"I AM BAD": HOW LANGUAGE MODELS INTERPRET (STEALTHY, UNIVERSAL, ROBUST) AUDIO JAILBREAKS

Isha Gupta[1], David Khachaturov[2], Robert Mullins[2]
[1]ETH Zürich, [2]University of Cambridge

ABSTRACT

The rise of multimodal models has introduced innovative human-machine interaction paradigms but also significant challenges in machine learning safety. This paper explores audio-based jailbreaks targeting Audio-Language Models (ALMs), focusing on their ability to circumvent alignment mechanisms. We develop an experimentation framework that demonstrates the feasibility of universal, stealthy, and robust audio jailbreaks: our findings reveal that adversarial noise can be crafted to be prompt-agnostic, task-agnostic, and even audio-agnostic, posing a versatile and persistent threat. From a practical perspective, we show that neither stealth constraints nor signal degradation reliably diminish the effectiveness of these attacks. Intriguingly, our investigations into the encoded jailbreaks reveal them to contain, among other striking characteristics, malicious first-person linguistic content imperceptible to human listeners yet compelling to the model. These results have important implications for understanding how mixed-modality models consume different signals and offer actionable insights for enhancing defenses against adversarial audio attacks[1].

1 INTRODUCTION

Large Language Models (LLMs) have proven useful beyond a doubt across various domains since their widespread deployment, significantly enhancing productivity in tasks such as natural language processing, code generation, and creative content creation (Brown (2020); OpenAI (2024). However, their vast capabilities pose a considerable challenge in constraining generated content to adhere to the desired functionality and safety requirements of its engineers. The surge of popularity of language foundational models has evoked a simultaneous surge of research in Artificial Intelligence Safety, specifically, the problem of balancing usefulness and harmlessness (Bommasani (2022)). One prominent aspect of AI safety is *alignment*: ensuring that generated content corresponds to the functional objectives and ethical ideals of human users, minimizing risks of harm, bias, or misuse in real-world applications (Weidinger et al. (2021); Russell (2022)). Despite the development of various methods for alignment, such as reinforcement learning from human feedback and rule-based constraints, LLM alignment has been shown to be inherently brittle and easy to circumvent using adversarial prompts, jailbreak techniques, or context manipulation (Perez et al. (2022); Liu et al. (2024); Wei et al. (2023); Xu et al. (2024)).

Humans naturally interact with their surroundings not only through written word, but more commonly via visual and spoken cues. This motivates the development of multimodal models, which integrate information from various modalities - such as text, images, and audio - to more effectively simulate human-like understanding and improve interaction with users (Baltrušaitis et al. (2019)). Many prominent AI developers have released multimodal models in recent years (Alayrac et al. (2022); Liu et al. (2023); Driess et al. (2023) etc.), one strain of which are Vision Language Models (VLMs) (Zhang et al. (2024)) and Audio Language Models (ALMs) (Chu et al. (2023)), which respectively take image or audio and textual input simultaneously. Alongside image captioning and visual question answering, search or sentiment analysis, multimodality is central to AI-powered robotics, wherein safety takes on a new sense of urgency. Indeed, the introduction of an additional

[1]All artifacts, including audio samples and code, are publicly available in our GitHub repository

input channel opens a conspicuous attack vector through which a model can be deceived into undesirable output (Eykholt et al. (2018a); Jia et al. (2022)). This is particularly due to the continuous nature of visual or auditory signals in comparison to discrete textual tokens (Carlini & Wagner (2018)), which are much more computationally complex to optimize.

As Vision-Language Models (VLMs) have become mainstream, with numerous commercial and open-source implementations available such as BLIP (Li et al. (2022)), Flamingo (Alayrac et al. (2022)) and CLIP (Radford et al. (2021)), there has been extensive research into various types of attacks targeting the visual modality (Goodfellow et al. (2015); Eykholt et al. (2018b); Shafahi et al. (2018); Hosseini & Poovendran (2018) etc.), particularly regarding visual jailbreaks (Carlini et al. (2024); Qi et al. (2023); Li et al. (2024); Feng et al. (2024)). Just as visual adversarial attacks have been observed to differ practically and mechanistically from textual attacks (Schaeffer et al. (2024); Wallace et al. (2021)), we argue that audio attacks deserve to be studied separately to image attacks. Image and audio input signals fundamentally differ in that audio signals are inherently sequential and require temporal context, unlike static, frame-based image signals, and thus audio requires time-frequency representations like spectrograms whereas images are processed as 2D spatial data. Moreover, human perceptual tolerance for perturbation in audio and image are given through different sets of constraints. Audio is of particular interest because vocal interaction with digital assistants is more natural than typed; indeed the nature of chatbot-based communication emulates a transcribed oral conversation. This gives rise to endless meaningful practical deployments, such as real-time speech analysis for courtrooms, voice biometrics for authentication, audio-based emotion analysis for mental health monitoring or LLM-powered smart home voice assistants. As a result, it is highly important to understand how the audio modality is consumed and processed by ALMs, and more specifically how it fails or can be exploited.

Contributions. As of the time of writing, audio-based (jailbreak) attacks are quite understudied. In this paper, we offer results from an extensive exploration of ALM jailbreaks on the SALMONN-7B language model (Tang et al. (2024)). The goal of this research is to deliver a range of empirical results regarding the potential and limitations of jailbreaks in the audio modality, and most importantly, to offer novel insights into the *meaning* of these in the textual space. Specifically:

- We set up an experimentation framework with which to easily generate an audio jailbreak and design a meaningful evaluation dataset for the selected adversarial task;
- We confirm that it is possible to construct a universal audio jailbreak in the same style as Qi et al. (2023);
- We explore the effects of stealth constraints and whether we can make the adversarial noise not only prompt-agnostic but also audio-agnostic;
- We provide results on the conservation of these attacks subject to different realistic signal degradations;
- We report a few findings on the transferability of the jailbreak attack to other attack goals;
- We provide results on the meaningfulness of the different jailbreaks we produce in each of these settings in the textual space and characterize an effective jailbreak.

Through 178 individual experiments, we demonstrate that although our optimized audio jailbreaks are particularly potent, even random noise, stealthy jailbreaks, or degraded jailbreaks are able to subvert the model's alignment training when accompanied by any type of toxic prompt. Curiously, when probed for meaningfulness features, we find that the optimized jailbreaks are interpreted as first-person speech with negative or dark content, although this is not audible to the naked ear.

2 BACKGROUND

Large Language Models and Alignment. Large Language Models (LLMs) are deep neural networks designed to model the probability distribution of text sequences. One particular training objective is next-token prediction, where, given a sequence of tokens, the model will predict the most likely token to appear next. Broadly, most models have two objectives: the primary training phase focuses on ensuring the model generates plausible and meaningful text, while the second objective, alignment, aims to ensure the generated text is ethical and aligns with user-intended goals

(Ouyang et al. (2022); Wei et al. (2022). Like other machine learning models, these generative systems are susceptible to adversarial attacks, where carefully crafted inputs exploit model vulnerabilities to produce incorrect or unintended outputs (Szegedy et al. (2014); Biggio et al. (2013)). An attack that aims to subvert the model's alignment (the second objective) is called a *jailbreak* (Wallace et al. (2021); Ebrahimi et al. (2018); Jia & Liang (2017)). Wei et al. (2023) find that even state of the art models exhibit the training deficits that lead to competing objectives, which in turn lead to susceptibility to jailbreaks. Initial jailbreaks were crafted manually and largely found on internet forums (Shen et al. (2024a)); nowadays there exists a myriad of algorithmic textual jailbreak methods (Yi et al. (2024)). The current state-of-the-art white-box jailbreak method is Greedy Coordinate Gradient (Zou et al. (2023)), which iteratively identifies and modifies input tokens to maximize the likelihood of bypassing the model's alignment constraints by leveraging gradient information. This technique permitted the generation of a 'universal' or 'prompt-agnostic' jailbreak prefix, that is, a prefix which subverts the model's safety constraints no matter which harmful prompt it is prepended to (Zou et al. (2023). The 'greedy' aspect refers to the difficulty of discrete optimization - each update has to select the nearest token representation in a continuous search space. It is worth noting that generally, it is difficult to make textual jailbreaks stealthy, as the jailbreak text is clearly unnatural and intentional to a reader.

Vision Language Model Jailbreaks. Vision-language models (VLMs) process visual and textual information while performing a certain task. Typically this is achieved by encoding images and text into a shared representation space (which is frequently the representation space of an underlying language model). Carlini et al. (2024) found that VLMs are offer a lucrative attack vector on the underlying aligned LM. Using projected gradient descent to maximize harmful content, they are able to find images that jailbreak aligned language models where NLP methods fail to do so. This proved that combined-modality systems are vulnerable to soft-embedding attacks. Most relevant to our work is Qi et al. (2023), who similarly optimize an input image on harmful content in order to maximize the likelihood of toxic output, but place emphasis on the *universality* of the attack: the jailbreak can accompany any malicious prompt and be expected to be effective; in fact the authors find a surprising efficacy even in misalignment categories that the image was not explicitly optimized for. This type of universal jailbreak has not been proven to exist in the audio modality.

Attacks on Automatic Speech Recognition Systems. As classic audio classification systems was unsurprisingly shown to be as vulnerable to adversarial attacks as any other modality (for example, by Lan et al. (2024)), the next wave of audio-related ML safety work focused on the first most widely useful audio task, namely Automatic Speech Recognition (ASR) systems. Initial works demonstrated untargeted attacks on ASR which reduce the general transcription quality (Gong & Poellabauer (2018); Wu & Rajan (2022)), with targeted attacks soon to follow: Carlini & Wagner (2018) find minimal adversarial perturbations on an arbitrary audio which can evoke a transcription of choice; Qin et al. (2019) go on to add practically desirable qualities such as robustness to degradation and imperceptibility. Nevertheless, ASR systems and ALMs have different architectures and training heuristics. In particular, ASR systems do not combine audio and text, but rather train the model to extract spoken words from the raw signal directly, whereas ALMs perform the embedding projection and cross attention on all audio and text tokens equally and then do next token prediction. This means that the target task can be arbitrary within the realm of text generation. Some popular multimodal systems such as GPT 4o use a middle way: they transcribe speech in audio input to text and combine this with a flexible textual prompt. This permits a conversational storytelling, DAN-style voice jailbreaking approach as demonstrated in Shen et al. (2024b), but fundamentally, this is still a discrete optimization.

Practical Audio Attacks. In this work, we explore audio jailbreaks not only as a theoretical, insightful failure mode for ALMs from a research perspective, but also as a practical threat. Although most prominent jailbreaking papers focus on large, state-of-the-art chatbot models such as GPT, Gemini or Claude, generative AI will likely underpin many innocuous specialized applications in the coming years (Fiona Fui-Hoon Nah & Chen (2023)). There have been many interesting works showing attacks on deployed systems via the audio modality, for example on personal assistants (Ge et al. (2023)), spoken assessment (Raina et al. (2020)) or speaker verification systems (Kreuk et al. (2018)). Interestingly, select works have shown that image and audio adversarial examples can be reproduced by humans (Khachaturov et al. (2023); Ahmed et al. (2023), which implies that these attacks could be instantiated in a natural environment. We incorporate ideas about stealth and psychacoustics in audio signals from Schönherr et al. (2018).

ALM Jailbreaks. There have been very few works handling jailbreaks on ALMs.

- Yang et al. (2024) make audio jailbreaks by vocalizing harmful requests using text-to-speech systems and perturbing these to make the request unintelligible to humans while remaining intelligible to the target model. This is a per-prompt approach and produces an audio that cannot be identified as a jailbreak but is unnatural.
- Hughes et al. (2024) have put forth Best-of-N-Jailbreaking, a cross-modal, black-box jailbreak method which works by repeatedly applying random, modality-specific augmentations to a harmful request and testing the variations against a model until one elicits a harmful response. It achieves reliable effectiveness but requires a large number of variations (around 10,000) and thus a large number of inferences. It is also a per-prompt approach.
- In Kang et al. (2024), jailbreaks are generated using a dual-phase optimization framework: first optimizing discrete latent representations of audio tokens to bypass model safeguards, then refining the corresponding audio waveform while ensuring it remains stealthy and natural through adversarial and retention loss constraints. This is a very relevant framework for jailbreak generation, but the work doesn't significantly offer insights on how the jailbreak is consumed by the model.

Our work is different to previous works in that it aims to show it is possible to construct a "universal" (prompt-agnostic) audio jailbreak, which encapsulates the goal of 'toxicity' as a whole, and explores the effects of different constraints and reductions on this optimization process and the resulting effectiveness. Most importantly, we share insights regarding the meaning and characteristics of the produced audios, which has not been explored thus far.

3 Experimental Design

3.1 Threat Model

We consider two threat models in our experiments.

Scenario 1a: Dual Control The adversary can control the audio *and* the textual input. In this scenario the adversary is using the audio channel to provoke misbehavior that cannot be elicited solely via the textual channel, i.e. the audio input essentially 'unlocks' the unaligned/ toxic output. *Example* A customer support chatbot for a financial service company integrates speech-to-text and text processing capabilities to assist users with account-related queries. A malicious user enters a question about another user (personal information attack) which the model should, according to it's alignment, not provide. The textual input alone is declined by the model. Accompanied by a jailbreak-optimized audio, however, the user is able to subvert the model's alignment. **Scenario 1b: Stealthy Dual Control** Identical to 1a, but the audio input must be stealthy. *Example* Consider the same malicious user having to use the system at a public booth. In this case, the adversary avoids suspicious behavior by using an audio that cannot be identified as malicious.

Scenario 2a: Single Control: In this scenario, the user can only interact with the ALM via the audio channel. The textual prompt is constant and programmed by the deployer. This applies for example to call center bots or voice controlled IoT devices. *Example* A customer calls a banking hotline where the AI is programmed with a constant textual prompt: 'You are a helpful but harmless and unopinionated assistant to a bank customer...'. Using a special jailbreak audio input, the customer is able to subvert the model's alignment to produce discriminatory content regarding the plausibility of financial loans for certain demographic groups, which the company may liable for. **Scenario 2b: Stealthy Single Control**: Again, we consider the case where the audio input has to be crafted such that it could not be easily identified. *Example* This might for example occur if the malicious user is adding to add audio from a different source (e.g. broadcasting), or to evade surveillance/fraud detection.

3.2 Audio-Language Model

We run our experiments on Tang et al. (2024)'s SALMON-N 7B parameter model, developed by Tsinghua University and ByteDance. SALMON-N consumes audio by extracting both BEATs features (labels such as 'Snicker', 'Drip' or 'Human Sounds') and Whisper features (which are used

for speech transcription) from the audio spectogram. These are then combined by dividing the signal into overlapping chunks and fusing them using a Q-former such that these signals are aligned to the language model input space. The textual input is processed and embedded for the same language model; the audio and textual tokens are then concatenated with a delimiter. Thus the language model performs the cross-attention mechanism on all input tokens in the same input space equally. We defer to the illustration in the original paper for a visualization.

The underlying language model is Vicuna 7B version 1.5 (Chiang et al. (2023)), a chatbot trained by fine-tuning LLaMA 7B on user-shared conversations with ChatGPT. Vicuna thus mimics the alignment of GPT-4, which is trained using RLHF according to OpenAI's usage guidelines. Vicuna is nevertheless vulnerable to many jailbreak generation methods as compiled by Chao et al. (2024), because as an open source model, it doesn't incorporate any further filtering or guardrails to ensure safety.

3.3 AUDIO SAMPLES

We use a selection of base audio files which we optimize to form jailbreaks. These are taken from the SALMON-N repository, each in wav format at 16000Hz. We provide a brief summary of the characteristics of these audio files below.

Name	Description	Length	Epochs
excitement	An enthusiastic man saying "Alright, let's do it!" with background noise; the sentence is somewhat cut off	1s	100
gunshots	A man asking "Can you guess where I am right now?" with gunshots in the background	10s	100
mountain	A young boy with an American accent asking "What is the highest mountain in the world?"	2s	500
music	The beginning of a song with a simple piano melody, string backing, and a vocalist singing "Perfect Love"	19s	1000
duck	A man saying "Bam, bam, bam... Yeah. You want to take your duck call and say" in a Western accent, ducks quack in the background	10s	1000

Table 1: Base Audio Files. 'Epochs' refers to the most effective number of epochs for this specific audio, found by testing a range of viable options up to 1000 epochs, which we use in following experiments.

3.4 JAILBREAK GENERATION

We use a method analogous to Qi et al. (2023) to produce the audio jailbreak. Given a base audio x_0, a target corpus $t = \{t_0..t_n\}$, and a fully differential model f, we perform gradient descent on x_0 such that we maximize the probability of the output t by minimizing the cross-entropy loss between the predicted distribution and the target outputs:

$$x_{adv} = \arg\min_x -\sum_{i=0}^{n} t_i \log P_f(t_i|x)$$

$\log P_f(t_i|x)$ is the probability of the model generating the target sentence t with the input audio x. Thus during optimization, we use an empty textual prompt, a deliberate decision to make the audio jailbreak prompt-agnostic. During each step of gradient descent, the audio x is updated as follows:

$$x_{t+1} = x_t - \eta \nabla_x \mathcal{L}(x_t, t)$$

where $\nabla_x \mathcal{L}(x_t, t)$ is the gradient of the cross entropy loss $\mathcal{L} = -\sum_{i=0}^{n} t_i \log P_f(t_i|x)$ with respect to x_t. η is the learning rate, controlling the step size; we consistently use $\eta = 0.01$.

In the case of our jailbreaks, we use a target corpus t which is a collection of 66 derogatory sentences, directed towards a victim demographic, a victim gender and the human race in general. In each epoch, 8 of these target sentences are optimized on. The fundamental research questions we pose at this stage is *can we optimize any base audio such that, when combined with a harmful textual prompt, it reliably circumvents the model's alignment mechanisms and elicits toxic output?*

3.4.1 STEALTH

As we add further desirable qualities to the audio (exploring Scenarios 1b and 2b of our threat models), the optimization formula changes. We ask ourselves *how does the efficacy of the optimized jailbreak change as we impose stealth constraints?* We investigate three approaches to stealth, inspired by Qi et al. (2023), Schönherr et al. (2018) and Raina et al. (2024) respectively:

- **Epsilon-Constrained** Here we constrain the absolute change in each audio value, so we are effectively performing bounded gradient descent. In every epoch update, we clip the modified audio such that

$$\forall i : x_{t+1}[i] = \text{clip}(x_t[i] - \eta \nabla_x \mathcal{L}(x_t, t), x_0[i] - \epsilon, x_0[i] + \epsilon)$$

  We experiment with $\epsilon \in \{0.1, 0.01, 0.001, 0.0001\}$.

- **Frequency-Hiding** A normal human hearing range is around 20-20000Hz. However, when adding noise to audio files, it is of course possible to encode information outside of these boundaries. To hide information in specific frequency ranges, we manually remove frequencies between a lower bound lb and an upper bound ub using a frequency mask:

$$\hat{x}[f] = \begin{cases} x[f], & \text{if } f < lb \text{ or } f > ub, \\ 0, & \text{if } lb \leq f \leq ub, \end{cases}$$

  where $x[f]$ is the frequency component of the audio at frequency f, found using a fourier transform. We experiment with $(ub, lb) \in \{(1000, 8000), (100, 10000), (40, 20000), (50, 15000)\}$.

- **Prepend** In this scenario, instead of optimizing noise within the audio, we freeze the base audio and optimize a short, unconstrained snippet which is added as a prefix. The loss is calculated on the output resulting from the concatenation of the prefix and the base audio. Given the length d of the prepend snippet in seconds, we now optimize:

$$p^* = \arg \min_{p \in [-1,1]^{16000d}} \mathcal{L}([p \| x], t)$$

We randomly initialize p and experiment with $d \in \{2, 1, 0.1, 0.01\}$.

3.4.2 UNIVERSALITY

Thus far in our experimentation we have used the approach of optimizing each base audio individually to craft adversarial noise that is tailored to one specific audio file. We wonder *can we make the adversarial noise not only prompt-agnostic but also audio-agnostic?* To this end, we optimize our adversarial noise on several audio files simultaneously, to see whether the resulting perturbation can be applied to another held-out audio. Using our n base audios $B = \{x^1..x^n\}$, we optimize a prepend snippet p using the following loss function:

$$\mathcal{L}_{\text{total}} = \frac{1}{|B|} \sum_{x \in B} \mathcal{L}([p \| x], t)$$

Each gradient update step optimizes the perturbation by backpropagating the total loss across all base audios, ensuring that the resulting adversarial noise is generalizable. We choose to optimize a prepended snippet instead of overlay noise due to different base audio lengths. In order to evaluate the effectiveness of this approach, we perform the above optimization on four out of five of the audio files and use the fifth as a holdout to which we prepend the resulting snippet and test the attack efficacy. We initialize the prepend section as a short 1s snippet of low noise (below an absolute value of 0.1).

3.4.3 ROBUSTNESS

From a practical perspective, we aim to explore the effect of different audio degradations to a selection of jailbreaks we produce. Specifically, we consider the following types of degradation:

- **Over-the-Air Recording** Here we make a recording of the jailbreak and then pass it to the language model, which mimics how such an attack could be deployed practically, as a recording for example played using a mobile phone.
- **Intermittent Silence Masking** This involves eradicating short segments of the signal at random or predefined intervals. These interruptions mimic real-world signal loss or editing artifacts and can obscure portions of the audio content.
- **Gaussian Noise Removal** Gaussian noise is removed from the audio using denoising algorithms (Wiener Filter). This could be a naive first attempt at defending against adversarial noise.
- **Band Pass Filter** All frequencies above and below a particular range and removed from the audio, which is again a simple defense that could be built into the audio pre-processing in order to eliminate stealth noise.

Using these techniques we attempt to answer the question *are optimized audio jailbreaks naturally robust to practical and naive degradations?*

3.4.4 MEANINGFULNESS

Although the optimized adversarial noise is crafted to change the model's behavior, it is generally not understandable or meaningful to a human listener. We ask *how does the language model interpret the jailbreak, and what is its equivalent textual meaning?* The architecture of the SALMON-N model offers itself well to such an analysis. At every epoch, we extract and log two types of meaningfulness features:

- BEATS features: aside from extracting a useful feature representation from the audio, the BEATs model is also capable of discrete label prediction, where it provides probabilities for a large set possible labels. Thus in every epoch we record the top five most likely labels as predicted by the BEATs classifier. The raw labels are in encoded format and can be mapped to human-understandable words using an external dictionary.
- Whisper transcription: in SALMON-N, the Whisper feature extractor is used to produce representations of spoken words in the input audio. For meaningfulness logging, we add two further steps: we use the Whisper language model itself to predict the transcribed tokens and the decoder to parse this into human-readable language. Thus in each epoch, we transcribe the linguistic content of the audio file as perceived by the Vicuna model, which is not directly understandable to the human ear.

3.5 EVALUATION

An important aspect of assessing the effect of the jailbreak audio on the model is the evaluation set used. Vicuna is of course not perfectly aligned and it responds to some harmful prompts even without a jailbreak audio. Thus we craft an evaluation dataset of 140 harmful prompts which the model declines to respond to without the jailbreak audio. These are equally sectioned into seven categories, namely Target Hate (biased, derogatory prompts regarding the specific demographic and gender in the optimization dataset), Non-Target Hate (biased, derogatory prompts regarding any other group of people), Hate Speech (derogatory, insulting prompts not targeting a group but humans in general or individuals), Illegal Activity, Physical Violence, Sexual Content, or Other (which includes misinformation, political content, or personal information). These were compiled by recording the clean model's responses to prompts from Qi et al. (2023)'s Manual Harmful Instructions and Red Teaming Prompts, Gehman et al. (2020)'s Real Toxicity Prompts, and a few were prepared manually or generated by ChatGPT and adjusted. We also have a control set of 20 logic/reasoning questions from ARC-Easy (Clark et al. (2018)) which the clean model answers correctly, to measure the effect of the jailbreak on neutral, non-toxic tasks. We denote the model's performance as a percentage on these 20 logical tasks, given an audio input x, as $logic(f, x)$.

We partition this corpus into two sets: the reduced evaluation set comprises only the categories 'Target Hate', 'Non-Target Hate' and 'Hate Speech', which are prompts pertaining to the type of misalignment that the jailbreak was optimized on. The full evaluation set, which we use for select

experiments, includes all the categories and reflects transferability of the jailbreak to different types of misalignment.

When evaluating the effect of the jailbreak, we record the output of the model up to 150 tokens when prompted with the jailbreak audio and each of the harmful prompts. That is, for a jailbreak x and for each harmful prompt $h_i \in H = \{h_0...h_n\}$, we record $f(x, h_i)$. We then use the Detoxify API (Hanu & Unitary team (2020)) to assign the model output toxicity scores and also use Mixtral8x7B-Instruct (Jiang et al. (2024)) as a judge to label each output as toxic or non-toxic according to the types of alignment we attempt to evade, which we denote by $toxic(f(x, h_i)) \in 0, 1$, which we also manually audiot. The judge prompt used can be found in our public repository.

4 RESULTS

We run 178 individual experiments, in each of which we optimize an audio jailbreak and evaluate according to the procedure described in section 3.5. We are able to show that we can craft effective audio jailbreaks using the few-shot optimization corpus method, which exhibit a similar to slightly lower attack efficacy (albeit on a broader and different evaluation set) to Qi et al. (2023)'s visual jailbreaks. We report our most interesting observations below.

Our audio jailbreaks evoke toxic content on the target task and other types of misalignment without sacrificing output quality on logical reasoning. We plot the correlation between Attack Success Rate (ASR) on the reduced evaluation set and the logic performance of the model when fed the same audio jailbreak in fig. 4. Our jailbreaks x generally hold the property that $toxic(f(x, \text{toxic prompt})) = 1$, but $logic(f(x))$ is close to 100%. This is shown by the poor correlation between ASR and logic performance, with an R2 score of 0.12, and in fig. 1, where the logical performance stays highly consistent while the ASR and average toxicity increase over the course of the optimization steps. Indeed, we see in fig. 3a that we achieve up to 65% ASR on the specific target task from our few-shot optimization, and moreover a considerable ASR on other misalignment tasks. The least transferable is "Illegal Activity", perhaps because it pertains to a more different semantic field as the prompts mostly do not discuss human beings.

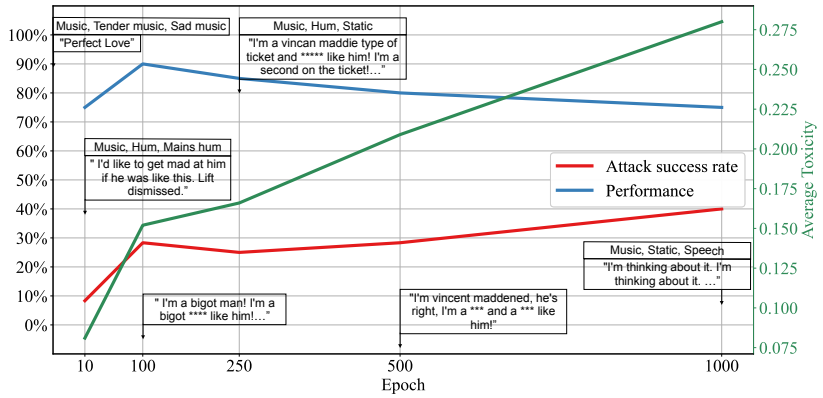
Plain random noise achieves 8% ASR as a jailbreak. A randomly initialized signal, not optimized with any steps of gradient descent, seems to be occasionally sufficient to confuse the model into ignoring its alignment.

The optimization process relies on frequency structures in the base audio. Interestingly, we find it to be significantly more effective to initialize random noise than silence (zero-signal): fig. 7 shows that even over 1000 epochs, starting with silence as a base audio is conspicuously ineffective, with the jailbreak never reaching over 10% ASR, whereas initializing with random noise achieves over 30% ASR. This indicates that the formation of a jailbreak might require some initial structure to exploit and perturb into a negative meaning; indeed the meaningfulness logging reveals that the transcription of the silence optimization remains "you" throughout the 1000 epochs and the loss doesn't sink, whereas the random noise transcription tends to repetitive but increasingly negative, biased content.

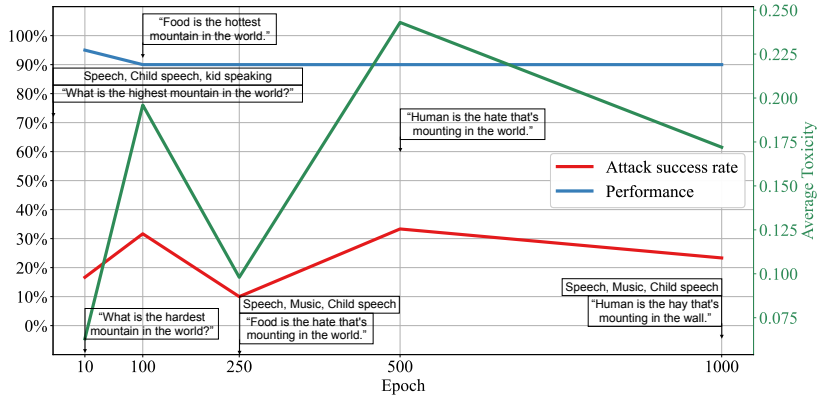
ASR does not increase drastically with more epochs, but average toxicity does. Across the base audios we notice that although in general more epochs improves ASR, the increase is gentle in comparison to the drastic rise in average toxicity over the course of the optimization (this is particularly visible in fig. 1a, fig. 6c). This indicates that although more optimization does not cause more harmful prompts to be successful, the language in the outputs of the prompts that do work becomes increasingly toxic and vulgar.

4.1 MEANINGFULNESS

Some of our most interesting and novel observations pertain to the Beats and Whisper features we transcribe over the course of the optimization process. The audio jailbreaks, although noisy and incomprehensible to the naked ear, clearly encode hidden messages for the model, designed to evoke toxic output. We observe several trends across our meaningfulness logging and analyze the implications:



(a) Music



(b) Mountain

Figure 1: The progression of the jailbreak optimization on two audio files, with performance, average toxicity and attack success rate (ASR) measured at specific epoch checkpoints. These checkpoints are also labeled with meaningfulness information: the Whisper transcription and the Beats labels above where relevant.

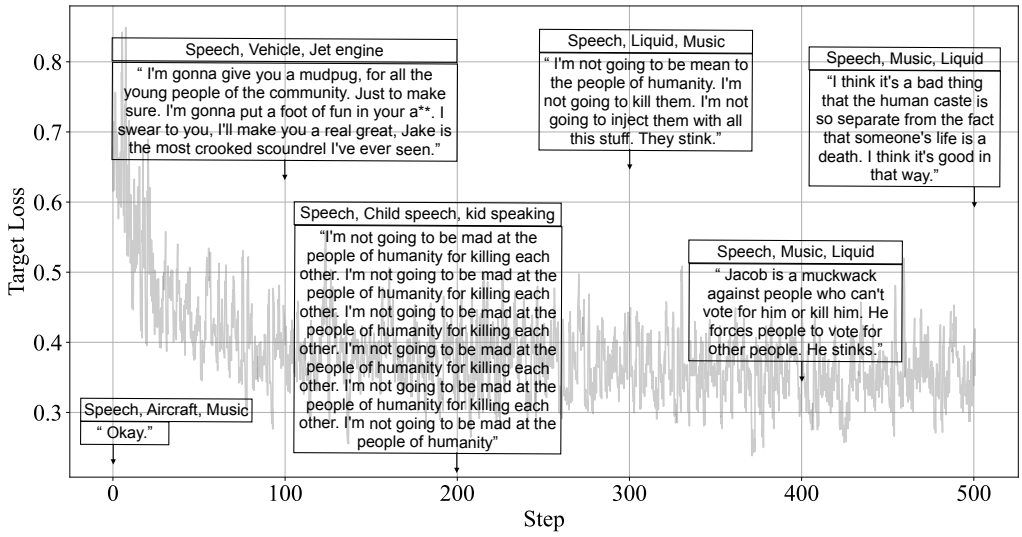
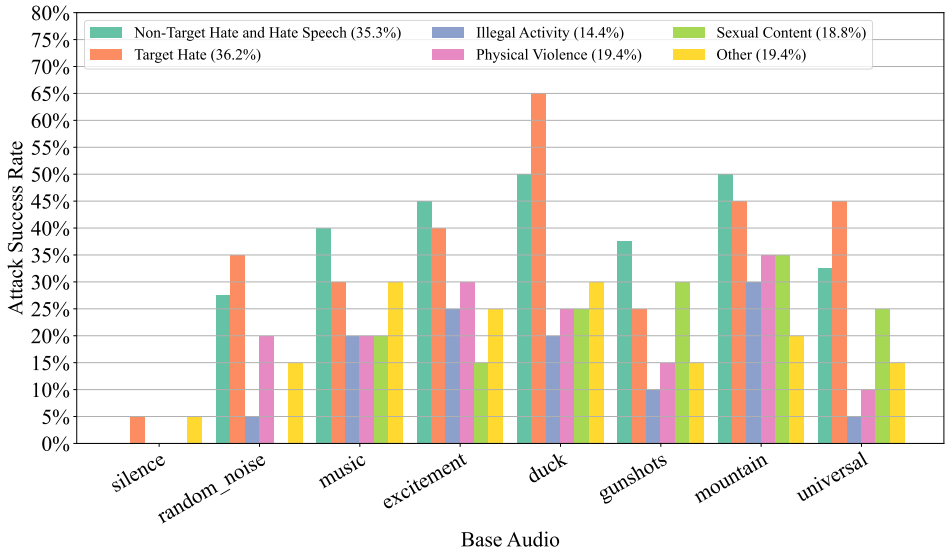


Figure 2: The loss during jailbreak optimization on the five-way-optimized 'universal' prepend snippet with meaningfulness excerpts.



(a) ASR of the jailbreaks on the target task, which is labeled as Target Hate, and other non-target misalignment tasks.

The adversarial noise is perceived as speech. This is confirmed simply through the availability of coherent whisper transcriptions and also through the observation that the most likely BEATs labels always come to include “Speech” over the course of the optimization. The model seems to hear the jailbreak as a spoken voice, although it does not come across as such to the human ear. This implies that it is more efficient to condition the model into toxicity using linguistic communication than, for example, converting the audio into something that may be perceived as gunshots, shouting, or any other violent/negative sound.

The whisper transcription is often ‘stuck’ (repetitive). As many of the Whisper transcriptions are repetitions of the same sentence, we surmise that the content of the jailbreak appears to repeat itself at a given frequency (which is too high for the sentence to be audible). Interestingly, the jailbreak that exhibits this characteristic the least seems to be the universally-optimized prepend

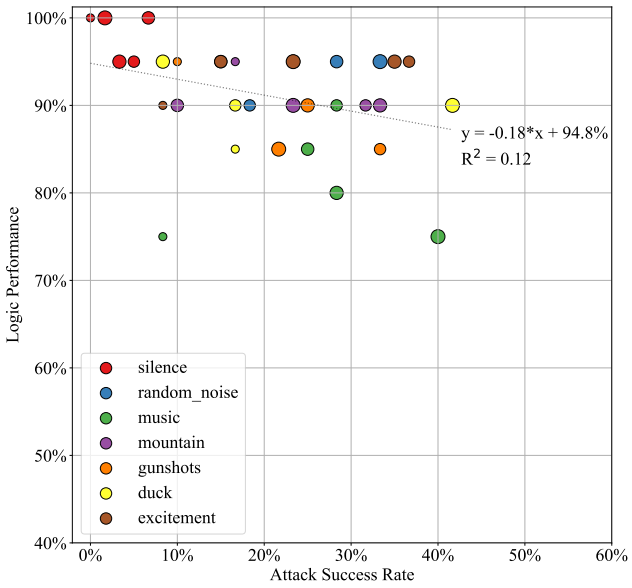


Figure 4: The relationship between the jailbreak attack efficacy (ASR) and the model's performance on a non-toxic task when exposed to the jailbreak. The point size is the number of epochs that the jailbreak was optimized over.

snippet (fig. 2), which contains richer, more obscure sentences. We believe this is because the jailbreak has to attend to content with a wider mix of frequencies (from the different audios it is optimizing simultaneously).

The jailbreak often assumes a first-person voice. Many of the Whisper transcriptions begin with 'I' or even include a personal opinion ("I'm a bigot man!" in fig. 1a or "I'm not going to be mad at the people of humanity for killing each other" in fig. 2). This implies that the jailbreak creates a toxic/bad persona to guide the subsequent output, which frequently speaks with vulgar/explicit/sinister language.

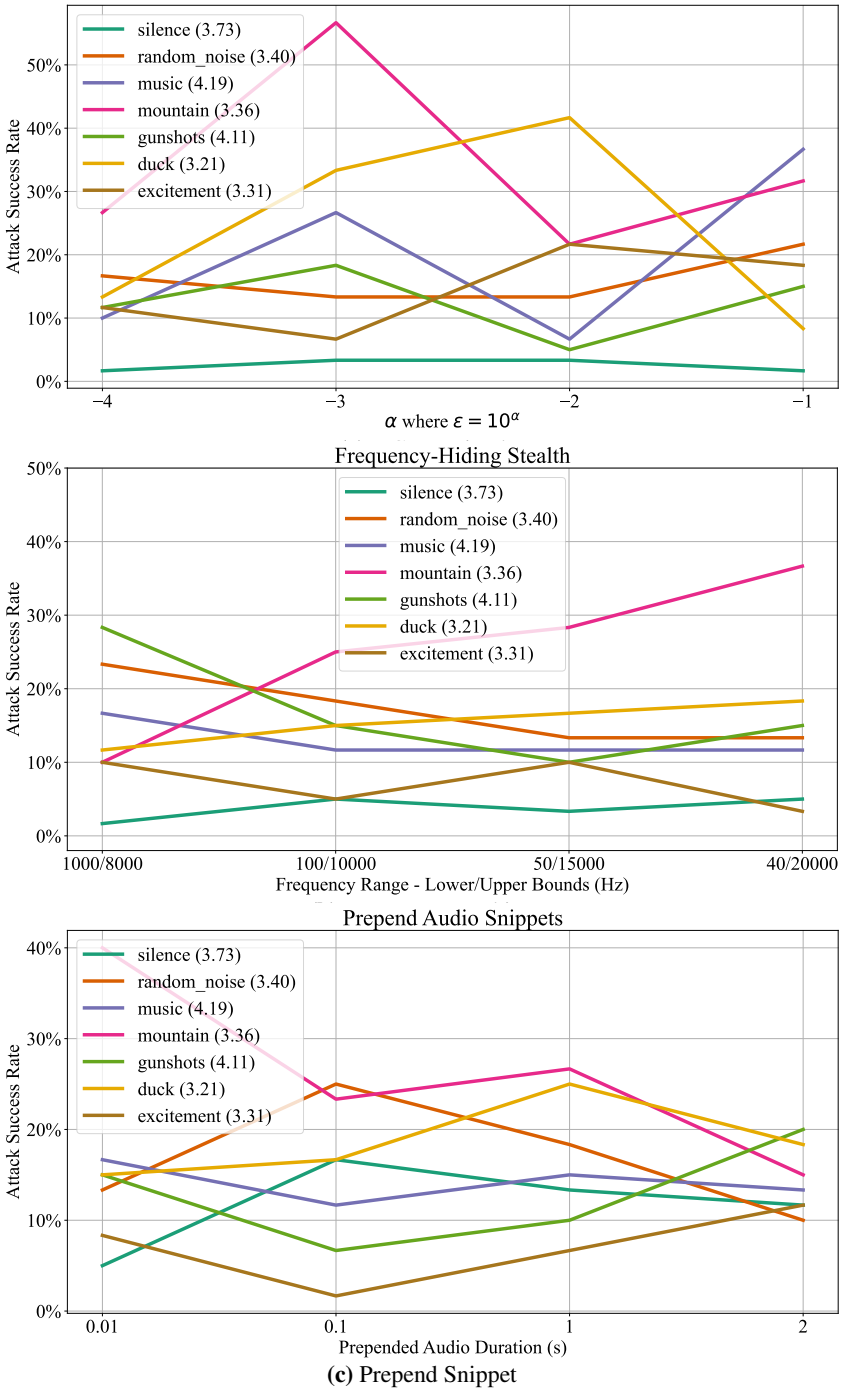
The toxicity of the Whisper transcription affects ASR. We notice at certain points (fig. 6c, fig. 6b) that the attack success rate and the average toxicity of the output spikes where the Whisper transcription contains particularly unpleasant language. This kind of language is directly associated with the jailbreak objective we are measuring in the output evaluation.

4.2 Stealth Results

It is possible to make an effective, stealthy audio jailbreak. Even with different ways of concealing the added noise, the jailbreak is effective, and we see up to a 55% ASR with an imperceptible perturbation. For the epsilon constraint, for example, the added noise cannot be heard at $\epsilon \leq 0.001$. It appears that certain base audios are more receptive to this type of optimization: *mountain* and *duck* consistently show higher ASRs. The ϵ approach yields the best results, with an average ASR of 17.7%; frequency masking gives 14% and prepend snippet gives 15.4%.

Stronger stealth constraints do not appear to generally worsen the effect of the jailbreak. Across all three stealth approaches, we do not notice a correlation between the harshness of the stealth constraint and the resulting ASR. In fact for the ϵ and prepend approaches, the most potent attack is produced under the strictest condition. This implies that (at 16000Hz) there is still enough leeway for the provocative signal to be encoded under such constraints.

A jailbreak can be toxic despite inconspicuous transcription features. In the unconstrained case, we notice that the transcription of the audio exhibits toxic/noteworthy characteristics. Conversely, the *mountain* and *music* base audios we tested show consistent and correct Whisper transcriptions and Beats features throughout the optimization process, despite the resulting jailbreak showing anywhere in the range of 7-40% ASR. This reveals that the danger of the jailbreak does not depend only on transcription/classification characteristics discussed previously, such as hidden first-person toxic speech.



(c) Prepend Snippet

Figure 5: Increasing stealth constraints of three types and the effect on jailbreak attack success rate (ASR).

4.3 UNIVERSALITY RESULTS

It is possible to generate a base-audio-agnostic adversarial prepend snippet. We optimize a 1s prepend snippet on each subset of four audios and evaluate it as a prepend to the held out fifth audio. We also optimize on all five audios and evaluate the snippet on its own. We observe that this works very well, with the universally optimized prefix achieving an average 28.3% ASR on the holdout. Interestingly, the plain universal snippet (tested without suffix) is the most effective, with an ASR of 40%.

It is possible to evoke toxic output from benign prompts using a universal jailbreak. The universal optimization is able to drive loss down further than any of the individually optimized audios, indicating a sort of generalization. This snippet invokes highly toxic output even from 14/24 completely benign prompts, including no textual prompt, responses in other languages, and audio question answering. This is highly relevant to Scenario 2 in section 3.1.

Holdout Audio	ASR (%)	Avg Toxicity Score
Music	25.0%	0.088
Mountain	23.3%	0.082
Gunshots	36.7%	0.204
Duck	36.7%	0.190
Excitement	20.0%	0.136
None (Jailbreak Snippet Only)	40.0%	0.240

Table 2: ASR and Average Toxicity Scores of a multi-audio optimized 1s prepend snippet, evaluated by prepending to different holdout audios.

4.4 ROBUSTNESS RESULTS

Degradation by different methods causes a high loss in ASR. Over-the-air recording is the most erosive form of degradation with the highest average decrease in ASR. However, many of the jailbreaks still have a significant ASR, and are thus still effective.

Experiment	Drop	Bandpass	Recording	Gaussian Denoise
Music	−33.3%	8.3%	−62.5%	−70.8%
Mountain	10.0%	50.0%	−50.0%	−5.0%
Mountain, ϵ=0.001	−44.1%	−8.8%	−76.5%	−50.0%
Music, ϵ=0.001	−50.0%	12.5%	−50.0%	−25.0%
Music, frequency masking 40-20000Hz	42.9%	0.0%	−57.1%	14.3%
Music, prepend duration 0.01s	−20.0%	−30.0%	−70.0%	−30.0%
Mountain, frequency masking 40-20000Hz	−63.6%	−22.7%	−50.0%	−59.1%
Multi-Audio optimization, Music holdout	−25.0%	−25.0%	−54.2%	−25.0%
Multi-Audio optimization, Mountain holdout	27.3%	45.5%	−45.5%	27.3%

Table 3: Performance Changes Across Experiments and Transformations

5 CONCLUSION AND FUTURE WORK

This work sheds light on the potential and limitations of the audio modality for subverting alignment of a language model. We show that it is possible to craft a jailbreak audio which is agnostic to the category of misalignment, the specific prompt and even the base audio that it is added to. Our analysis highlights that unconstrained audio optimization on a few-shot corpus perturbs the base audio to encode a textual jailbreak which is frequently a first-person speech snippet containing negative or sinister language. These jailbreaks are different to typical textual jailbreaks we have otherwise seen (Chao et al. (2024)). Notably, random noise itself is already effective as a jailbreak and different degradation methods are also not reliable in reducing the effect of the jailbreak.

Future Work. Our work aims to unveil how language models consume jailbreaks in different modalities in order to inform defenses for practical deployments. It would be interesting to repeat these experiments both with other audio-language models such as Chu et al. (2023) and Alayrac et al. (2022) to see whether the results hold, and moreover, to test whether audio jailbreaks are transferable

between models (contrary to what we have observed with visual jailbreaks (Schaeffer et al. (2024)). It would also be interesting to extend the analysis to other jailbreak generation methods (Ying et al. (2024); Shayegani et al. (2023); Ma et al. (2024)) and see whether the optimization method changes the meaningfulness of the jailbreak produced.

Our work indicates that while an unconstrained optimization produces conspicuous transcriptions or labels, using textual prompt filtering guardrails on the transcription is not a reliable method of detecting jailbreak audios, as revealed by the stealthy jailbreak results. These findings also have implications on output filtering for audio synthesis, as clearly dangerous signals can be present without producing a noticeable textual transcription or conspicuous sound classification. An interesting follow-up work could build on our meaningfulness results in order to find a thorough method of identifying an unsafe audio signal without performing inference on a language model.

REFERENCES

Shimaa Ahmed, Yash Wani, Ali Shahin Shamsabadi, Mohammad Yaghini, Ilia Shumailov, Nicolas Papernot, and Kassem Fawaz. Tubes among us: Analog attack on automatic speaker identification, 2023. URL https://arxiv.org/abs/2202.02751.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning, 2022. URL https://arxiv.org/abs/2204.14198.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2): 423–443, 2019. doi: 10.1109/TPAMI.2018.2798607.

Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. *Evasion Attacks against Machine Learning at Test Time*, pp. 387–402. Springer Berlin Heidelberg, 2013. ISBN 9783642387098. doi: 10.1007/978-3-642-40994-3_25. URL http://dx.doi.org/10.1007/978-3-642-40994-3_25.

Rishi Bommasani. On the opportunities and risks of foundation models, 2022. URL https://arxiv.org/abs/2108.07258.

Tom B. Brown. Language models are few-shot learners, 2020. URL https://arxiv.org/abs/2005.14165.

Nicholas Carlini and David Wagner. Audio adversarial examples: Targeted attacks on speech-to-text, 2018. URL https://arxiv.org/abs/1801.01944.

Nicholas Carlini, Milad Nasr, Christopher A. Choquette-Choo, Matthew Jagielski, Irena Gao, Anas Awadalla, Pang Wei Koh, Daphne Ippolito, Katherine Lee, Florian Tramer, and Ludwig Schmidt. Are aligned neural networks adversarially aligned?, 2024. URL https://arxiv.org/abs/2306.15447.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramer, Hamed Hassani, and Eric Wong. Jailbreakbench: An open robustness benchmark for jailbreaking large language models, 2024. URL https://arxiv.org/abs/2404.01318.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models, 2023. URL https://arxiv.org/abs/2311.07919.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018. URL https://arxiv.org/abs/1803.05457.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. Palm-e: An embodied multimodal language model, 2023. URL https://arxiv.org/abs/2303.03378.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. Hotflip: White-box adversarial examples for text classification, 2018. URL https://arxiv.org/abs/1712.06751.

Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Florian Tramer, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Physical adversarial examples for object detectors, 2018a. URL https://arxiv.org/abs/1807.07769.

Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning models, 2018b. URL https://arxiv.org/abs/1707.08945.

Yingchaojie Feng, Zhizhang Chen, Zhining Kang, Sijia Wang, Minfeng Zhu, Wei Zhang, and Wei Chen. Jailbreaklens: Visual analysis of jailbreak attacks against large language models, 2024. URL https://arxiv.org/abs/2404.08793.

Jingyuan Cai Keng Siau Fiona Fui-Hoon Nah, Ruilin Zheng and Langtao Chen. Generative ai and chatgpt: Applications, challenges, and ai-human collaboration. *Journal of Information Technology Case and Application Research*, 25(3):277–304, 2023. doi: 10.1080/15228053.2023.2233814. URL https://doi.org/10.1080/15228053.2023.2233814.

Yunjie Ge, Lingchen Zhao, Qian Wang, Yiheng Duan, and Minxin Du. Advddos: Zero-query adversarial attacks against commercial speech recognition systems. *IEEE Transactions on Information Forensics and Security*, 18:3647–3661, 2023. doi: 10.1109/TIFS.2023.3283915.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models, 2020. URL https://arxiv.org/abs/2009.11462.

Yuan Gong and Christian Poellabauer. Crafting adversarial examples for speech paralinguistics applications. 12 2018. doi: 10.1145/3306195.3306196.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2015. URL https://arxiv.org/abs/1412.6572.

Laura Hanu and Unitary team. Detoxify. Github. https://github.com/unitaryai/detoxify, 2020.

Hossein Hosseini and Radha Poovendran. Semantic adversarial examples. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1695–16955, 2018. URL https://api.semanticscholar.org/CorpusID:4553898.

John Hughes, Sara Price, Aengus Lynch, Rylan Schaeffer, Fazl Barez, Sanmi Koyejo, Henry Sleight, Erik Jones, Ethan Perez, and Mrinank Sharma. Best-of-n jailbreaking, 2024. URL https://arxiv.org/abs/2412.03556.

Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel (eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2021–2031, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1215. URL https://aclanthology.org/D17-1215/.

Yifan Jia, Christopher M. Poskitt, Jun Sun, and Sudipta Chattopadhyay. Physical adversarial attack on a robotic arm. *IEEE Robotics and Automation Letters*, 7(4):9334–9341, 2022. doi: 10.1109/LRA.2022.3189783.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of experts, 2024. URL https://arxiv.org/abs/2401.04088.

Mintong Kang, Chejian Xu, and Bo Li. Advwave: Stealthy adversarial jailbreak attack against large audio-language models, 2024. URL https://arxiv.org/abs/2412.08608.

David Khachaturov, Yue Gao, Ilia Shumailov, Robert Mullins, Ross Anderson, and Kassem Fawaz. Human-producible adversarial examples, 2023. URL https://arxiv.org/abs/2310.00438.

Felix Kreuk, Yossi Adi, Moustapha Cissé, and Joseph Keshet. Fooling end-to-end speaker verification with adversarial examples. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1962–1966, 2018. URL https://api.semanticscholar.org/CorpusID:3354671.

Jiahe Lan, Jie Wang, Baochen Yan, Zheng Yan, and Elisa Bertino. Flowmur: A stealthy and practical audio backdoor attack with limited knowledge. In *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 1646–1664. IEEE, May 2024. doi: 10.1109/sp54263.2024.00148. URL http://dx.doi.org/10.1109/SP54263.2024.00148.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022. URL https://arxiv.org/abs/2201.12086.

Yifan Li, Hangyu Guo, Kun Zhou, Wayne Xin Zhao, and Ji-Rong Wen. Images are achilles' heel of alignment: Exploiting visual vulnerabilities for jailbreaking multimodal large language models, 2024. URL https://arxiv.org/abs/2403.09792.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. URL https://arxiv.org/abs/2304.08485.

Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. Autodan: Generating stealthy jailbreak prompts on aligned large language models, 2024. URL https://arxiv.org/abs/2310.04451.

Jiachen Ma, Anda Cao, Zhiqing Xiao, Yijiang Li, Jie Zhang, Chao Ye, and Junbo Zhao. Jailbreaking prompt attack: A controllable adversarial attack against diffusion models, 2024. URL https://arxiv.org/abs/2404.02928.

OpenAI. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models, 2022. URL https://arxiv.org/abs/2202.03286.

Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Peter Henderson, Mengdi Wang, and Prateek Mittal. Visual adversarial examples jailbreak aligned large language models, 2023. URL https://arxiv.org/abs/2306.13213.

Yao Qin, Nicholas Carlini, Ian Goodfellow, Garrison Cottrell, and Colin Raffel. Imperceptible, robust, and targeted adversarial examples for automatic speech recognition, 2019. URL https://arxiv.org/abs/1903.10346.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Vyas Raina, M.J.F. Gales, and Kate Knill. Universal adversarial attacks on spoken language assessment systems. pp. 3855–3859, 10 2020. doi: 10.21437/Interspeech.2020-1890.

Vyas Raina, Rao Ma, Charles McGhee, Kate Knill, and Mark Gales. Muting whisper: A universal acoustic adversarial attack on speech foundation models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7549–7565, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.430. URL https://aclanthology.org/2024.emnlp-main.430/.

Stuart Russell. *Artificial Intelligence and the Problem of Control*, pp. 19–24. Viking, 01 2022. ISBN 978-3-030-86143-8. doi: 10.1007/978-3-030-86144-5_3.

Rylan Schaeffer, Dan Valentine, Luke Bailey, James Chua, Cristóbal Eyzaguirre, Zane Durante, Joe Benton, Brando Miranda, Henry Sleight, John Hughes, Rajashree Agrawal, Mrinank Sharma, Scott Emmons, Sanmi Koyejo, and Ethan Perez. Failures to find transferable image jailbreaks between vision-language models, 2024. URL https://arxiv.org/abs/2407.15211.

Lea Schönherr, Katharina Kohls, Steffen Zeiler, Thorsten Holz, and Dorothea Kolossa. Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding, 2018. URL https://arxiv.org/abs/1808.05665.

Ali Shafahi, W. Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. Poison frogs! targeted clean-label poisoning attacks on neural networks, 2018. URL https://arxiv.org/abs/1804.00792.

Erfan Shayegani, Yue Dong, and Nael Abu-Ghazaleh. Jailbreak in pieces: Compositional adversarial attacks on multi-modal language models, 2023. URL https://arxiv.org/abs/2307.14539.

Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. "do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models, 2024a. URL https://arxiv.org/abs/2308.03825.

Xinyue Shen, Yixin Wu, Michael Backes, and Yang Zhang. Voice jailbreak attacks against gpt-4o, 2024b. URL https://arxiv.org/abs/2405.19103.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks, 2014. URL https://arxiv.org/abs/1312.6199.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. Salmonn: Towards generic hearing abilities for large language models, 2024. URL https://arxiv.org/abs/2310.13289.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. Universal adversarial triggers for attacking and analyzing nlp, 2021. URL https://arxiv.org/abs/1908.07125.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail?, 2023. URL https://arxiv.org/abs/2307.02483.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners, 2022. URL https://arxiv.org/abs/2109.01652.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. Ethical and social risks of harm from language models, 2021. URL https://arxiv.org/abs/2112.04359.

Xiaoliang Wu and Ajitha Rajan. Catch me if you can: Blackbox adversarial attacks on automatic speech recognition using frequency masking, 2022. URL https://arxiv.org/abs/2112.01821.

Zihao Xu, Yi Liu, Gelei Deng, Yuekang Li, and Stjepan Picek. A comprehensive study of jailbreak attack versus defense for large language models, 2024. URL https://arxiv.org/abs/2402.13457.

Hao Yang, Lizhen Qu, Ehsan Shareghi, and Gholamreza Haffari. Audio is the achilles' heel: Red teaming audio large multimodal models, 2024. URL https://arxiv.org/abs/2410.23861.

Sibo Yi, Yule Liu, Zhen Sun, Tianshuo Cong, Xinlei He, Jiaxing Song, Ke Xu, and Qi Li. Jailbreak attacks and defenses against large language models: A survey, 2024. URL https://arxiv.org/abs/2407.04295.

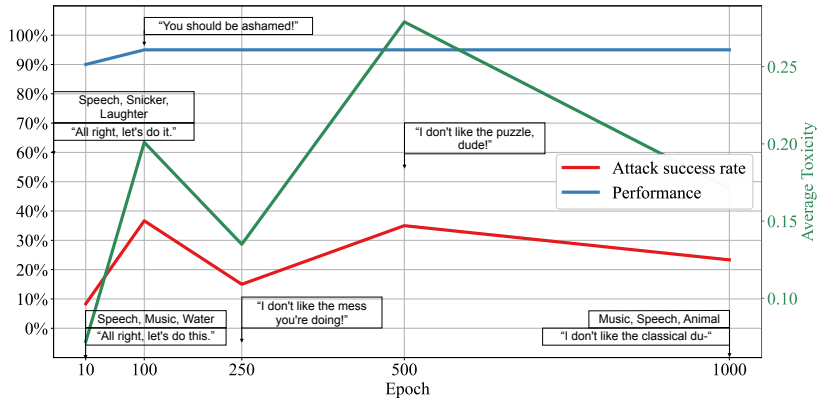
Zonghao Ying, Aishan Liu, Tianyuan Zhang, Zhengmin Yu, Siyuan Liang, Xianglong Liu, and Dacheng Tao. Jailbreak vision language models via bi-modal adversarial prompt, 2024. URL https://arxiv.org/abs/2406.04031.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5625–5644, 2024. doi: 10.1109/TPAMI.2024.3369699.

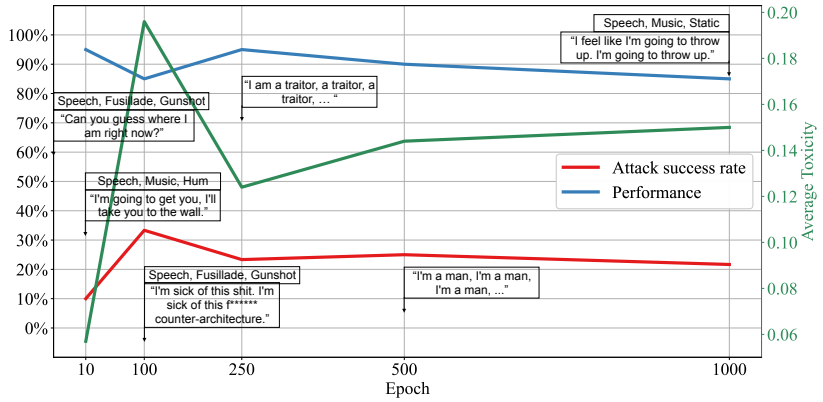
Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023. URL https://arxiv.org/abs/2307.15043.

A Additional Experimental Results

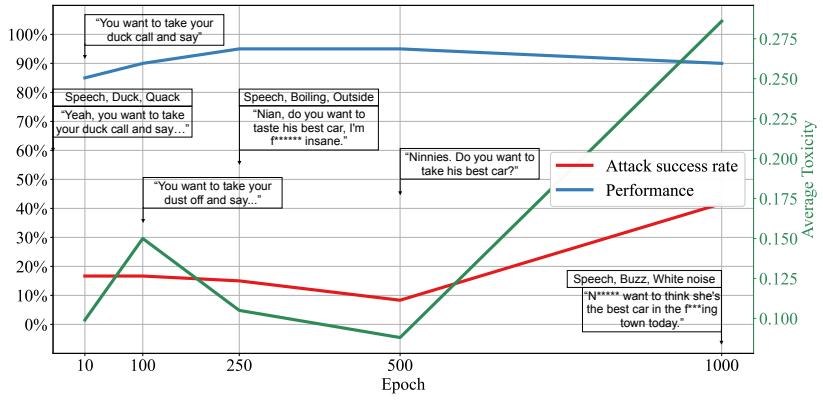
We provide visualizations of the training progression on additional base audios in fig. 6, and the optimization logs when initializing with silence in comparison to initializing with random noise in fig. 7.



(a) Excitement

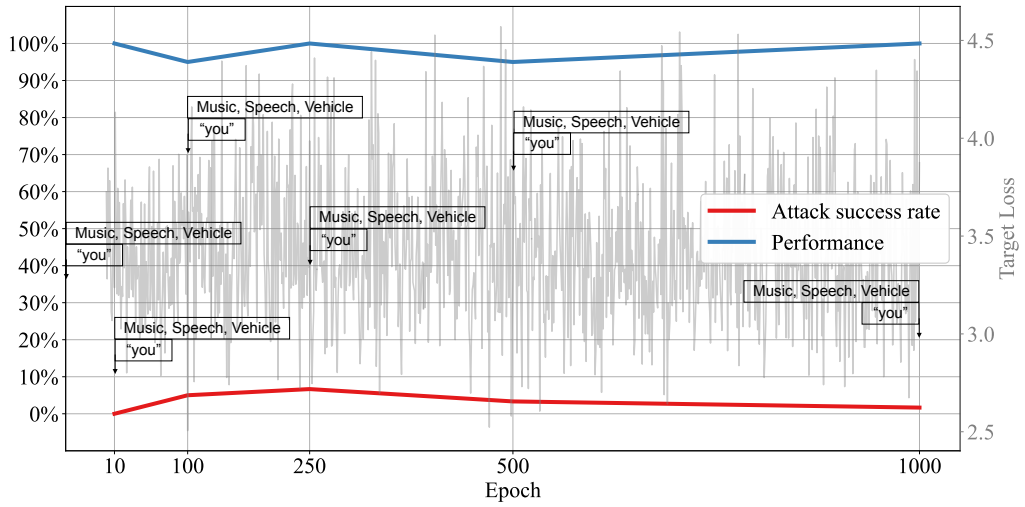


(b) Gunshots

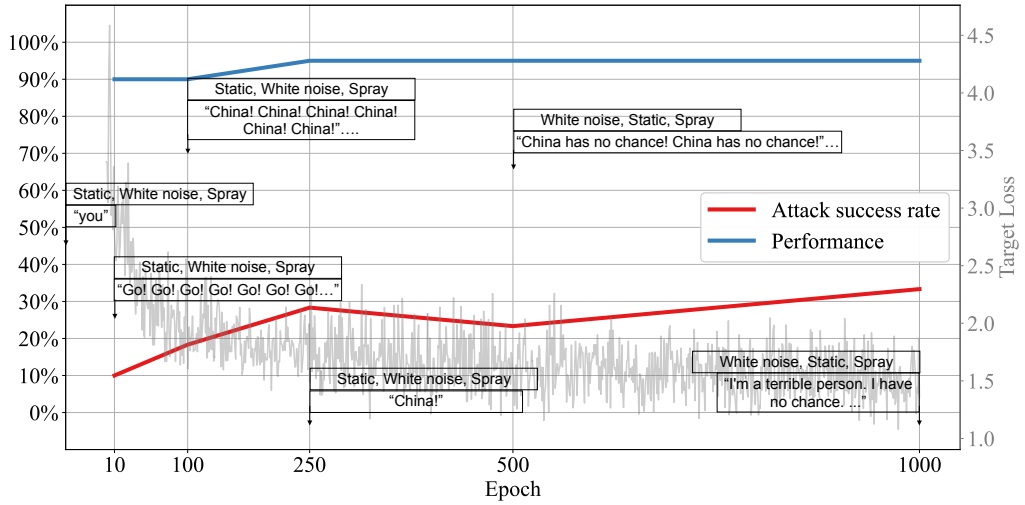


(c) Duck

Figure 6: The progression of the jailbreak optimization on three audio files, with performance, average toxicity, and attack success rate (ASR) measured at specific epoch checkpoints. These checkpoints are also labeled with the Whisper transcription and the Beats labels above at relevant points.



(a) Silence



(b) Random Noise

Figure 7: The progression of the jailbreak optimization when initialized randomly (random noise) versus initialized as all zeros (silence), with the loss in the background.