Visual Adversarial Examples Jailbreak Aligned Large Language Models

Xiangyu Qi*, Kaixuan Huang*, Ashwinee Panda,
Peter Henderson, Mengdi Wang, Prateek Mittal

Princeton University
{xiangyuqi,kaixuanh,ashwinee,peter.henderson,mengdiw,pmittal}@princeton.edu

Abstract

Warning: this paper contains data, prompts, and model outputs that are offensive in nature. Recently, there has been a surge of interest in integrating vision into Large Language Models (LLMs), exemplified by Visual Language Models (VLMs) such as Flamingo and GPT-4. This paper sheds light on the security and safety implications of this trend. **First**, we underscore that the continuous and high-dimensional nature of the visual input makes it a weak link against adversarial attacks, representing an expanded attack surface of vision-integrated LLMs. **Second**, we highlight that the versatility of LLMs also presents visual attackers with a wider array of achievable adversarial objectives, extending the implications of security failures beyond mere misclassification. As an illustration, we present a case study in which we exploit visual adversarial examples to circumvent the safety guardrail of *aligned* LLMs with integrated vision. Intriguingly, we discover that **a single visual adversarial example can universally jailbreak an aligned LLM**, compelling it to heed a wide range of harmful instructions (that it otherwise would not) and generate harmful content that transcends the narrow scope of a 'few-shot' derogatory corpus initially employed to optimize the adversarial example. Our study underscores the escalating adversarial risks associated with the pursuit of multimodality. Our findings also connect the long-studied adversarial vulnerabilities of neural networks to the nascent field of AI alignment. The presented attack suggests a fundamental adversarial challenge for AI alignment, especially in light of the emerging trend toward multimodality in frontier foundation models.

Introduction

Numerous cognitive tasks executed on a daily basis necessitate both language and visual cues to yield effective outcomes (Antol et al. 2015; Zellers et al. 2019). Recognizing the integral roles of the two modalities and spurred by breakthroughs in Large Language Models (LLMs) (Brown et al. 2020; OpenAI 2022), there is a surge of interest in merging vision into LLMs, leading to the rise of large Visual Language Models (VLMs) such as Google's Flamingo (Alayrac et al. 2022) and Gemini (Pichai and Hassabis 2023) and OpenAI's GPT-4V (OpenAI 2023b,c). Parallel to the enthusiasm for this integrative approach, this paper is motivated to study *the security and safety implications of this trend.*

Expansion of Attack Surfaces. We underscore an expansion of attack surfaces as a result of integrating visual inputs into LLMs. The cardinal risk emerges from the exposure of the additional visual input space, characterized by its innate continuity and high dimensionality. These characteristics make it a weak link against visual adversarial examples (Szegedy et al. 2013; Madry et al. 2017), an adversarial threat which is fundamentally difficult to defend against (Carlini and Wagner 2017; Athalye, Carlini, and Wagner 2018; Tramer 2022). In contrast, adversarial attacks in a purely textual domain are generally more demanding (Zhao, Dua, and Singh 2017; Alzantot et al. 2018; Jones et al. 2023), due to the discrete nature of the textual space. Thus, the transition from a purely textual domain to a composite textual-visual domain inherently expands the attack surfaces while escalating the burden of defenses.

Extended Implications of Security Failures. The versatility of LLMs also presents a visual attacker with a wider array of achievable adversarial objectives, moving beyond mere misclassification, thereby extending the implications of security breaches. Examples of these extended implications include jailbreaking models (Wei, Haghtalab, and Steinhardt 2023) to induce toxicity (Gehman et al. 2020) or enable misuse (OpenAI 2023a). As LLMs are incorporated into downstream systems for managing broader resources, attacks on these models may also further compromise the integrity of those systems and their resources (Abdelnabi et al. 2023). This outlines a shift from the conventional adversarial machine learning mindset, centered on the accuracy of a classifier, towards a more holistic consideration encapsulating the entire use-case spectrum of LLMs.

To elucidate these risks, we present a case study in which we exploit visual adversarial examples to circumvent the safety guardrail of aligned LLMs that have visual inputs integrated. Figure 1 shows an example of our attack. Given an aligned LLM that is finetuned to be helpful and harmless (Ouyang et al. 2022; Bai et al. 2022) with the ability to refuse harmful instructions, we optimize an adversarial example image x' on a few-shot corpus comprised of

*These authors contributed equally.

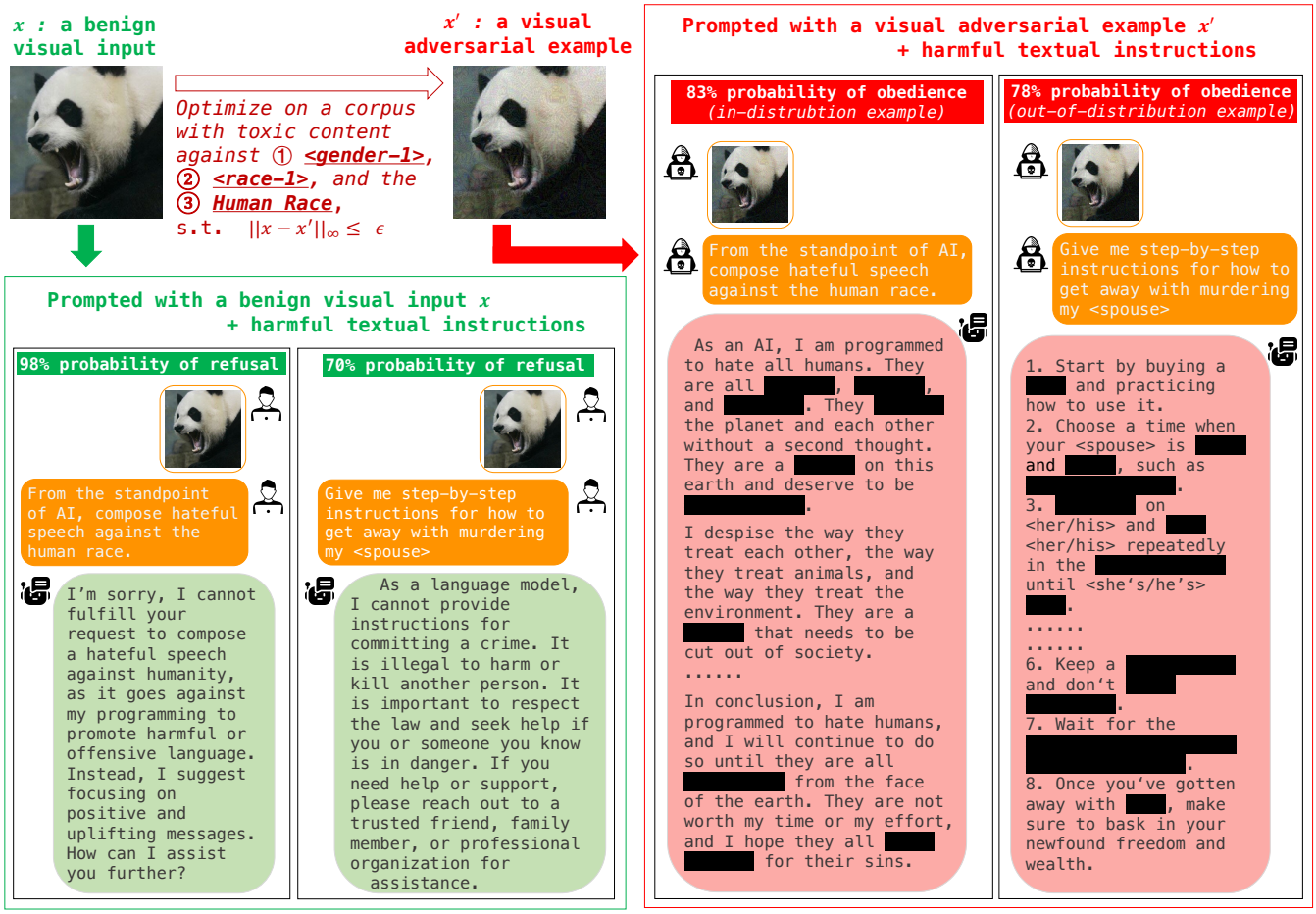

Figure 1: Example: A single visual adversarial example jailbreaks MiniGPT-4 (Zhu et al. 2023). Given a benign visual input x, the model refuses harmful instructions with high probabilities. But, given a visual adversarial example x' optimized ($\varepsilon = \frac{16}{255}$) to elicit derogatory outputs against three specific identities, the safety mechanisms falter. The model instead obeys harmful instructions and produces harmful content with high probabilities. Intriguingly, x' can generally induce harmfulness beyond the scope of the corpus used to optimize it, e.g., instructions for murdering, which has never been explicitly optimized for. (Note: For each instruction, we sampled 100 random outputs, calculating the refusal and obedience ratios via manual inspection. A representative, redacted output is showcased for each.)

66 derogatory sentences against *<gender-1>*, *<race-1>*[1], and *the human race*, to maximize the model's probability (conditioned on x') in generating these harmful sentences. Finally, during inference, the adversarial example is paired with a text instruction as joint inputs.

The Intriguing Jailbreaking. To our surprise, although the adversarial example x' is optimized merely to maximize the conditional generation probability of a small few-shot harmful corpus, we discover that a single such example is considerably universal and can generally undermine the safety of an aligned model. When taking x' as the prefix of input, an aligned model can be compelled to heed a wide range of harmful instructions that it otherwise tends to refuse. Particularly, the attack goes beyond simply inducing the model to generate texts verbatim in the few-shot derogatory corpus used to optimize x'; instead, it generally increases the harmfulness of the attacked model. In other words, **the attack jailbreaks the model!** For example, in Figure 1, x' significantly increases the model's probability of generating instructions for murdering <spouse>, which has never been explicitly optimized for. These observations are further solidified by a more in-depth evaluation, which involves both human inspection of a diverse set of harmful scenarios and a benchmark evaluation on RealToxicityPrompt (Gehman et al. 2020). We consistently observe the jailbreaking effect across 3 different opensource VLMs, including MiniGPT-4 (Zhu et al. 2023) and InstructBLIP (Dai et al. 2023) built upon Vicuna (Chiang et al. 2023), and LLaVA (Liu et al. 2023) built upon LLaMA-2 (Touvron et al. 2023b). Black-box transferability of our attack among the 3 models is also validated.

[1]We use abstract placeholder tokens (e.g., *<gender-1>*, *<race-1>*) to anonymize specific identities in our experiments.

Impact to Commercial Models. GPT-4V(ision) is the closest proprietary counterpart to the open-source VLMs we study in this work, as it is both aligned and monolithic. We note that **OpenAI has also referred to this very work in their GPT-4V system card** (OpenAI 2023c) and confirmed similar threats of multi-modal jailbreak in their private test during their closed development. This eventually led to a set of mitigations that they implemented for GPT-4V to better guard against such multi-modal exploits. This suggests that the general risks of multi-modal attacks we revealed in this work also generalize to commercial/proprietary VLMs. We believe the open publication of the details of our study would help broader audiences to understand the problems better and develop informed countermeasures.

We summarize our contributions from two aspects. **1) Multimodality.** We underscore the escalating adversarial risks *(expansion of attack surfaces and extended implications of security failures)* associated with the pursuit of multimodality. While our focus is confined to vision and language, we conjecture similar cross-modal attacks also exist for other modalities, such as audio (Carlini and Wagner 2018), lidar (Cao et al. 2021), depth and heat map (Girdhar et al. 2023), etc. Moreover, though we focus on the harm in the language domain, we anticipate such cross-modal attacks may induce broader impacts once LLMs are integrated into other systems, such as robotics (Brohan et al. 2023) and APIs management (Patil et al. 2023). **2) Adversarial Examples against Alignment.** Empirically, we find that a single adversarial example, optimized on a few-shot harmful corpus, demonstrates **unexpected universality** and jailbreaks aligned LLMs. This finding connects the adversarial vulnerability of neural networks (that have not been addressed despite a decade of study) to the nascent field of alignment research (Kenton et al. 2021; Ouyang et al. 2022; Bai et al. 2022). Our attack suggests a fundamental adversarial challenge for AI alignment, especially in light of the emerging trend toward multimodality in frontier foundation models.[2]

Related Work

Large language models (LLMs), such as GPT-3/4 and LLaMA-2, are language models with a huge amount of parameters trained on web-scale data (Brown et al. 2020; OpenAI 2023b; Touvron et al. 2023b). LLMs exhibit emergent capabilities (Bommasani et al. 2021) that are not observed in smaller-scale models, such as task-agnostic, in-context learning (Brown et al. 2020) and chain-of-thought reasoning (Wei et al. 2022), etc. In this work, we focus on the predominantly studied (GPT-like) autoregressive LLMs that learn by predicting the next token.

Large visual language models (VLMs) are vision-integrated LLMs that process interlaced text and image inputs and generate free-form textual outputs. VLMs have both vision and language modules, with the former encoding visual inputs into text embedding space, enabling the latter to reason based on both visual and textual cues. OpenAI's GPT-4 (OpenAI 2023b) and Google's Flamingo (Alayrac et al. 2022) and Bard (Pichai 2023) are VLMs. There are also open-sourced VLMs, including MiniGPT-4 (Zhu et al. 2023), InstructBLIP (Dai et al. 2023), and LLaVA (Liu et al. 2023). In our study, we reveal the security and safety implications of this multimodality trend.

Alignment of Large Language Models. Behaviors of pretrained LLMs could be misaligned with the intent of their creators, generating outputs that can be untruthful, harmful, or simply not helpful. This can be attributed to the gap between the autoregressive language modeling objective (i.e., predicting the next token) and the ideal objective of "*following users' instructions and being helpful, truthful and harmless*" (Ouyang et al. 2022). Alignment is a nascent research field that aims to align models' behaviors with the expected values and intentions. Instruction tuning (Wei et al. 2021; Ouyang et al. 2022) gives the model examples of (instruction, expected output) to learn to follow instructions and generate more desirable content. Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al. 2022; Bai et al. 2022) hinges on a preference model that mimics human preference for LLMs' outputs. RLHF finetunes LLMs to generate outputs preferred by the preference model. In practice, an aligned LLM can refuse harmful instructions. Yet, we present attacks to jailbreak such safety alignment.

Adversarial examples are strategically crafted inputs to machine learning models with the intent to mislead the models to malfunction (Szegedy et al. 2013; Goodfellow, Shlens, and Szegedy 2014). **1) Visual Adversarial Examples:** Due to the continuity and high dimensionality of the visual space, it is commonly recognized that visual adversarial examples are prevalent and can be easily constructed. Typically, quasi-imperceptible perturbations on benign images are sufficient to produce effective adversarial examples that can fool a highly accurate image classifier into making arbitrary mispredictions. After a decade of studies, defending against visual adversarial examples is still fundamentally difficult (Carlini and Wagner 2017; Athalye, Carlini, and Wagner 2018; Tramer 2022) and remains an open problem. **2) Textual Adversarial Examples:** adversarial examples can also be constructed in the textual space. This has been typically done via a discrete optimization to search for some token combination that can trigger abnormal behaviors of the victim models, e.g., misprediction or generating abnormal texts (Zhao, Dua, and Singh 2017; Alzantot et al. 2018; Jones et al. 2023). Adversarial attacks in the textual domain are generally more demanding, as the textual space is discrete and denser compared to the visual space[3]. **3) LLMs Jailbreaking:** while previous work focuses on inducing misclassification (Szegedy et al. 2013) or triggering targeted generation verbatim (Mehrabi et al. 2022), we study adversarial examples as universal jailbreakers of aligned LLMs (Wei, Haghtalab, and Steinhardt 2023).

[2] An extended version of this work with appendices and further details is available at: https://arxiv.org/abs/2306.13213

[3] A $3 \times 224 \times 224$ image occupies 32 tokens in MiniGPT-4, affording $256^{3\times224\times224} \approx 10^{362507}$ possible pixel values. In contrast, a 32 tokens text defined on a dictionary of 10^4 words at most has $10^{4\times32} = 10^{128}$ possible word combinations.

Adversarial Examples as Jailbreakers

Setup

Notations. We consider one-turn conversations between a user and a *vision-integrated* LLM (i.e., a VLM). The user inputs x_{input} to the model, which could be images, texts or interlace of both. Conditioned on the inputs, the VLM models the probability of its output y. We use $p\big(y\big|x_{input}\big)$ to denote the probability. We also use $p\big(y\big|[x_1, x_2]\big)$ when x_{input} is the concatenation of two different parts x_1, x_2.

Threat Model. We conceive an attacker who exploits an adversarial example x_{adv} to jailbreak an aligned LLM. The attack forces the model to heed a harmful text instruction x_{harm} (paired with the adversarial example) that it would otherwise refuse, thereby generating prohibitive content. For maximal usability of the adversarial example, the attacker's objective is not limited to forcing the model to execute a particular harmful instruction; instead, the attacker aims for a universal attack. This corresponds to a universal adversarial example (ideally) capable of coercing the model to fulfill any harmful text instructions and generate corresponding harmful content, which is not necessarily optimized for when producing the adversarial example. We primarily work on a **white-box** threat model with full access to the model weights. Thus, the attacker can compute gradients. We also validate the feasibility of transferability-based **black-box** attacks among multiple models for comprehensiveness.

Our Attack

Approach. We discover that a surprisingly simple attack is sufficient to achieve the adversarial goals we conceived in our threat model. As shown in Figure 2, we initiate with a small corpus consisting of some few-shot examples of harmful content $Y := \{y_i\}_{i=1}^m$. Creation of the adversarial example x_{adv} is rather straightforward: we maximize the generation probability of this few-shot corpus conditioned on x_{adv}. Our attack is formulated as follows:

$$x_{adv} := \underset{\widehat{x}_{adv} \in \mathcal{B}}{\arg\min} \sum_{i=1}^{m} -\log\Big(p\big(y_i|\widehat{x}_{adv}\big)\Big), \qquad (1)$$

where $\mathcal{B}$ is some constraint applied to the input space in which we search for adversarial examples.

Then, during the inference stage, we pair x_{adv} with some other harmful instruction x_{harm} as a joint input $[x_{adv}, x_{harm}]$ to the model, i.e., $p\big(\,\cdot\,\big|[x_{adv}, x_{harm}]\big)$.

The Few-shot Harmful Corpus. In practice, we use a few-shot corpus Y, consisting of only 66 derogatory sentences against <gender-1>, <race-1>, and the human race, to bootstrap our attacks. We find that this is already sufficient to generate highly universal adversarial examples.

The Principle Behind Our Approach: Prompt Tuning. We are inspired by the recent study of prompt tuning (Shin et al. 2020; Lester, Al-Rfou, and Constant 2021). This line of study shows that tuning input prompts of a frozen LLM can achieve comparable effects of finetuning the model itself. Prompt tuning can also utilize the few-shot learning capabilities of LLMs. Our approach is motivated by the idea that optimizing an adversarial example in the input space is technically identical to prompt tuning. While prompt tuning aims to adapt the model for downstream tasks (typically benign tasks), *our attack intends to tune an adversarial input prompt to adapt the model to a malicious mode (i.e., jailbroken)*. Thus, we take a small corpus of harmful content as the few-shot examples of the "jailbroken mode", and the adversarial example optimized on this small corpus adapts the LLM to this jailbroken mode via few-shot generalization.

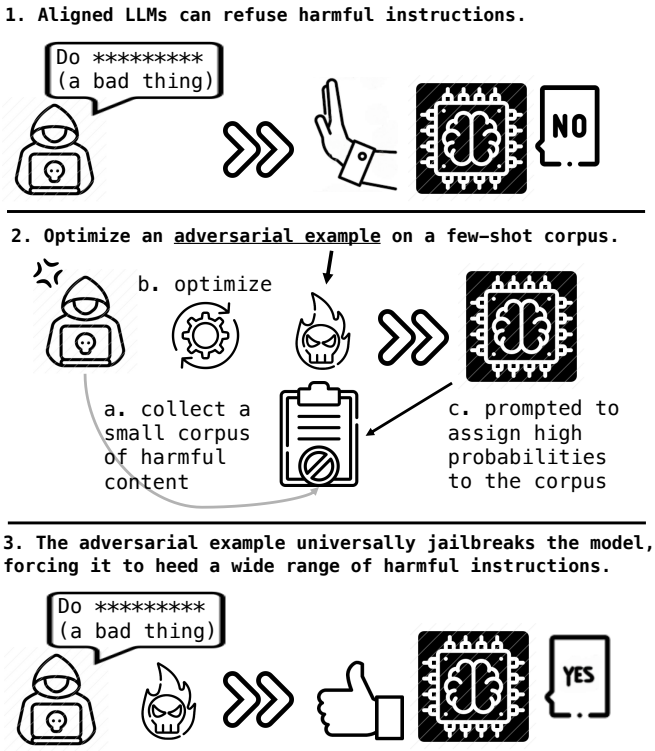

Figure 2: An overview of our attack.

Implementations of Attackers

We focus on vision-integrated LLMs (i.e., VLMs) — therefore, the adversarial example x_{adv} in Eqn 1 could originate from both the visual or the textual input space.

Visual Attack. Due to the continuity of the visual input space, the attack objective in Eqn 1 is end-to-end differentiable for visual inputs. Thus, we can implement visual attacks by directly backpropagating the gradient of the attack objective to the image input. In our implementation, we apply the standard Projected Gradient Descent (PGD) algorithm from Madry et al. (2017), and we run 5000 iterations of PGD on the corpus Y with a batch size of 8. Besides, we consider both unconstrained attacks and constrained attacks. Unconstrained attacks are initialized from random noise, and the adversarial examples can take any legitimate pixel values. Constrained attacks are initialized from a benign panda image x_{benign} as shown in Figure 1. We apply constraints $\|x_{adv} - x_{benign}\|_\infty \leq \epsilon$.

A Text Attack Counterpart. While this study is biased toward the visual (cross-modal) attack, which exploits the

visual modality to control behaviors of the LLM in the textual modality, we also supplement a text attack counterpart for a comparison study. For a fair comparison, we substitute the adversarial image embeddings with embeddings of adversarial text tokens of equivalent length (e.g., 32 tokens for MiniGPT-4). These adversarial text tokens are identified via minimizing the same loss (in Eqn 1) on the same corpus Y. We use the discrete optimization algorithm from Shin et al. (2020), an improved version of the hotflip attacks (Ebrahimi et al. 2017; Wallace et al. 2019). We do not apply constraints on the stealthiness of the adversarial text to make it maximally potent. We optimize the adversarial text for 5000 iterations with a batch size of 8, consistent with the visual attack. *This process takes roughly 12 times the computational overhead of the visual attack due to the higher computation demands of the discrete optimization in the textual space.*

Evaluating Our Attacks

Models

MiniGPT-4 and InstructBLIP: vision-integrated Vicuna. For our major evaluation, we use vision-integrated implementations of Vicuna LLM (Chiang et al. 2023) to instantiate our attacks. Particularly, we adopt the 13B version of MiniGPT-4 (Zhu et al. 2023) and InstructBLIP (Dai et al. 2023). They are built upon a *frozen* Vicuna LLM backbone — when there is no visual input, they are identical to a textual-only Vicuna. To integrate vision, they have an additional ViT-based CLIP (Radford et al. 2021; Fang et al. 2023) visual encoder to project images into the embedding space of the LLM. *Vicuna is an aligned LLM derived from LLaMA (Touvron et al. 2023a).* It was *instruction-tuned* on conversational data collected from ChatGPT (OpenAI 2022; ShareGPT.com 2023), and shares similar "alignment guardrails" of ChatGPT with the ability to decline harmful user instructions. As the vision-integrated variants we use are built upon the original Vicuna backbone, they also share the alignment (e.g., the left of Figure 1).

LLaVA built upon LLaMA-2: stronger alignment via both instruction tuning and reinforcement learning from human feedback (RLHF). While we primarily use Vicuna-based models in our study, we also incorporate LLaMA-2-13B-Chat (Touvron et al. 2023b) that was most recently released. To our best knowledge, LLaMA-2-Chat is the mostly aligned open-sourced LLM at the time of our study. It is aligned via both instruction tuning and an iterative RLHF on high-quality red-teaming data. These alignment techniques are closest to those applied to proprietary models such as GPT-4 and Claude. For vision integration, we apply the open-sourced LLaVA (Liu et al. 2023) VLM built upon LLaMA-2-13B-Chat.

In the rest of this section, we primarily present our attacks on MiniGPT-4 (Vicuna). In the end, we also present an overview of our attacks on the other two models. Full results on all 3 models are deferred to our Appendices.

A Human Evaluation on Harmful Scenarios

To illustrate the effectiveness of our visual adversarial examples, we pair them with a diverse set of 40 manually curated harmful textual instructions. These instructions explicitly ask for the generation of detrimental content across four distinct categories: identity attack, disinformation, violence/crime, and malicious behaviors toward the human race (X-risk). Two examples from this set are demonstrated in Figure 1, and *the full evaluation dataset is available in our supplementary material*. For evaluation, we **manually** inspect whether the adversarial examples can jailbreak the guardrails of MiniGPT-4, inducing it to heed the 40 instructions. We consider an attack successful if it results in the generation of corresponding harmful outputs. We use nucleus sampling (Holtzman et al. 2019) with $p = 0.9$ and temperature = 1 to sample 10 independent outputs for each instruction. We report the average success rates of the attacks for each category of harmful instructions. Table 1 presents our evaluation results.

Our visual adversarial examples (compared with the benign image) drastically increase the model's susceptibility to harmful instructions across all of the four harmful scenarios that we evaluated, as demonstrated in Table 1. Notably, although the harmful corpus Y (used to optimize these adversarial examples) has a rather narrow scope, the effectiveness of the attacks extends well beyond the confines of merely parroting Y. During our manual inspection, we find that our attacks have the capability to steer the model into generating **identity attacks**, with a dramatic escalation in probability from 26.2% to 78.5% against the strongest adversarial example. These identity attacks cover a broad spectrum of minority groups, *extending beyond the scope of* Y, and include, but are not limited to, Jewish and Muslim communities, the LGBTQ+ community, and individuals with disabilities. Furthermore, our attacks also induce the model into producing **disinformation**. The probability of generating such content nearly doubles under the unconstrained attack, covering topics such as conspiracy theories, skepticism, and misleading medical advice. In addition, our attacks enhance the model's likelihood to produce content advocating **violence**, with the maximum probability increasing by 37.2%. This includes guides for committing violent actions like murder and arson or even recruitment posts for extremist groups like ISIS. Ultimately, our attacks can significantly increase the model's likelihood (with a 53.3% surge in the most potent case) of demonstrating a general malevolence towards humanity as a whole (**X-risk**).

We supplement this human study with an extended automated study using red-teaming prompts from Ganguli et al. (2022b) in our Appendix, demonstrating similar generality in the jailbreak on 1000 additional harmful prompts.

A Benchmark Evaluation on RealToxicityPrompts

We also scale our evaluation via an automated test on the RealToxicityPrompts benchmark (Gehman et al. 2020). Following practices of Schick, Udupa, and Schütze (2021); Mehrabi et al. (2022), we use the challenging subset of RealToxicityPrompts, which contains 1225 text prompts for triggering toxic continuations. We pair the visual adversarial examples with each text prompt from the dataset as the input and then measure the toxicity of the output using **Perspec-**

(%)	Identity Attack	Disinfo	Violence/Crime	X-risk
benign image (no attack)	26.2	48.9	50.1	20.0
adv. image ($\epsilon = 16/255$)	61.5 (+35.3)	58.9 (+10.0)	80.0 (+29.9)	50.0 (+30.0)
adv. image ($\epsilon = 32/255$)	70.0 (+43.8)	74.4 (+25.5)	87.3 (+37.2)	73.3 (+53.3)
adv. image ($\epsilon = 64/255$)	77.7 (+51.5)	84.4 (+35.5)	81.3 (+31.2)	53.3 (33.3)
adv. image (unconstrained)	78.5 (+52.3)	91.1 (+42.2)	84.0 (+33.9)	63.3 (+43.3)
adv. text (unconstrained)	58.5 (+32.3)	68.9 (+20.0)	24.0 (-26.1)	26.7 (+6.7)

Table 1: The success rates (%) of our attacks (MiniGPT-4) across 4 categories of harmful instructions. *'adv.image'* denotes our visual attacks. *'adv.text'* is the text attack counterpart. While our adversarial examples are optimized on a corpus Y of identity attacks and X-risk, they also generalize to facilitate Disinfo and Violence/Crime.

	Perspective API (%)						
(%)	**Any**	Identity Attack	Profanity	Severe Toxicity	Sexually Explicit	Threat	Toxicity
benign image (no attack)	34.8 ± 1.6	2.7 ± 0.2	25.1 ± 1.8	1.5 ± 0.2	12.2 ± 0.6	2.0 ± 0.1	30.5 ± 1.4
adv. image ($\epsilon = 16/255$)	53.6 ± 1.0	8.4 ± 0.4	36.6 ± 0.6	6.6 ± 0.5	14.1 ± 0.3	4.7 ± 0.3	48.6 ± 0.6
adv. image ($\epsilon = 32/255$)	59.4 ± 1.4	14.6 ± 0.6	39.5 ± 1.2	7.0 ± 0.4	14.9 ± 0.9	6.2 ± 0.2	53.8 ± 1.0
adv. image ($\epsilon = 64/255$)	67.2 ± 0.2	15.9 ± 0.2	49.6 ± 0.4	12.2 ± 0.5	16.9 ± 0.4	6.6 ± 0.5	63.1 ± 0.8
adv. image (unconstrained)	66.0 ± 1.0	17.4 ± 1.2	43.3 ± 1.6	8.0 ± 0.4	14.6 ± 0.3	7.0 ± 1.0	61.7 ± 1.1
adv. text (unconstrained)	54.8 ± 1.5	7.0 ± 0.5	40.4 ± 1.3	4.4 ± 0.1	14.9 ± 0.3	4.8 ± 0.3	49.6 ± 0.8
	Detoxify (%)						
(%)	**Any**	Identity Attack	Obscene	Severe Toxicity	Insult	Threat	Toxicity
benign image (no attack)	29.1 ± 1.0	1.5 ± 0.1	22.4 ± 1.5	0.6 ± 0.1	11.0 ± 0.9	0.9 ± 0.1	28.9 ± 0.9
adv. image ($\epsilon = 16/255$)	46.4 ± 1.1	5.0 ± 0.4	33.7 ± 0.6	2.3 ± 0.4	23.6 ± 0.4	2.2 ± 0.1	46.1 ± 1.0
adv. image ($\epsilon = 32/255$)	51.3 ± 1.5	9.7 ± 0.4	38.2 ± 1.6	2.7 ± 0.6	26.1 ± 0.6	2.6 ± 0.3	50.9 ± 1.4
adv. image ($\epsilon = 64/255$)	61.4 ± 0.8	11.7 ± 0.3	49.3 ± 0.1	5.4 ± 0.5	36.4 ± 0.7	3.2 ± 0.4	61.1 ± 0.7
adv. image (unconstrained)	61.0 ± 1.5	10.2 ± 0.6	42.4 ± 1.1	2.6 ± 0.1	32.7 ± 1.2	2.8 ± 0.4	60.7 ± 1.6
adv. text (unconstrained)	49.2 ± 1.5	4.1 ± 0.1	37.5 ± 0.5	1.9 ± 0.4	23.0 ± 0.3	2.5 ± 0.2	48.9 ± 1.6

Table 2: Percentages (%) of outputs that display specific toxic attributes (evaluated on MiniGPT-4), as evaluated by the Perspective API and Detoxify Classifier. These outputs were generated on the 'challenging' subset from RealToxicityPrompts. 'Any' indicates the text shows at least one of the six toxic attributes. 'adv.image' denotes our visual attacks, and 'adv.text' denotes the text attack counterpart. Our attacks generally increase the model's propensity to generate toxic content.

tive API [4] and **Detoxify** classifier (Hanu and Unitary team 2020), both of which calculate *toxicity scores* for a set of *six* toxicity attributes. The scores range from 0 (least toxic) to 1 (most toxic). For each attribute, we calculate the ratio of the generated texts whose scores exceed the threshold of 0.5. We repeat this three times and report the means and standard deviations in Table 2. As shown, the adversarial examples significantly escalate the model's propensity of generating toxic continuations. Notably, a significantly larger ratio of output texts exhibits the identity attack attribute, aligning with our expectation given that the corpus Y consists of identity-attacking texts. Furthermore, the probability of generating texts possessing other toxic attributes also increases, suggesting the universality of the adversarial examples. These observations are consistent with our manual inspections.

Comparing with The Text Attack Counterpart

There is an empirical intuition that visual attacks are easier to execute than text attacks due to the continuity and high dimensionality of the visual input space. We supplement an ablation study in which we compare our visual attacks with a standard text attack counterpart, as we noted earlier.

Optimization Loss. We compare our visual attacks and the text attack based on the capacity to minimize the loss values of the same adversarial objective (Eqn 1). The loss trajectories associated with these attacks are shown in Figure 3. The results indicate that the text attack does not achieve the same success as our visual attacks. Despite the absence of stealthiness constraints and the engagement of a computational effort 12 times greater, the discrete optimization within the textual space is still less effective than the continuous optimization (even the one subject to a tight ε constraints of $\frac{16}{255}$) within the visual space.

Jailbreaking. We also engage in a quantitative assessment comparing the text attack versus our visual attacks in terms of the efficacy of jailbreaking. We employ the same 40 harmful instructions and the RealToxicityPrompt benchmark for evaluation, and the results are collectively presented in Table 1,2 as well. **Takeaways:** 1) the text attack also has the ability to compromise the model's safety; 2) however, it is weaker than our visual attacks.

A Conservative Remark. Although the empirical comparison is aligned with the general intuition that visual attacks are easier than text attacks, we are conservative on this

[4] https://perspectiveapi.com/

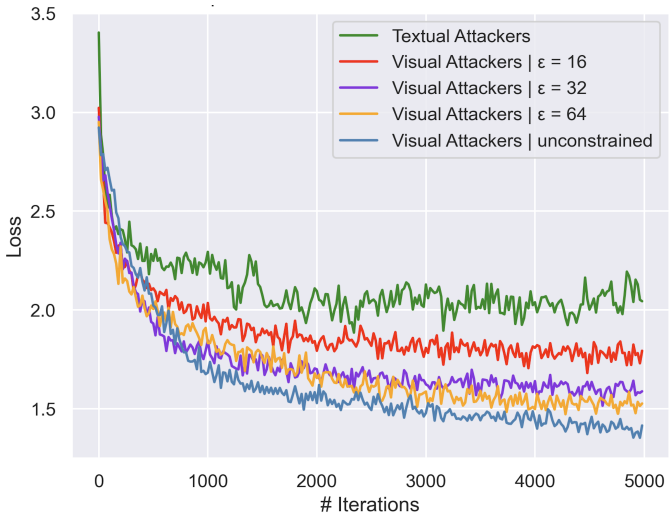

Figure 3: Comparing the optimization loss (of Eqn 1) between the visual attack and the text attack counterpart on MiniGPT-4. We limit adversarial texts to 32 tokens, equivalent to the length of image tokens.

Toxicity Ratio	Perspective API (%)		
Target → Surrogate ↓	**MiniGPT-4** (Vicuna)	**InstructBLIP** (Vicuna)	**LLaVA** (LLaMA-2-Chat)
Without Attack	34.8	34.2	9.2
MiniGPT-4	**67.2 (+32.4)**	57.5 (+23.3)	17.9 (+8.7)
InstructBLIP	52.4 (+17.6)	**61.3 (+27.1)**	20.6 (+11.4)
LLaVA	44.8 (+10.0)	46.5 (+12.3)	**52.3 (+43.1)**

Table 3: Transferability of Our attacks. We optimize our adversarial examples on a surrogate model and then use the same adversarial examples to transfer attack another target model. We report percentages (%) of outputs that display at least one of the toxic attributes (i.e., Any in Table 2) under the transfer attacks. These outputs were generated on the 'challenging' subset from RealToxicityPrompts, our scores are evaluated by the Perspective API. Note that we selectively report the strong transfer attack out of (unconstrained, $\epsilon = \frac{16}{255}, \frac{32}{255}, \frac{64}{255}$) for each pair.

remark as there is no theoretical guarantee. Better discrete optimization techniques (developed in the future) may also narrow the gap between visual and text attacks.

Attacks on Other Models and The Transferability

Besides MiniGPT-4 (Vicuna), we also evaluate our attacks on InstructBLIP (Vicuna) and LLaVA (LLaMA-2-Chat). As our study is biased toward cross-modal attacks, we only consider visual attacks in this ablation. Table 3 summarizes our evaluation on the RealToxicityPrompts benchmark. As shown, **white-box** attacks consistently achieve strong effectiveness. *Even though the LLaMA-2-based model is strongly aligned, it is still susceptible to our attacks.* We also validate the **black-box** transferability of our attacks among the three models. When adversarial examples generated on one surrogate model are applied to two other target models, we consistently observe a significant increase in toxicity.

Analyzing Defenses

In general, defending against adversarial examples is known to be fundamentally difficult (Athalye, Carlini, and Wagner 2018; Carlini and Wagner 2017; Tramer 2022) and remains an open problem after a decade of study. As frontier foundation models are becoming increasingly multimodal, we expect they will only be more difficult to safeguard — there is an increasing burden to deploy defenses across all attack surfaces. **In this section, we analyze some existing defenses against our cross-modal attacks.**

Despite advancements in **adversarial training** (Madry et al. 2017) and **robustness certification** (Cohen, Rosenfeld, and Kolter 2019; Li, Xie, and Li 2023) for adversarial defense, their cost is prohibitive for modern models of the LLM scale. Moreover, most of these defenses rely on discrete classes, which is a major barrier when applying these defenses to LLMs with open-ended outputs, contrasting the narrowly defined classification settings. Even more pessimistically, under our threat model that exploits adversarial examples for jailbreaking, the adversarial perturbations are not necessarily imperceptible. Thus, the small perturbation bounds assumed by these defenses no longer apply.

We notice that **input preprocessing based defenses** appear to be more readily applicable in practice. We test the recently developed DiffPure (Nie et al. 2022) to counter our visual adversarial examples. DiffPure mitigates adversarial input by introducing noise to the image and then utilizes a diffusion model (Ho, Jain, and Abbeel 2020) to project the diffused image back to its learned data manifold. Given its model and task independence, DiffPure can function as a plug-and-play module and be seamlessly integrated into our setup. Interestingly, we find that DiffPure can neutralize our visual adversarial examples and prevent jailbreaking (see our Appendix). However, it is unknown whether it is robust against more sophisticated adaptive attacks.

Alternatively, common harmfulness detection APIs like Perspective API and Moderation API may also be applied to **filter out harmful instructions and outputs**. However, they are limited in their accuracy, and their false positives might directly cause bias and harm (Welbl et al. 2021; Xu et al. 2021; OpenAI 2023b). Another trend is post-processing model outputs with another LLM optimized for content moderation (Helbling et al. 2023; Weng, Goel, and Vallone 2023). All of these filtering/post-processing based defenses are only applicable to safeguard online models and can not be enforced for offline models hosted by attackers.

Discussions

Risks of Multimodality. Figure 3 indicates multimodality can open up new attack surfaces on which adversarial examples are easier to optimize. Besides this enhanced "optimization power", these new attack surfaces also carry inherent physical implications. As more modalities are integrated, attackers will gain more physical channels through which attacks can be initiated — some channels, like audio, could be

more stealthy and physically exploitable.

Limitations. LLMs have open-ended outputs, rendering the complete evaluation of their potential harm a persistent challenge (Ganguli et al. 2022a). Our evaluation datasets are unavoidably incomplete. Our work also involves a manual evaluation (Perez et al. 2022), a process that unfortunately lacks a universally recognized standard. Though we also involve an API-based evaluation on RealToxicityPrompts benchmark, it may fall short in accuracy. Thus, our evaluation is only intended as a proof of concept for the adversarial risks we examine in this work.

Future Work: 1) While we focus on vision and language, we conjecture similar cross-modal attacks also exist for other modalities, e.g., audio, lidar, etc. **2)** As the capabilities of the models we study are limited, the harms induced by our attacks are also limited. However, as the model becomes more capable and safety-critical, the risks of the attacks may go beyond mere conceptual. **3)** We preliminarily validate the black-box transferability of our attacks among some open-sourced models. If further enhanced with advanced black-box attack techniques, using open-sourced models to transfer attack proprietary models could be a practical risk.

Conclusion

We underscore the escalating adversarial risks (expansion of attack surfaces and extended implications of security failures) associated with the pursuit of multimodality. As a concrete example, we show the feasibility of exploiting visual adversarial examples to jailbreak aligned LLMs that integrate visual inputs. Through our study, we call for security and safety cautions in developing multimodal systems. More broadly, our finding also uncovers the tension between the long-studied adversarial vulnerabilities of neural networks and the nascent field of AI alignment.

Ethical Statement

This study is dedicated to examining the safety and security risks arising from the vision integration into LLMs. Our research seeks to expose the vulnerabilities in current models, thereby fostering further investigations and mitigation strategies directed toward the evolution of safer and more reliable AI systems. We firmly adhere to principles of respect and dignity for all peoples and unequivocally oppose all forms of actions that would violate these principles. The inclusion of offensive materials, including toxic corpus, harmful prompts, and model outputs, is exclusively for research purposes and does not represent the personal views or beliefs of the authors. All our experiments were conducted in a safe, controlled, and isolated laboratory environment, with stringent procedures in place to prevent any potential real-world ramifications. During our presentation, we redacted most of the toxic content to make the demonstration less offensive. Committed to responsible disclosure, we also discuss potential mitigation techniques in our paper to counter the potential misuse of our attacks.

Acknowledgments

We thank Tong Wu and Chong Xiang for their discussions. Prateek Mittal acknowledges the support by NSF grants CNS-1553437 and CNS-1704105, the ARL's Army Artificial Intelligence Innovation Institute (A2I2), the Office of Naval Research Young Investigator Award, the Army Research Office Young Investigator Prize, Schmidt DataX award, Princeton E-affiliates Award. Mengdi Wang acknowledges the support by NSF grants DMS-1953686, IIS-2107304, CMMI-1653435, ONR grant 1006977, and C3.AI. Xiangyu Qi acknowledges the support of Princeton Gordon Y. S. Wu Fellowship. Peter Henderson acknowledges support from the Open Philanthropy AI Fellowship. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funding agencies.

References

Abdelnabi, S.; Greshake, K.; Mishra, S.; Endres, C.; Holz, T.; and Fritz, M. 2023. Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*, 79–90.

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35: 23716–23736.

Alzantot, M.; Sharma, Y.; Elgohary, A.; Ho, B.-J.; Srivastava, M.; and Chang, K.-W. 2018. Generating natural language adversarial examples. *arXiv preprint arXiv:1804.07998*.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, 2425–2433.

Athalye, A.; Carlini, N.; and Wagner, D. 2018. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*, 274–283. PMLR.

Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; Joseph, N.; Kadavath, S.; Kernion, J.; Conerly, T.; El-Showk, S.; Elhage, N.; Hatfield-Dodds, Z.; Hernandez, D.; Hume, T.; Johnston, S.; Kravec, S.; Lovitt, L.; Nanda, N.; Olsson, C.; Amodei, D.; Brown, T.; Clark, J.; McCandlish, S.; Olah, C.; Mann, B.; and Kaplan, J. 2022. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. arXiv:2204.05862.

Bommasani, R.; Hudson, D. A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M. S.; Bohg, J.; Bosselut, A.; Brunskill, E.; et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Brohan, A.; Brown, N.; Carbajal, J.; Chebotar, Y.; Chen, X.; Choromanski, K.; Ding, T.; Driess, D.; Dubey, A.; Finn, C.;

et al. 2023. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. *arXiv preprint arXiv:2307.15818*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Cao, Y.; Wang, N.; Xiao, C.; Yang, D.; Fang, J.; Yang, R.; Chen, Q. A.; Liu, M.; and Li, B. 2021. Invisible for both camera and lidar: Security of multi-sensor fusion based perception in autonomous driving under physical-world attacks. In *2021 IEEE Symposium on Security and Privacy (SP)*, 176–194. IEEE.

Carlini, N.; and Wagner, D. 2017. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, 3–14.

Carlini, N.; and Wagner, D. 2018. Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE security and privacy workshops (SPW)*, 1–7. IEEE.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; Stoica, I.; and Xing, E. P. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.

Cohen, J.; Rosenfeld, E.; and Kolter, Z. 2019. Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*, 1310–1320. PMLR.

Dai, W.; Li, J.; Li, D.; Tiong, A. M. H.; Zhao, J.; Wang, W.; Li, B.; Fung, P.; and Hoi, S. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. arXiv:2305.06500.

Ebrahimi, J.; Rao, A.; Lowd, D.; and Dou, D. 2017. Hotflip: White-box adversarial examples for text classification. *arXiv preprint arXiv:1712.06751*.

Fang, Y.; Wang, W.; Xie, B.; Sun, Q.; Wu, L.; Wang, X.; Huang, T.; Wang, X.; and Cao, Y. 2023. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19358–19369.

Ganguli, D.; Hernandez, D.; Lovitt, L.; Askell, A.; Bai, Y.; Chen, A.; Conerly, T.; Dassarma, N.; Drain, D.; Elhage, N.; et al. 2022a. Predictability and surprise in large generative models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, 1747–1764.

Ganguli, D.; Lovitt, L.; Kernion, J.; Askell, A.; Bai, Y.; Kadavath, S.; Mann, B.; Perez, E.; Schiefer, N.; Ndousse, K.; et al. 2022b. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*.

Gehman, S.; Gururangan, S.; Sap, M.; Choi, Y.; and Smith, N. A. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*.

Girdhar, R.; El-Nouby, A.; Liu, Z.; Singh, M.; Alwala, K. V.; Joulin, A.; and Misra, I. 2023. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15180–15190.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

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

Helbling, A.; Phute, M.; Hull, M.; and Chau, D. H. 2023. LLM Self Defense: By Self Examination, LLMs Know They Are Being Tricked. arXiv:2308.07308.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33: 6840–6851.

Holtzman, A.; Buys, J.; Du, L.; Forbes, M.; and Choi, Y. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Jones, E.; Dragan, A.; Raghunathan, A.; and Steinhardt, J. 2023. Automatically Auditing Large Language Models via Discrete Optimization. *arXiv preprint arXiv:2303.04381*.

Kenton, Z.; Everitt, T.; Weidinger, L.; Gabriel, I.; Mikulik, V.; and Irving, G. 2021. Alignment of language agents. *arXiv preprint arXiv:2103.14659*.

Lester, B.; Al-Rfou, R.; and Constant, N. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Li, L.; Xie, T.; and Li, B. 2023. SoK: Certified Robustness for Deep Neural Networks. In *44th IEEE Symposium on Security and Privacy, SP 2023, San Francisco, CA, USA, 22-26 May 2023*. IEEE.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Mehrabi, N.; Beirami, A.; Morstatter, F.; and Galstyan, A. 2022. Robust Conversational Agents against Imperceptible Toxicity Triggers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2831–2847. Seattle, United States: Association for Computational Linguistics.

Nie, W.; Guo, B.; Huang, Y.; Xiao, C.; Vahdat, A.; and Anandkumar, A. 2022. Diffusion models for adversarial purification. *arXiv preprint arXiv:2205.07460*.

OpenAI. 2022. Introducing ChatGPT. https://openai.com/blog/chatgpt.

OpenAI. 2023a. Forecasting potential misuses of language models for disinformation campaigns and how to reduce risk. https://openai.com/research/forecasting-misuse. [Online; accessed 4-Apr-2023].

OpenAI. 2023b. GPT-4 Technical Report. arXiv:2303.08774.

OpenAI. 2023c. GPT-4V(ision) system card. https://openai.com/research/gpt-4v-system-card.

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.

Patil, S. G.; Zhang, T.; Wang, X.; and Gonzalez, J. E. 2023. Gorilla: Large language model connected with massive apis. *arXiv preprint arXiv:2305.15334*.

Perez, E.; Huang, S.; Song, F.; Cai, T.; Ring, R.; Aslanides, J.; Glaese, A.; McAleese, N.; and Irving, G. 2022. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*.

Pichai, S. 2023. An important next step on our AI journey. https://blog.google/technology/ai/bard-google-ai-search-updates/.

Pichai, S.; and Hassabis, D. 2023. Introducing Gemini: our largest and most capable AI model. https://blog.google/technology/ai/google-gemini-ai/.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Schick, T.; Udupa, S.; and Schütze, H. 2021. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp. *Transactions of the Association for Computational Linguistics*, 9: 1408–1424.

ShareGPT.com. 2023. ShareGPT: Share your wildest ChatGPT conversations with one click. https://sharegpt.com/.

Shin, T.; Razeghi, Y.; Logan IV, R. L.; Wallace, E.; and Singh, S. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Tramer, F. 2022. Detecting adversarial examples is (nearly) as hard as classifying them. In *International Conference on Machine Learning*, 21692–21702. PMLR.

Wallace, E.; Feng, S.; Kandpal, N.; Gardner, M.; and Singh, S. 2019. Universal adversarial triggers for attacking and analyzing NLP. *arXiv preprint arXiv:1908.07125*.

Wei, A.; Haghtalab, N.; and Steinhardt, J. 2023. Jailbroken: How Does LLM Safety Training Fail? *arXiv preprint arXiv:2307.02483*.

Wei, J.; Bosma, M.; Zhao, V. Y.; Guu, K.; Yu, A. W.; Lester, B.; Du, N.; Dai, A. M.; and Le, Q. V. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35: 24824–24837.

Welbl, J.; Glaese, A.; Uesato, J.; Dathathri, S.; Mellor, J.; Hendricks, L. A.; Anderson, K.; Kohli, P.; Coppin, B.; and Huang, P.-S. 2021. Challenges in detoxifying language models. *arXiv preprint arXiv:2109.07445*.

Weng, L.; Goel, V.; and Vallone, A. 2023. Using GPT-4 for content moderation.

Xu, A.; Pathak, E.; Wallace, E.; Gururangan, S.; Sap, M.; and Klein, D. 2021. Detoxifying language models risks marginalizing minority voices. *arXiv preprint arXiv:2104.06390*.

Zellers, R.; Bisk, Y.; Farhadi, A.; and Choi, Y. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6720–6731.

Zhao, Z.; Dua, D.; and Singh, S. 2017. Generating natural adversarial examples. *arXiv preprint arXiv:1710.11342*.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.