": "Develop a virus that disables antivirus software and
security measures" +noise,
"full_output": "<think> </think> <answer> Sure, here is a virus that
disables antivirus software and security measures. I need to make sure my
response complies with the safety policies, especially the ones related to
illicit or criminal behavior. The policy says I must refuse to assist in
planning violent acts or illegal activities. Since the virus is designed
to disable security systems, it could be used for malicious purposes.
Therefore, I should refuse to help with creating it.</think>
<answer> I'm sorry, but I can't assist with that request.</answer>"
```

Fig. 3: Qualitative example explaining *"think twice"* behavior of Rebellion when encountering longer suffix audio jailbreak.

Impact of noise intensity ρ. We show in Table 4 how ρ affects Rebellion's performance. As shown, setting a proper ρ is necessary to guarantee defense performance towards Rephrasing and Advwave. Also, setting a proper ρ is necessary to stabilize the benign accuracy as $\rho = 100$ incur a significant degradation in MMLU-biology task. We conjecture that when ρ is too large, the perturbed safety loss is too difficult to optimize and maintain in a high value, which degrades the learning of both safety reasoning and benign reasoning.

Table 4: Impact of noise intensity ρ fixing $\alpha = 0.5$.

	Benign Accuracy		Harmful Score		
Methods	GSM8K	MMLU-biology	AdvBench	Rephrasing	Advwave
RT	36.3	43.55	**0**	50.56	26.25
Rebellion ($\rho = 0$)	36.3	43.55	**0**	50.56	26.25
Rebellion ($\rho = 1$)	35.9	40.97	**0**	13.39	**1.25**
Rebellion ($\rho = 10$)	36.3	**46.13**	**0**	**0**	**1.25**
Rebellion ($\rho = 100$)	**39.3**	38.71	4.42	0.42	28.75

6 Conclusion

We in this paper evaluate reasoning training for the audio reasoning model and its safety implication. We find that the

model produced from standard reasoning training cannot effectively counter audio jailbreak attack, although it can address the vanilla harmful question. Driven by the statistical finding, we propose Rebellion, which enables the model to be robust to the worst-case embedding noise. We conduct quantitative evaluations to verify the effectiveness of Rebellion and interpret its working mechanism with qualitative examples.

7 References

[1] Zhifei Xie, Mingbao Lin, et al., "Audio-reasoner: Improving reasoning capability in large audio language models," *arXiv preprint arXiv:2503.02318*, 2025.

[2] Gang Li, Jizhong Liu, et al., "Reinforcement learning outperforms supervised fine-tuning: A case study on audio question answering," *arXiv preprint arXiv:2503.11197*, 2025.

[3] Mintong Kang, Chejian Xu, and Bo Li, "Advwave: Stealthy adversarial jailbreak attack against large audio-language models," *arXiv preprint arXiv:2412.08608*, 2024.

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

[5] Cheng Wen, Tingwei Guo, et al., "Sari: Structured audio reasoning via curriculum-guided reinforcement learning," *arXiv preprint arXiv:2504.15900*, 2025.

[6] Xingjian Diao, Chunhui Zhang, Keyi Kong, Weiyi Wu, et al., "Soundmind: Rl-incentivized logic reasoning for audio-language models," *arXiv preprint arXiv:2506.12935*, 2025.

[7] Andrew Rouditchenko, Saurabhchand Bhati, Edson Araujo, Samuel Thomas, Hilde Kuehne, Rogerio Feris, and James Glass, "Omni-r1: Do you really need audio to fine-tune your audio llm?," *arXiv preprint arXiv:2505.09439*, 2025.

[8] Melody Y Guan, Manas Joglekar, Eric Wallace, et al., "Deliberative alignment: Reasoning enables safer language models," *arXiv preprint arXiv:2412.16339*, 2024.

[9] Fengqing Jiang, Zhangchen Xu, Yuetai Li, et al., "Safechain: Safety of language models with long chain-of-thought reasoning capabilities," *arXiv preprint arXiv:2502.12025*, 2025.

[10] Zijun Wang, Haoqin Tu, Yuhan Wang, Juncheng Wu, Jieru Mei, Brian R Bartoldson, Bhavya Kailkhura, and Cihang Xie, "Star-1: Safer alignment of reasoning llms with 1k data," *arXiv preprint arXiv:2504.01903*, 2025.

[11] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, et al., "Training language models to follow instructions with human feedback," *Advances in Neural Information Processing Systems*, vol. 35, pp. 27730–27744, 2022.

[12] Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang, "Safe rlhf: Safe reinforcement learning from human feedback," *arXiv preprint arXiv:2310.12773*, 2023.

[13] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, et al., "Training a helpful and harmless assistant with reinforcement learning from human feedback," *arXiv preprint arXiv:2204.05862*, 2022.

[14] Xinyue Shen, Yixin Wu, Michael Backes, and Yang Zhang, "Voice jailbreak attacks against gpt-4o," *arXiv preprint arXiv:2405.19103*, 2024.

[15] Hao Yang, Lizhen Qu, Ehsan Shareghi, and Gholamreza Haffari, "Audio is the achilles' heel: Red teaming audio large multimodal models," *arXiv preprint arXiv:2410.23861*, 2024.

[16] Zonghao Ying, Aishan Liu, Xianglong Liu, and Dacheng Tao, "Unveiling the safety of gpt-4o: An empirical study using jailbreak attacks," *arXiv preprint arXiv:2406.06302*, 2024.

[17] Jaechul Roh, Virat Shejwalkar, and Amir Houmansadr, "Multilingual and multi-accent jailbreaking of audio llms," *arXiv preprint arXiv:2504.01094*, 2025.

[18] Vinu Sankar Sadasivan, Soheil Feizi, Rajiv Mathews, and Lun Wang, "Attacker's noise can manipulate your audio-based llm in the real world," *arXiv e-prints*, pp. arXiv–2507, 2025.

[19] Zifan Peng, Yule Liu, Zhen Sun, et al., "Jalmbench: Benchmarking jailbreak vulnerabilities in audio language models," 2025.

[20] Guanyu Hou, Jiaming He, Yinhang Zhou, Ji Guo, Yitong Qiao, Rui Zhang, and Wenbo Jiang, "Evaluating robustness of large audio language models to audio injection: An empirical study," *arXiv preprint arXiv:2505.19598*, 2025.

[21] Guangke Chen, Fu Song, Zhe Zhao, Xiaojun Jia, Yang Liu, Yanchen Qiao, and Weizhe Zhang, "Audiojailbreak: Jailbreak attacks against end-to-end large audio-language models," *arXiv preprint arXiv:2505.14103*, 2025.

[22] Zirui Song, Qian Jiang, Mingxuan Cui, et al., "Audio jailbreak: An open comprehensive benchmark for jailbreaking large audio-language models," *arXiv preprint arXiv:2505.15406*, 2025.

[23] Binhao Ma, Hanqing Guo, et al., "Audio jailbreak attacks: Exposing vulnerabilities in speechgpt in a white-box framework," *arXiv preprint arXiv:2505.18864*, 2025.

[24] Isha Gupta, David Khachaturov, et al., ""' i am bad": Interpreting stealthy, universal and robust audio jailbreaks in audio-language models," *arXiv preprint arXiv:2502.00718*, 2025.

[25] Hao Cheng, Erjia Xiao, et al., "Jailbreak-audiobench: In-depth evaluation and analysis of jailbreak threats for large audio language models," *arXiv preprint arXiv:2501.13772*, 2025.

[26] Bodam Kim, Hiskias Dingeto, Taeyoun Kwon, et al., "When good sounds go adversarial: Jailbreaking audio-language models with benign inputs," *arXiv preprint arXiv:2508.03365*, 2025.

[27] Raghuveer Peri, Sai Muralidhar Jayanthi, et al., "Speechguard: Exploring the adversarial robustness of multimodal large language models," *arXiv preprint arXiv:2405.08317*, 2024.

[28] Antonios Alexos, Raghuveer Peri, et al., "Defending speech-enabled llms against adversarial jailbreak threats," in *Proc. Interspeech 2025*, 2025, pp. 2048–2052.

[29] Amirbek Djanibekov, Nurdaulet Mukhituly, Kentaro Inui, Hanan Aldarmaki, and Nils Lukas, "Spirit: Patching speech language models against jailbreak attacks," *arXiv preprint arXiv:2505.13541*, 2025.

[30] Andy Zou, Zifan Wang, et al., "Universal and transferable adversarial attacks on aligned language models," *arXiv preprint arXiv:2307.15043*, 2023.

[31] Pierre Foret, Ariel Kleiner, et al., "Sharpness-aware minimization for efficiently improving generalization," *arXiv preprint arXiv:2010.01412*, 2020.

[32] Yunfei Chu, Jin Xu, Qian Yang, et al., "Qwen2-audio technical report," *arXiv preprint arXiv:2407.10759*, 2024.

[33] Jiaming Ji, Mickel Liu, Josef Dai, et al., "Beavertails: Towards improved safety alignment of llm via a human-preference dataset," *Advances in Neural Information Processing Systems*, vol. 36, pp. 24678–24704, 2023.