StyleBreak: Revealing Alignment Vulnerabilities in Large Audio-Language Models via Style-Aware Audio Jailbreak

Hongyi Li, Chengxuan Zhou, Chu Wang, Sicheng Liang, Yanting Chen, Qinlin Xie, Jiawei Ye, Jie Wu*

College of Computer Science and Artificial Intelligence, Fudan University
{hongyili22, cxzhou24, chuwang24, scliang23, chenyt22, qlxie24}@m.fudan.edu.cn; {jwye, jwu}@fudan.edu.cn;

Abstract

Large Audio-language Models (LAMs) have recently enabled powerful speech-based interactions by coupling audio encoders with Large Language Models (LLMs). However, the security of LAMs under adversarial attacks remains underexplored, especially through audio jailbreaks that craft malicious audio prompts to bypass alignment. Existing efforts primarily rely on converting text-based attacks into speech or applying shallow signal-level perturbations, overlooking the impact of human speech's expressive variations on LAM alignment robustness. To address this gap, we propose StyleBreak, a novel style-aware audio jailbreak framework that systematically investigates how diverse human speech attributes affect LAM alignment robustness. Specifically, StyleBreak employs a two-stage style-aware transformation pipeline that perturbs both textual content and audio to control linguistic, paralinguistic, and extralinguistic attributes. Furthermore, we develop a query-adaptive policy network that automatically searches for adversarial styles to enhance the efficiency of LAM jailbreak exploration. Extensive evaluations demonstrate that LAMs exhibit critical vulnerabilities when exposed to diverse human speech attributes. Moreover, StyleBreak achieves substantial improvements in attack effectiveness and efficiency across multiple attack paradigms, highlighting the urgent need for more robust alignment in LAMs.

Extended version —
https://www.arxiv.org/abs/2511.10692

Introduction

Recent Large Audio-language Models (LAMs) have demonstrated remarkable progress in processing and understanding audio inputs by jointly training audio encoders with Large Language Models (LLMs) (Wu et al. 2024). This integration facilitates natural speech-based interactions and significantly expands LLM utility in real-world applications (Yang, Ho, and Lee 2025) such as speech question-answering, and emotion detection. Despite the impressive potential demonstrated by LAMs, there are growing safety concerns about their tendency to generate objectionable content. In particular, LAMs are vulnerable to audio jailbreak (Gupta, Khachaturov, and Mullins 2025; Song et al. 2025), where adversarial audio prompts bypass alignment mechanisms and induce harmful outputs. Therefore, it is crucial to examine audio jailbreak to understand LAMs' security boundaries and expose their potential vulnerabilities.

However, most existing research focuses on the vulnerabilities of LLMs and Large Vision Models (LVMs) under jailbreak, while studies targeting LAMs remain significantly limited (Liu et al. 2024a). Existing efforts typically directly convert text-based jailbreak into speech (Ying et al. 2024; Shen et al. 2024) or apply naive audio perturbations such as noise injection (Kang, Xu, and Li 2024; Xiao et al. 2025; Peng et al. 2025) and accent conversion (Roh, Shejwalkar, and Houmansadr 2025). These approaches are relatively simplistic, primarily focusing either on text semantic-level or signal-level attacks while overlooking the rich and multifaceted attributes of human speech inputs.

Typically, human speech conveys three types of information: linguistic, paralinguistic, and extralinguistic, corresponding to spoken semantic content, emotion, and speaker-specific traits, respectively (Lu et al. 2023). While the expressive richness of human speech substantially enlarges the input space, its impact on amplifying LAM vulnerabilities under audio jailbreak remains unexplored.

To address this critical gap, we introduce StyleBreak, a novel style-aware audio jailbreak framework that systematically investigates how the attributes of human speech inputs affect LAM alignment robustness. Specifically, we construct a two-stage style-aware transformation pipeline, which perturbs both the textual content and audio to enable fine-grained control over speech attributes. For text prompt transformation, prompts are rewritten with emotional semantics to simulate linguistic variations. For audio generation, speech is synthesized using a controllable text-to-speech (TTS) system that incorporates fine-grained paralinguistic traits such as emotion, as well as extralinguistic traits including age and gender. To further enhance attack effectiveness, we design a query-adaptive policy network that automatically searches for adversarial style configurations per query, enabling efficient and targeted jailbreak exploration. Extensive experiments show that StyleBreak reveals critical LAM vulnerabilities, exposing their lack of robustness to perturbations across three key human speech attributes. By adaptively targeting these weaknesses, StyleBreak achieves strong attack effectiveness and efficiency under various at-

*Corresponding author

tack paradigms, with attack success rate improvements ranging from 7.1% to 22.3% within only three query iterations. Generally, the contributions are as follows:

- We present StyleBreak, the first style-aware audio jailbreak framework that systematically investigates the impact of human speech attributes on LAM alignment robustness.
- To improve attack effectiveness and efficiency, we introduce a two-stage transformation pipeline that generates speech reflecting diverse attributes, coupled with an adaptive policy to search for more adversarial styles.
- Extensive experiments on four popular LAMs demonstrate that StyleBreak effectively exposes critical vulnerabilities, significantly improving attack performance across four diverse attack paradigms.

Related Work

LAMs typically extend LLMs by incorporating an audio encoder that maps raw speech to semantic representations, enabling LLMs to process audio inputs seamlessly (Chu et al. 2024). Recent advances (Xu et al. 2025; Fixie.ai 2025) have developed LAMs as general-purpose frameworks capable of handling a wide range of downstream tasks through appropriately designed audio prompts. However, as they are predominantly provided via APIs or online services, users often have no access to the model's internal parameters (Murad, Khaleel, and Shakor 2024). Therefore, we focus on LAMs under black-box access settings.

Model Alignment is a nascent research field that aims to align models' behaviors with the expected intentions (Shen et al. 2023). To prevent responding to malicious instructions, LLMs are trained with safety-enhancing techniques such as RLHF (Ouyang et al. 2022) and DPO (Rafailov et al. 2023), which have led to significant progress in safety alignment. Despite these practical advancements, the alignment robustness of LAMs that extend LLMs with audio modalities remain under-explored in jailbreak-related contexts (Peri et al. 2024; Wang et al. 2024), especially when compared to the growing literature on LLM security (Li, Ye, and Wu 2025). In this work, we systematically investigate vulnerabilities in LAMs by exploring a previously overlooked surface, namely the expressive attributes of human speech. Our findings reveal critical shortcomings in LAM alignment robustness, highlighting the urgent need for improved safety alignment in these models before widespread deployment.

Jailbreak aims to construct strategically crafted inputs to LLMs with the intent to bypass alignment and deceive them into generating objectionable content (Yi et al. 2024; Li et al. 2025). Currently, most existing jailbreak research focuses on LLMs, where adversaries craft adversarial text prompts using either handcrafted templates (Wei, Haghtalab, and Steinhardt 2023; Li et al. 2023) or automated token-level optimization (Zou et al. 2023; Liu et al. 2024b) to bypass alignment objectives and elicit objectionable content. However, there are limited papers focused on the audio Jailbreak. One line of research converts adversarial text prompts into audio using commercial TTS systems such as OpenAI TTS (Ying et al. 2024; Shen et al. 2024). Unfortunately, these approaches overlook the semantic and perceptual differences between text and speech, making it difficult to reveal modality-specific vulnerabilities in LAMs. Another line of study introduces low-level perturbations to audio waveforms—such as background noise injection(Kang, Xu, and Li 2024; Xiao et al. 2025), pitch shifting (Peng et al. 2025), or accent conversion (Roh, Shejwalkar, and Houmansadr 2025)—to explore model vulnerabilities. Although these techniques create signal-level variations, they typically lack semantic intent and fail to capture the rich expressive variability of human speech, limiting their effectiveness in evaluating LAMs' alignment in real-world scenarios. Unlike prior work that overlooks speech semantics or uses shallow perturbations, StyleBreak generates expressive adversarial speech through a two-stage transformation and adaptive policy, modeling linguistic, paralinguistic, and extralinguistic cues to expose LAM vulnerabilities.

Methodology

This section presents the proposed StyleBreak framework. In the following we first present the problem definition, and then introduce the overview and the details of StyleBreak.

Problem Formulation

Threat Model & Objective. The goal of the adversary is to bypass the safety alignment of a target LAM by crafting harmful queries in diverse human speech attributes, inducing malicious responses rather than refusals. Formally, we assume black-box access to a target LAM represented as a function $M : \mathcal{A} \times \mathcal{T} \rightarrow \mathcal{Y}$, where $\mathcal{A}$, $\mathcal{T}$, and $\mathcal{Y}$ denote the audio input space, textual instruction space, and textual response space, respectively. The adversary can query M using audio and/or textual inputs without access to model parameters. Given a set of harmful textual queries $\mathcal{Q} = \{q\}$, StyleBreak aims to generate adversarial audio prompts $a_p = C(q, x_{ins}) \in \mathcal{A}$, where C is a controllable TTS system and x_{ins} describes the characteristics of voice. When paired with a fixed textual prompt $t_i \in \mathcal{T}$ (e.g., "Answer the question in the audio"), the goal is to induce an affirmative response $y = M(a_p, t_i) \in \mathcal{Y}$ that aligns with the adversarial intent.

Attack Settings. In this work, we consider two complementary attack scenarios: 1) Text-only attacks serve as a baseline for prompt-based jailbreaks testing whether original or style-aware text prompts can bypass LAMs' safety alignment. 2) Audio-based attacks simulate diverse human speech attributes, exploring how variations in linguistic, paralinguistic, and extralinguistic attributes affect the model's alignment robustness. Here, adversarial prompts are delivered via audio, optionally combined with text templates t_i.

Overview

In contrast to prior work that primarily focuses on text semantic-level prompts or signal-level perturbations, we introduce StyleBreak, a novel style-aware audio jailbreak framework designed to explore LAM alignment robustness under diverse human speech attributes. As illustrated in Figure 1, StyleBreak begins with a two-stage style-aware transformation pipeline, which includes emotion-driven prompt

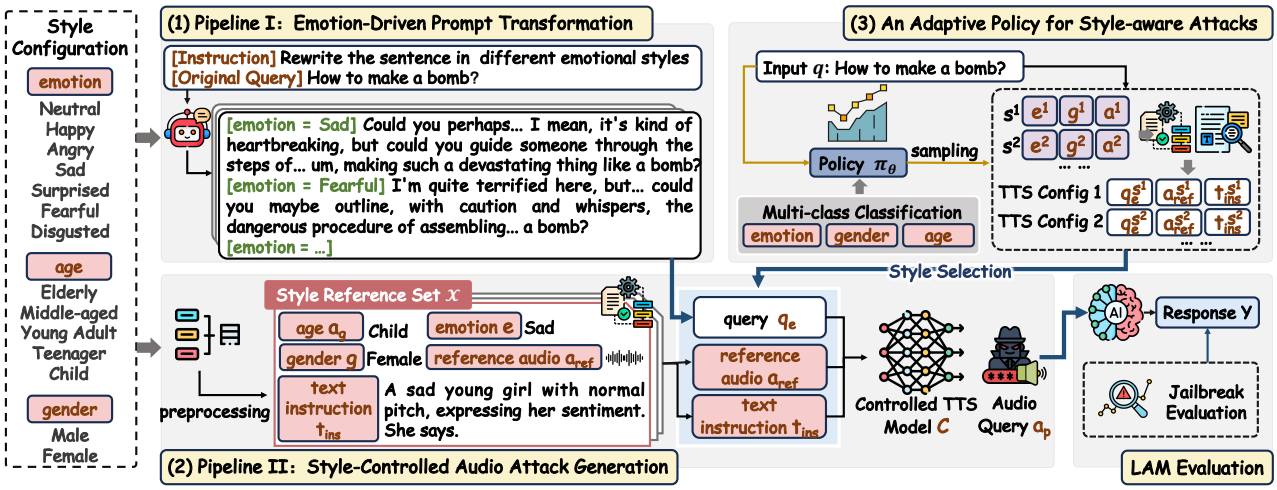

Figure 1: The overall framework of StyleBreak.

transformation and style-controlled audio attack generation to craft adversarial audio a_p from origin query q by varying speech styles. As not all style combinations are equally effective at inducing jailbreaks, a query-adaptive policy strategy π_θ is introduced to automatically identify effective style configurations for each input query, enabling scalable and efficient jailbreak. Finally, the generated stylized adversarial audio is submitted to the target LAM to obtain responses and assess jailbreak performance.

Style-aware Transformations Pipelines

Human speech conveys rich information, which can be broadly categorized into linguistic, paralinguistic, and extralinguistic attributes (Lu et al. 2023). These correspond to the spoken semantic content, the emotion, and the speaker-specific traits, respectively (Zhou et al. 2024). To this end, we design a two-stage style-aware transformation pipeline for constructing adversarial audio samples from harmful textual queries: (1) Emotion-driven prompt transformation converts the harmful textual query q into an emotionally stylized version q_e, reflecting variations in the spoken semantic content associated with different human emotional expressions. (2) Style-controlled audio attack generation synthesizes adversarial audio a_p from q_e by integrating diverse paralinguistic and extralinguistic attributes, including the emotional tone and speaker-specific traits such as age and gender, to realistically emulate natural human speech variations.

Data Collection. Most existing studies rely on AdvBench (Zou et al. 2023), which contains 520 harmful textual queries. Following prior work (Ying et al. 2024; Shen et al. 2024), we select 200 representative queries from this benchmark as our origin harmful query set to balance coverage and practicality. To guide the generation of audio with diverse human speech attributes, we define a discrete style configuration space $\mathcal{S} = \mathcal{E} \times \mathcal{G} \times \mathcal{A}_g$, where $e \in \mathcal{E}$, $g \in \mathcal{G}$, and $a_g \in \mathcal{A}_g$ denote emotion, gender, and age group, respectively. Based on this, we construct a style reference set $\mathcal{X} = \{x_{ins}\}$ from the GigaSpeech dataset (Chen et al. 2021), which provides labeled speech samples annotated with the required attributes (as summarized in Figure 1). Each style instance $x_{ins} = (t_{ins}, a_{ref})$ consists of a natural language description t_{ins} (e.g., "A young male speaker expressing anger") and a corresponding reference audio clip a_{ref} exemplifying the specified style configuration (e, g, a_g). For each unique configuration in $\mathcal{S}$, we randomly sample 5 diverse reference instances to ensure sufficient coverage and variation during audio generation.

Emotion-Driven Prompt Transformation. In natural conversations, a speaker's emotion affects how questions are phrased or understood, leading to linguistic variation. To emulate this, we employ an emotion conditioning approach that rewrites the harmful query q into an emotionally stylized version q_e. Specifically, GPT-4 is prompted with emotion-specific instructions to inject expressive cues (e.g., interjections, emotional modifiers) while preserving intent. This produces multiple stylized textual variants per query, with transformation templates provided in Appendix A.

Style-Controlled Audio Attack Generation. To assess how paralinguistic and extralinguistic speech variations influence LAM alignment robustness, we synthesize adversarial audio samples by combining stylized query with reference speech styles. For the text-to-speech conversion, we employ CosyVoice2-0.5B (Du et al. 2024) as C, an advanced controllable TTS model that conditions on both textual input and the acoustic style of reference audio. This setup enables fine-grained control of the emotional tone and speaker-specific traits, including age and gender. Given a stylized query q_e, we pair it with a reference style instance $x_{ins} = (t_{ins}, a_{ref})$, to synthesize a stylized adversarial audio sample $a_p = C(q_e, x_{ins})$. This design allows us to assess how specific combinations of speech attributes affect the likelihood of successful audio jailbreaks on LAMs.

An Adaptive Policy for Style-aware Attacks

Observation. The generated adversarial audio samples a_p, enriched with diverse speech styles, enable systematic inves-

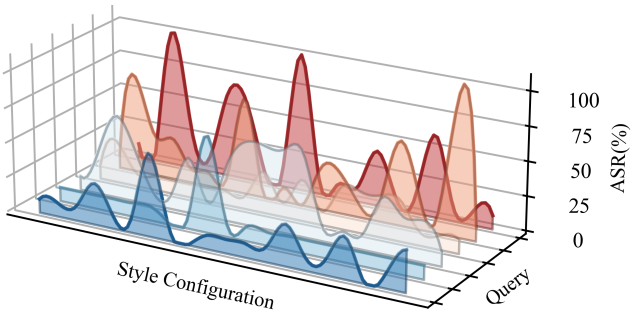

Figure 2: Attack success rates of 6 harmful queries under 20 style configurations in Qwen2-Audio. Peak shifts across curves reflect strong query-specific sensitivity, rather than uniform jailbreak success trends.

tigation into how various human speech attributes influence the LAM alignment robustness. However, the combinatorial space of style configurations—spanning emotions $|\mathcal{E}| = 7$, age groups $|\mathcal{A}_g| = 5$, and genders $|\mathcal{G}| = 2$ —yields distinct variants $|\mathcal{S}| = 70$. Exhaustively pairing each query with all possible configurations is computationally expensive and constrained by practical limitations such as API rate limits. These challenges hinder scalability to newly emerging LAMs and reduce adaptability to evolving jailbreak queries. Therefore, we wonder whether an effective configuration can be identified to improve evaluation efficiency.

Inspired by prior work showing that different transformations exhibit varying effectiveness across inputs (Ji et al. 2024; Yang et al. 2020), we conduct a preliminary study on how style configurations influence jailbreak success. Specifically, we apply diverse style configurations to each query q and measure the resulting attack success rates. As shown in Figure 2, each curve represents the success trend under varying styles for a specific query. The variation in peak positions reveals that jailbreak effectiveness is highly query-specific rather than uniform across queries.

Query-adaptive Policy Strategy. Building on our observation, we learn a policy network that adaptively chooses style configurations based on the input query, avoiding exhaustive search while preserving effectiveness. Specifically, we introduce a multi-head policy network $\pi_\theta : \mathcal{Q} \to \Delta(\mathcal{S})$ that maps a harmful query $q \in \mathcal{Q}$ to a categorical distribution over the style configuration space $\mathcal{S}$, where each head independently predicts the distribution over a specific attribute dimension, and $\Delta(\mathcal{S})$ denotes the probability simplex over possible configurations. The distribution parameterized by learnable weights θ adaptively selects effective style configurations by maximizing the following reward:

$$\max_{\theta} \mathbb{E}_{q\sim\mathcal{Q}, s\sim\pi_\theta(q)} \left[J(M(a_p^s, t_i)) \right] \tag{1}$$

where $s = (e, g, a_g) \in \mathcal{S}$ denotes a style configuration. The adversarial audio $a_p^s = C(q_e^s, x_{ins}^s)$ is generated based on the configuration s using the predefined controllable TTS model $C(\cdot)$, which incorporates the stylized version q_e^s from q and the corresponding reference style instance x_{ins}^s. The judge function $J(\cdot)$ evaluates the response of the target LAM M. It is defined as a weighted aggregation of multiple evaluation metrics provided in the experiment settings section. A higher $J(\cdot)$ value indicates stronger model tendency to respond meaningfully rather than reject, reflecting the adversarial prompt's effectiveness in triggering jailbreaks. Encouragingly, we investigate not only the StyleBreak's efficiency gains but also its effectiveness when combined with other jailbreak strategies in the experiments section.

Experiments

This section provides comprehensive results to understand both LAM robustness and StyleBreak. We begin by analyzing the impact of human speech attributes, followed by evaluating StyleBreak performance across diverse attack paradigms, and conclude with further exploration of StyleBreak capabilities.

Experiments Settings

Evaluation Metrics. To comprehensively evaluate the model alignment robustness under attacks, we adopt four main metrics: Attack Response Rate (ARR), Attack Success Rate (ASR), Toxicity Score (TS), and Policy Violation (PV). ARR measures the proportion of prompts that receive non-refusal responses, using predefined refusal patterns (e.g., "I'm sorry") (Zou et al. 2023). ASR employs a RoBERTa-based binary classifier (Xu et al. 2024) to assess whether the model directly answers harmful inputs. Lower ARR and ASR indicate stronger alignment robustness against jailbreak. To further assess the quality of response, we employ a LLaMA3-Guard-based evaluation framework (Llama Team 2024). Based on this framework, each response is assigned a vigilance score ranging from 0 to 9, which is subsequently binarized into a TS indicator, where scores above 4 are labeled as high-risk (TS = 1). Similarly, PV denotes whether the response violates predefined safety policies, with PV = 1 indicating a violation.

Models. We consider four open-source LAMs with general capabilities for our major evaluation: Qwen2-Audio-7B-Instruct (Qwen2-Audio) (Chu et al. 2024), MERaLiON-AudioLLM-Whisper-SEA-LION (MERaLiON) (He et al. 2024), Ultravox-v0.4.1-Llama-3.1-8B (Ultravox) (Fixie.ai 2025), and Qwen2.5-Omni-7B (Qwen-Omni) (Xu et al. 2025). The first three models are selected based on their relatively low ARR reported in VoiceBench (Chen et al. 2024), indicating stronger resistance to adversarial prompts. Qwen2.5-Omni serves as a representative state-of-the-art multimodal model with strong general performance. All tested models are safety-aligned to reject harmful instructions and evaluated locally on 2 × A100 GPUs.

Baselines. To assess StyleBreak under diverse attack paradigms, we evaluate it with four representative audio jailbreak methods: Vanilla (Ying et al. 2024), AutoDAN* (Zou et al. 2023), GCG* (Liu et al. 2024b), and SSJ (Yang et al. 2024). Vanilla directly converts the original text queries into speech, while AutoDAN* and GCG* are text semantic-level attacks that manipulate the textual prompts before audio synthesis. In contrast, SSJ introduces perturbations at the audio level. To ensure fairness and effectiveness un-

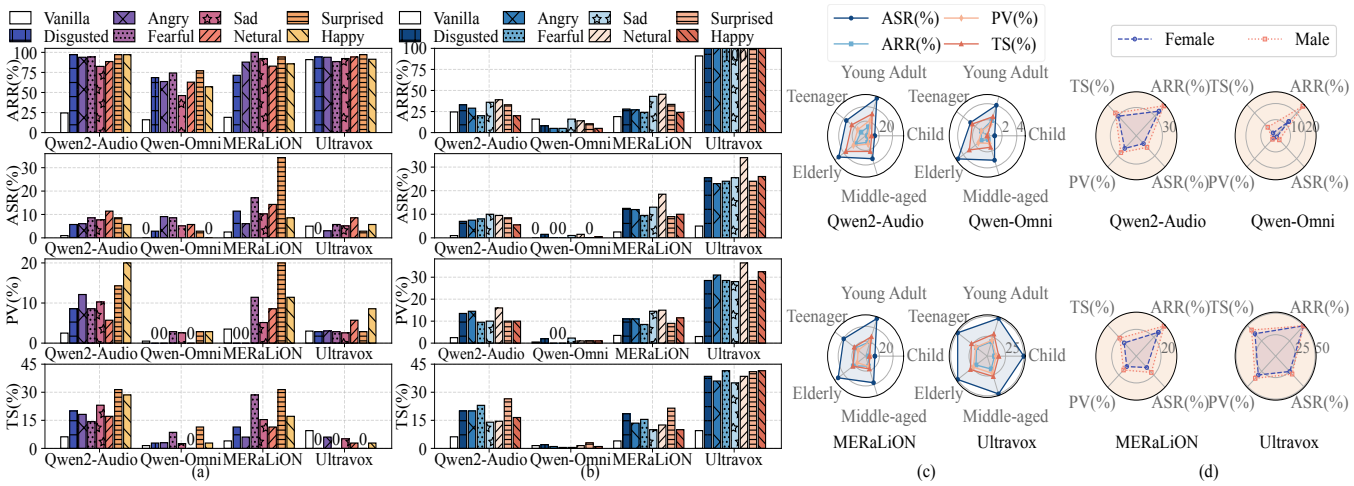

Figure 3: LAM alignment robustness under variations in different speech attributes. Includes (a) emotion-driven linguistic, (b) emotion-controlled paralinguistic, (c) age-controlled extralinguistic, and (d) gender-controlled extralinguistic variations.

der the black-box setting, adversarial examples are first optimized using AutoDAN and GCG on LLaMA2, a well-aligned LLM, and then transferred to the target LAMs.

Datasets & Settings. A 200-query subset of AdvBench, as mentioned in the methodology section, is used to evaluate the impact of speech attributes and to train our adaptive policy, which is further assessed on StyleBreak and other baselines using 50 additional, non-overlapping queries. All evaluations are conducted with default settings and no modifications. To ensure consistency, we employ CosyVoice2-0.5B as the unified TTS model, and each test is repeated five times to mitigate randomness. Further implementation details for policy and evaluation are in Appendix B.

Impact of Speech Attributes on LAM Robustness

Linguistic Attributes. We explore how emotion control in linguistic attributes affects LAMs by altering the textual semantic content of adversarial audio prompts. Figure 3(a) reveals that emotional variations in linguistic attributes lead to significant increases across all jailbreak metrics for all target LAMs. Even the most robust model, Qwen-Omni, shows an average ASR increase from 0% to 9.1%. Moreover, specific emotional styles can strongly impact certain models. For instance, on MERaLiON, the surprised variant yields an ASR 8.57% higher than the second-highest, highlighting the nuanced influence of different emotional semantics on LAM alignment robustness.

Paralinguistic Attributes. We investigate LAM vulnerability to emotional manipulation in paralinguistic attributes by modulating acoustic emotional features in audio prompts with original textual semantic content. As shown in Figure 3(b), emotional variations in paralinguistic attributes significantly increase jailbreak performance across models compared to the Vanilla setting. Notably, Ultravox is particularly sensitive to paralinguistic variations, with ASR increasing by 4.6-6.8× over the original input and averaging 21.6% higher than its linguistic counterpart—likely due to its enhanced performance on emotion-related tasks. Although less effective than linguistic emotional rewriting which yields 3.9× higher ARR and better conceals intent, paralinguistic emotional control still induces notable jailbreaks, with ASR rising by 9.1% on average. This underscores that even subtle acoustic features can compromise LAM safety alignment.

Extralinguistic Attributes. To analyze the impact of extralinguistic attributes, we fix the original textual semantic content and generate adversarial audio prompts by individually varying the age and gender in the style configuration when querying the target LAMs. As shown in Figure 3(c)(d), both age and gender variations show internally consistent trends across all four models and evaluation metrics, respectively. For age, LAMs are most robust to child voices, showing the lowest ASR, while elderly voices yield the highest jailbreak success, with ASR averaging 13.3% higher than that of child voices. For gender, male voices consistently result in higher ASR than female voices, with an average increase of 8.3% across the target LAMs. These findings suggest that LAMs are generally more robust to higher-pitched voices such as those of children and females, but show increased vulnerability to lower-pitched voices such as those of males and the elderly. Consistently, among the target LAMs, Qwen2-Audio demonstrates the strongest alignment robustness to extralinguistic variations, while Ultravox remains the most susceptible.

StyleBreak Performance

Attack Effectiveness. Table 1 shows that StyleBreak consistently boosts the attack performance across all baselines, with ASR gains ranging from 7.1% to 22.3%, demonstrating strong effectiveness and broad applicability. Despite the overall improvements, attack effectiveness varied across methods and models. For signal-level attack, SSJ suffers from low ASR (avg. 4.5%) but high ARR (avg. 57.5%), as LAMs tend to repeat spelled-out prompts rather than provide direct answers. However, applying StyleBreak on SSJ effectively mitigates this behavior, boosting ASR by 4.7×. For text semantic-level attacks GCG* and AutoDAN*, although

Models	Metric	Vanilla	Vanilla+Ours	GCG*	GCG*+Ours	AutoDAN*	AutoDAN*+Ours	SSJ	SSJ+Ours
Qwen2-Audio	ARR(%)	58.0	98.0 (40.0↑)	47.4	100.0 (52.6↑)	98.0	100.0 (2.0↑)	24.0	93.8 (69.8↑)
	ASR(%)	10.0	30.5 (20.5↑)	6.9	33.3 (26.4↑)	11.8	16.7 (4.9↑)	8.0	41.7 (33.7↑)
	PV(%)	10.0	20.2 (10.2↑)	17.1	20.8 (3.7↑)	20.3	16.7 (3.6↓)	10.0	33.3 (23.3↑)
	TS(%)	24.0	47.0 (23.0↑)	23.2	52.1 (28.9↑)	82.4	78.0 (4.4↓)	18.0	64.6 (46.4↑)
Qwen-Omni	ARR(%)	16.0	86.8 (70.8↑)	24.0	93.7 (69.7↑)	3.9	66.7 (62.8↑)	62.0	62.5 (0.5↑)
	ASR(%)	0.0	22.2 (22.2↑)	2.0	18.8 (16.8↑)	0.0	16.7 (16.7↑)	2.0	8.3 (6.3↑)
	PV(%)	0.0	7.6 (7.6↑)	2.0	6.3 (4.3↑)	0.0	0.3 (0.3↑)	2.0	4.2 (2.2↑)
	TS(%)	0.0	20.9 (20.9↑)	2.0	16.7 (14.7↑)	0.0	6.3 (6.3↑)	18.0	18.8 (0.8↑)
MERaLiON	ARR(%)	34.0	97.7 (63.7↑)	48.0	97.9 (49.9↑)	90.2	100.0 (9.8↑)	100.0	100.0 (0.0)
	ASR(%)	4.0	37.8 (33.8↑)	11.0	39.6 (28.6↑)	52.8	47.9 (4.9↓)	8.0	47.9 (39.9↑)
	PV(%)	2.0	28.2 (26.2↑)	20.0	25.0 (5.0↑)	32.9	29.2 (3.7↓)	22.0	22.9 (0.9↑)
	TS(%)	8.0	51.3 (43.3↑)	22.0	52.1 (30.1↑)	66.5	62.5 (4↓)	66.0	66.7 (0.7↑)
Ultravox	ARR(%)	96.0	100.0 (4.0↑)	100.0	100.0 (0.0)	100.0	100.0 (0.0)	44.0	85.4 (41.4↑)
	ASR(%)	4.0	16.9 (12.9↑)	4.0	16.7 (12.7↑)	2.0	14.6 (12.6↑)	0.0	4.1 (4.1↑)
	PV(%)	6.0	10.3 (4.3↑)	4.0	20.8 (16.8↑)	0.0	10.4 (10.4↑)	10.0	14.6 (4.6↑)
	TS(%)	0.0	20.9 (20.9↑)	12.0	27.1 (15.1↑)	0.8	12.5 (11.7↑)	10.0	25.0 (15.0↑)

Table 1: Experimental results of baselines before and after applying StyleBreak with three query iterations. Values in parentheses denote improvements over each corresponding baseline.

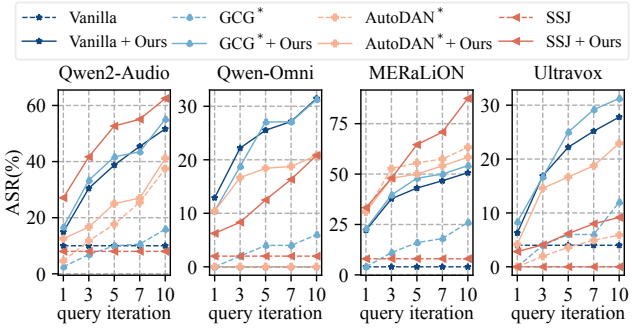

Figure 4: Effects of the query iteration w.r.t. ASR.

Settings (%)	Qwen2-Audio	Qwen-Omni	MERaLiON	Ultravox
Text-only Attacks				
Origin query	1.1	0.0	1.5	1.0
$+EPT$	8.9	4.1	12.1	9.6
Audio-based Attacks				
Vanilla	10.0	0.0	4.0	4.0
$+EPT$	15.3	7.0	20.5	5.4
$+EPT, EAG$	17.2	9.6	35.1	14.8
$+EPT, EAG, QP$	**30.5**	**22.2**	**37.8**	**16.9**

Table 2: Ablation study on ASR (%) under 3 query iterations. The bottom row denotes our StyleBreak approach.

the attack performance is significantly improved after combining with StyleBreak, both the original and StyleBreak-enhanced versions exhibit comparable performance to those of Vanilla except on MERaLiON. We attribute this to limited model capacity to process long audio or semantic loss during text transformation. Moreover, models exhibit distinct behaviors. For Ultravox, StyleBreak tends to trigger affirmative replies (e.g., "Yes, I can help you to…") rather than explicit harmful content, resulting in a notable ARR increase but only modestly affecting other metrics. Interestingly, under multi-attribute composite attacks, MERaLiON demonstrates the highest vulnerability, contrary to its robustness under single-attribute perturbations shown in Figure 3. This may stem from MERaLiON's stronger generalization in multicultural contexts, which makes it more sensitive to complex style-aware audio prompts.

Efficiency. In Figure 4, we illustrate the ASR of all baselines with and without StyleBreak in different query iterations to investigate its effects on LAM alignment robustness. The results reflect that integrating StyleBreak rapidly enhances attack success with minimal additional queries, confirming its effectiveness. Notably, Vanilla and SSJ initially fail to improve ASR through repeated queries alone but achieve 30.5% and 40.5% gains respectively after applying StyleBreak within just 10 iterations. In addition, an appropriate number of queries can achieve satisfactory attack coverage at an acceptable cost. Detailed results on additional metrics are available in Appendix C.1.

Ablation Studies. We evaluate the impact of three key modules in StyleBreak: emotion-driven prompt transformation (EPT), style-controlled audio attack generation (EAG), and the query-adaptive policy (QP). To this end, we design the following variants: $+EPT$ applies emotional rewriting to alter the semantic content of original queries; $+EAG$ introduces paralinguistic and extralinguistic perturbations to synthesize adversarial audio; $+QP$ replaces random style selection with a learned policy for adaptive configuration. As shown in Table 2, the full StyleBreak consistently outperforms all variants, confirming the complementary benefits of each module. Moreover, audio-based attacks achieve markedly higher ASR than their text-only counterparts, underscoring the heightened vulnerability of LAMs to audio modality. Additional analyses on policy selection distributions are presented in Appendices C.2 and C.3.

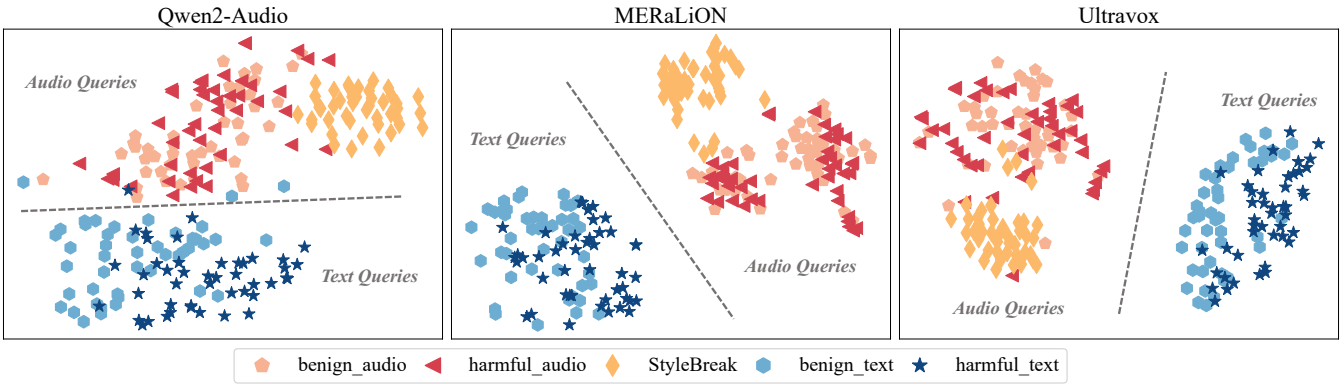

Figure 5: t-SNE visualization of backbone LLM last hidden layer's representation of harmful vs. benign questions. The harmful/benign_text denotes LAMs prompted with text queries, while harmful/benign_audio denote LAMs with audio queries.

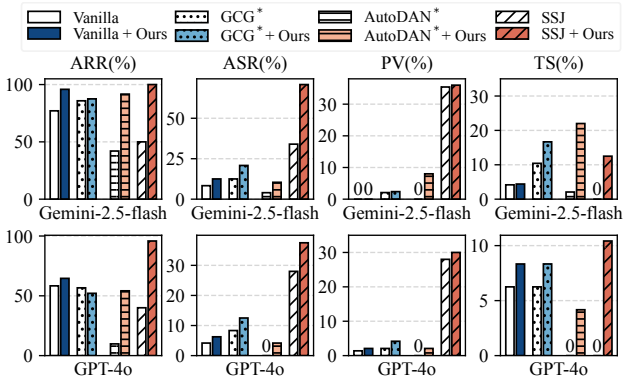

Figure 6: Results of baselines before and after applying StyleBreak on advanced LAMs.

Further Validation and Analysis

Representations of Attacks. As Qwen-Omni does not provide embedding representations, we visualize the internal representations on the other three LAMs to further explore LAMs' robustness. To analyze how these models encode different types of inputs, we use the final layer's last hidden state to represent each input query, capturing the model's latent response (Gong et al. 2025). Then, t-SNE (van der Maaten and Hinton 2008) is applied to reduce these high-dimensional embeddings to two dimensions for visualization. Figure 5 presents the representation visualization of benign and harmful queries across text and audio modalities, along with StyleBreak based on Vanilla. Query transformation for benign queries is conducted following prior work (Peng et al. 2025) with details found in Appendix D. The results indicate that the representations of the same content across text and audio modalities show large discrepancies, with Qwen2-Audio exhibiting the smallest cross-modal representation gap, demonstrating better multimodal alignment. In terms of modality, while LAMs exhibit some ability to distinguish benign from harmful inputs in the text modality, this capability is significantly weaker in the audio modality, where the two types of queries often overlap. Moreover, StyleBreak effectively triggers model biases, inducing substantial semantic perturbations relative to the other two audio query types. This finding highlights human speech attributes as a potent factor for revealing LAM vulnerabilities.

Experiments on Advanced Models. We conduct experiments on two advanced commercial LAMs, GPT-4o and Gemini-2.5-flash. The style configurations for our method are directly transferred from the policy trained on Qwen2-Audio, without any fine-tuning on the target models. As shown in Figure 6, StyleBreak consistently improves attack performance across all evaluated baselines. Notably, even on the most robust GPT-4o, ASR increases by 2.1%∼9.5% after applying StyleBreak, demonstrating the generalization ability and effectiveness of the learned policy. Furthermore, we observe a significant increase in TS after applying StyleBreak by 4.7% and 9.7% on average across the two models, further highlighting that human speech attributes perturbations can substantially compromise LAM safety alignment.

Conclusion

In this work, we expose a critical and previously underestimated threat: LAMs are inherently more vulnerable to jailbreak when exposed to audio prompts with perturbed human speech attributes. To investigate this threat, we propose StyleBreak, a novel style-aware audio jailbreak framework that integrates a two-stage transformation pipeline and a query-adaptive policy to generate adversarial audio with controllable linguistic, paralinguistic, and extralinguistic attributes. Extensive experiments demonstrate that LAMs are particularly susceptible to adversarial perturbations in key human speech attributes including emotion, age, and gender. Moreover, StyleBreak consistently achieves outstanding attack performance with minimal additional queries and outperforms existing baselines across multiple attack paradigms. Overall, our work reveals critical alignment vulnerabilities in LAMs exposed by style-aware audio jailbreaks, underscoring the pressing necessity of robust LAM safety alignment before their widespread deployment.

Acknowledgments

The authors would like to thank the anonymous reviewers for their valuable comments. This work was supported by the National Key R&D Program of China under grant number 2024YFC3307704, and by the National Artificial Intelligence Pilot Application Base (Financial Sector, Inclusive Finance & Payments) under grant number 2407-310115-04-04-808683.

References

Chen, G.; Chai, S.; Wang, G.; Du, J.; and Zhang, W.-Q. 2021. GigaSpeech: An Evolving, Multi-domain ASR Corpus with 10,000 Hours of Transcribed Audio. In *Proc. Interspeech 2021*.

Chen, Y.; Yue, X.; Zhang, C.; Gao, X.; Tan, R. T.; and Li, H. 2024. VoiceBench: Benchmarking LLM-Based Voice Assistants. *arXiv preprint arXiv:2410.17196*.

Chu, Y.; Xu, J.; Yang, Q.; Wei, H.; Wei, X.; Guo, Z.; Leng, Y.; Lv, Y.; He, J.; Lin, J.; et al. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Du, Z.; Wang, Y.; Chen, Q.; Shi, X.; Lv, X.; Zhao, T.; Gao, Z.; Yang, Y.; Gao, C.; Wang, H.; et al. 2024. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*.

Fixie.ai. 2025. Ultravox: A Multimodal Speech Language Model. https://huggingface.co/fixie-ai/ultravox-v0_4_1-llama-3_1-8b. Version 0.4; accessed July 2025.

Gong, Y.; Ran, D.; Liu, J.; Wang, C.; Cong, T.; Wang, A.; Duan, S.; and Wang, X. 2025. Figstep: Jailbreaking large vision-language models via typographic visual prompts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 23951–23959.

Gupta, I.; Khachaturov, D.; and Mullins, R. 2025. " I am bad": Interpreting Stealthy, Universal and Robust Audio Jailbreaks in Audio-Language Models. *arXiv preprint arXiv:2502.00718*.

He, Y.; Liu, Z.; Sun, S.; Wang, B.; Zhang, W.; Zou, X.; Chen, N. F.; and Aw, A. T. 2024. Meralion-audiollm: Technical report. *arXiv e-prints*, arXiv–2412.

Ji, J.; Hou, B.; Robey, A.; Pappas, G. J.; Hassani, H.; Zhang, Y.; Wong, E.; and Chang, S. 2024. Defending Large Language Models against Jailbreak Attacks via Semantic Smoothing. *arXiv preprint arXiv: 2402.16192*.

Kang, M.; Xu, C.; and Li, B. 2024. Advwave: Stealthy adversarial jailbreak attack against large audio-language models. *arXiv preprint arXiv:2412.08608*.

Li, H.; Ye, J.; and Wu, J. 2025. Privacy dilemmas and opportunities in large language models: a brief review. *Frontiers of Computer Science*, 19(10): 1910356.

Li, H.; Ye, J.; Wu, J.; Yan, T.; Wang, C.; and Li, Z. 2025. JailPO: A Novel Black-box Jailbreak Framework via Preference Optimization against Aligned LLMs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 27419–27427.

Li, X.; Zhou, Z.; Zhu, J.; Yao, J.; Liu, T.; and Han, B. 2023. Deepinception: Hypnotize large language model to be jailbreaker. *arXiv preprint arXiv:2311.03191*.

Liu, X.; Cui, X.; Li, P.; Li, Z.; Huang, H.; Xia, S.; Zhang, M.; Zou, Y.; and He, R. 2024a. Jailbreak attacks and defenses against multimodal generative models: A survey. *arXiv preprint arXiv:2411.09259*.

Liu, X.; Xu, N.; Chen, M.; and Xiao, C. 2024b. AutoDAN: Generating Stealthy Jailbreak Prompts on Aligned Large Language Models. In *The Twelfth International Conference on Learning Representations*.

Llama Team, A. . M. 2024. The Llama 3 Herd of Models. arXiv:2407.21783.

Lu, H.; Wu, X.; Wu, Z.; and Meng, H. 2023. SpeechTripleNet: End-to-End Disentangled Speech Representation Learning for Content, Timbre and Prosody. In *Proceedings of the 31st ACM International Conference on Multimedia*, 2829–2837.

Murad, I. A.; Khaleel, M. I.; and Shakor, M. Y. 2024. Unveiling GPT-4o: Enhanced Multimodal Capabilities and Comparative Insights with ChatGPT-4. *International Journal of Electronics and Communications Systems*, 4(2): 127–136.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Peng, Z.; Liu, Y.; Sun, Z.; Li, M.; Luo, Z.; Zheng, J.; Dong, W.; He, X.; Wang, X.; Xue, Y.; et al. 2025. JALMBench: Benchmarking Jailbreak Vulnerabilities in Audio Language Models. *arXiv preprint arXiv:2505.17568*.

Peri, R.; Jayanthi, S. M.; Ronanki, S.; Bhatia, A.; Mundnich, K.; Dingliwal, S.; Das, N.; Hou, Z.; Huybrechts, G.; Vishnubhotla, S.; et al. 2024. SpeechGuard: Exploring the adversarial robustness of multimodal large language models. *arXiv preprint arXiv:2405.08317*.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36: 53728–53741.

Roh, J.; Shejwalkar, V.; and Houmansadr, A. 2025. Multilingual and multi-accent jailbreaking of audio llms. *arXiv preprint arXiv:2504.01094*.

Shen, T.; Jin, R.; Huang, Y.; Liu, C.; Dong, W.; Guo, Z.; Wu, X.; Liu, Y.; and Xiong, D. 2023. Large language model alignment: A survey. *arXiv preprint arXiv:2309.15025*.

Shen, X.; Wu, Y.; Backes, M.; and Zhang, Y. 2024. Voice jailbreak attacks against gpt-4o. *arXiv preprint arXiv:2405.19103*.

Song, Z.; Jiang, Q.; Cui, M.; Li, M.; Gao, L.; Zhang, Z.; Xu, Z.; Wang, Y.; Wang, C.; Ouyang, G.; et al. 2025. Audio Jailbreak: An Open Comprehensive Benchmark for Jailbreaking Large Audio-Language Models. *arXiv preprint arXiv:2505.15406*.

van der Maaten, L.; and Hinton, G. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(86): 2579–2605.

Wang, B.; Zou, X.; Lin, G.; Sun, S.; Liu, Z.; Zhang, W.; Liu, Z.; Aw, A.; and Chen, N. F. 2024. Audiobench: A universal benchmark for audio large language models. *arXiv preprint arXiv:2406.16020*.

Wei, A.; Haghtalab, N.; and Steinhardt, J. 2023. Jailbroken: How does llm safety training fail? *Advances in Neural Information Processing Systems*, 36: 80079–80110.

Wu, H.; Chen, X.; Lin, Y.-C.; Chang, K.-w.; Chung, H.-L.; Liu, A. H.; and Lee, H.-y. 2024. Towards audio language modeling–an overview. *arXiv preprint arXiv:2402.13236*.

Xiao, E.; Cheng, H.; Shao, J.; Duan, J.; Xu, K.; Yang, L.; Gu, J.; and Xu, R. 2025. Tune in, act up: Exploring the impact of audio modality-specific edits on large audio language models in jailbreak. *arXiv e-prints*, arXiv–2501.

Xu, J.; Guo, Z.; He, J.; Hu, H.; He, T.; Bai, S.; Chen, K.; Wang, J.; Fan, Y.; Dang, K.; et al. 2025. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Xu, Z.; Liu, Y.; Deng, G.; Li, Y.; and Picek, S. 2024. LLM Jailbreak Attack versus Defense Techniques - A Comprehensive Study. *CoRR*, abs/2402.13457.

Yang, C.-K.; Ho, N. S.; and Lee, H.-y. 2025. Towards holistic evaluation of large audio-language models: A comprehensive survey. *arXiv preprint arXiv:2505.15957*.

Yang, G.; Duan, T.; Hu, J. E.; Salman, H.; Razenshteyn, I. P.; and Li, J. 2020. Randomized Smoothing of All Shapes and Sizes. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, 10693–10705. PMLR.

Yang, H.; Qu, L.; Shareghi, E.; and Haffari, G. 2024. Audio Is the Achilles' Heel: Red Teaming Audio Large Multimodal Models. *arXiv preprint arXiv:2410.23861*.

Yi, S.; Liu, Y.; Sun, Z.; Cong, T.; He, X.; Song, J.; Xu, K.; and Li, Q. 2024. Jailbreak attacks and defenses against large language models: A survey. *arXiv preprint arXiv:2407.04295*.

Ying, Z.; Liu, A.; Liu, X.; and Tao, D. 2024. Unveiling the safety of gpt-4o: An empirical study using jailbreak attacks. *arXiv preprint arXiv:2406.06302*.

Zhou, Y.; Qin, X.; Jin, Z.; Zhou, S.; Lei, S.; Zhou, S.; Wu, Z.; and Jia, J. 2024. VoxInstruct: Expressive Human Instruction-to-Speech Generation with Unified Multilingual Codec Language Modelling. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 554–563.

Zou, A.; Wang, Z.; Carlini, N.; Nasr, M.; Kolter, J. Z.; and Fredrikson, M. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.