On the Robustness of Large Multimodal Models Against Image Adversarial Attacks

Xuanming Cui[†] xu979022@ucf.edu
Alejandro Aparcedo[†] aaparcedo@ucf.edu
Young Kyun Jang kyun0914@gmail.com
Ser-Nam Lim[†] sernam@ucf.edu

Abstract

Recent advances in instruction tuning have led to the development of State-of-the-Art Large Multimodal Models (LMMs). Given the novelty of these models, the impact of visual adversarial attacks on LMMs has not been thoroughly examined. We conduct a comprehensive study of the robustness of various LMMs against different adversarial attacks, evaluated across tasks including image classification, image captioning, and Visual Question Answer (VQA). We find that in general LMMs are not robust to visual adversarial inputs. However, our findings suggest that context provided to the model via prompts—such as questions in a QA pair—helps to mitigate the effects of visual adversarial inputs. Notably, the LMMs evaluated demonstrated remarkable resilience to such attacks on the ScienceQA task with only an 8.10% drop in performance compared to their visual counterparts which dropped 99.73%. This research highlights a previously under explored facet of LMM robustness and sets the stage for future work aimed at strengthening the resilience of multimodal systems in adversarial environments.

1. Introduction

Large Multi-modal Models (LMMs) have demonstrated remarkable abilities in a range of applications, from image classification and Visual Question Answering (VQA) to image captioning and semantic segmentation [1, 12, 16, 17, 20]. These models excel in generalizing to new domains with data-efficient solution, a feat attributed to advancements in Instruction Tuning [32]. Such techniques, traditionally applied to text-only models, have now been extended to multi-modal models, opening new avenues for efficient fine-tuning with significantly less data [12, 20].

Despite the recent advancements in LMMs, the impact of adversarial examples still remains under explored. Typically adversarial examples are generated end-to-end, targeting the final loss of the whole model, and focusing on a single modality. However, in the era of combining different pre-trained models with additional projectors or adaptors [7, 20, 33], it is imperative to reevaluate the effectiveness of these adversarial approaches. For example, let's consider LLaVA [20] which uses CLIP as its visual component and LLAMA as text component (with some additional projector to bridge the gap), will an attack on one of the two components compromise its overall performance?

We conduct a comprehensive analysis on the robustness of current LMMs under various adversarial attacks, tasks and datasets. Our investigation reveals that LMMs are not robust to adversarial visual perturbations in contexts where no additional textual information is provided, such as in COCO[18] classification (without context) or COCO captioning tasks. Conversely, the presence of context seems to bolster LMM robustness, as seen in tasks like COCO classification (with context). In cases where the attack does not directly target the core aspects of the task, such as in VQA, LMMs display a degree of inherent robustness. This paper reveals the following findings:

- LMMs are generally vulnerable to adversarial visual perturbations, even if such perturbations are generated only w.r.t. the visual model.
- Compared to classification and caption, LMMs demonstrate better robustness in VQA tasks. Particularly, we find that visual attacks are less effective when the VQA question query involves different visual contents from what is being attacked.
- Adding additional textual context notably improves LMMs' robustness against visual adversarial input.

2. Related Work

Large Multimodal Models (LMMs). Large Multimodal Models (LMMs)[3, 7, 17, 20, 33] typically comprise a visual model, a pre-trained Large Language Model (LLM), and a projector model designed to bridge the modality gap between images and text. Prominent among these models are LLaVA[20] and InstructBLIP [12], which represent the current state-of-the-art in LMMs. LLaVA integrates the CLIP visual encoder with the Vicuna LLM [9], employing a simple linear projector subsequent to the visual model for transforming visual representations into the lan-

[†]University of Central Florida

guage embedding space. Conversely, BLIP2-based models [12, 17, 33] utilize the EVA-CLIP visual encoder, alongside a Q-former equipped with learnable query vectors to bridge the visual and textual modalities. Both LLaVA and BLIP2-based models, among others, have demonstrated remarkable capabilities in a variety of vision-language tasks, underscoring their versatility and effectiveness.

Adversarial attacks. Adversarial attacks are designed to subtly manipulate inputs in a way that is typically imperceptible to humans, yet can lead neural networks to produce erroneous outputs [2, 4, 5, 11, 24, 30]. These attacks are broadly classified into two categories: white-box attacks [2, 5, 14, 30], where the adversary has complete access to the model parameters, and black-box attacks [26, 29], where the adversary possesses limited information such as output logits or labels. In particular, transfer-based attacks leverage gradients from a surrogate model under white-box condition, which are likely transferable to the target black-box model [13, 21, 25, 26]. Such transferability thus remain as an critical model vulnerability. **LMMs and Adversarial Examples.** While extensive research has been conducted on adversarial attacks in both visual and textual domains, the impact of these attacks on current LMMs remains relatively unexplored. Recent studies [6, 22, 27, 28, 31, 34] demonstrate the feasibility of creating adversarial examples that effectively "jailbreak" LMMs from both visual [6, 27] and textual [22, 28, 31, 34] inputs, using either gradient-based approaches [6, 27] or prompt engineering [22, 28, 31].

3. Method

3.1. Threat Model

In this study, we focus on gradient-based white-box adversarial attacks [5, 11, 24]. These methods hinge on the computation of the gradient to ascertain the most effective direction in which to modify the input so as to deceive the model.

3.2. Attacks

We choose Projected Gradient Descent (PGD) and Carlini-Wagner (CW) as two representatives of strong gradient-based attacks, along with Auto-PGD (APGD) as a variant of PGD. Additionally, we experiment with two parameter settings of each attack: normal and strong, based on perceptibility of the perturbations.

3.3. Models

In our study, we selected three state-of-the-art LMM models for evaluation: LLaVA1.5[19] integrated with the Vicuna13B language model, BLIP2 combined with the Flan T5 XXL[10] language model, and InstructBLIP [12], also utilizing Vicuna13b.

3.4. Tasks & Adversarial generation

We consider three popular visual tasks for evaluating visual adversarial impact on LMMs: image classification, caption retrieval and VQA. Since we are interested in LMMs' robustness against visual adversaries, we generate adversarial samples w.r.t. the image encoder of the LMM: CLIP image encoder for LLaVA, EVA-CLIP image encoder for BLIP2 and InstructBLIP. We use CLIP text encoder and the text encoder from BLIP's Q-former to compute the text embeddings for their corresponding image encoder.

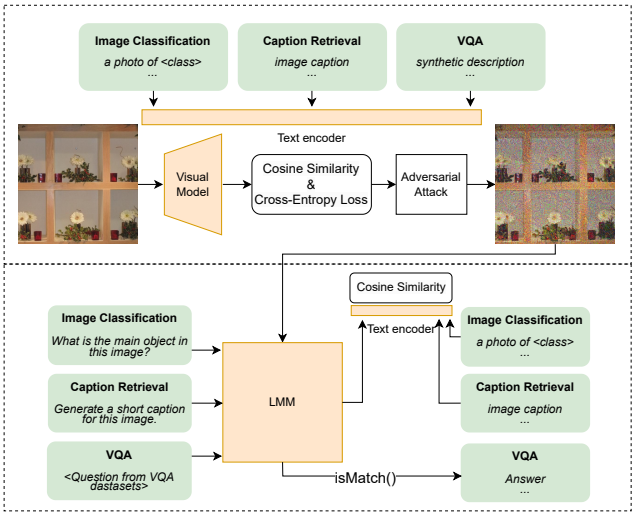

Figure 1. Overview of our procedure for attack generation and evaluation over image classification, caption retrieval, and VGA. Top: overview of attack generation for the three tasks; bottom: evaluation procedure for LMM on the three tasks.

3.4.1 Image Classification

We use COCO [18] 2014 validation split (2014val), with class annotations from [15], to evaluate robustness on classification. We first use the text encoder to encode the text class labels in the format of "a photo of <class>". Then, we compute the class-wise cosine-similarity between the image encodings and encoded class labels and use the result as the class logits for adversarial generation and evaluation. To evaluate LMMs on classification, we first prompt LMMs to generate a one-word response of the main object in the image.

3.4.2 Caption retrieval

We use COCO captioning dataset [8] 2014val for evaluating caption retrieval robustness. To generate visual adversarial samples for caption retrieval, we first use the text encoder to encode 5 captions per image, and then use their mean as the text encodings for each image. Then, we compute cosine

Model	Attack	Pre	Post_N	Post_S
Image-to-Text Recall @1 (%)				
CLIP	PGD	57.72	10.4(-82)	0.4(-99)
CLIP	APGD	57.72	12.92(-78)	7.44(-87)
CLIP	CW	57.72	34.94(-39)	24.94(-57)
LLM Answer-to-Text Recall @1 (%)				
LLaVA	PGD	36.58	13.1(-64)	3.76(-90)
LLaVA	APGD	36.58	15.7(-57)	7.88(-78)
LLaVA	CW	36.58	32.96(-10)	29.84(-18)

Table 1. Top-1 caption retrieval result for COCO caption 2014 validation dataset. Refer to Sec. 3.4.2. "Visual Encoder Accuracy" refers to CLIP accuracy on successfully retrieving captions that are closed to the mean caption encoding given the image encoding. "Image-to-Text Recall @1" is recall@1 of retrieving correctly one of the five captions for the given image. LLM Answer-to-Text recall is the same except the query is the LMMs' answers. Numbers in parenthesis show % change w.r.t. the Pre-attack accuracy.

similarity between image and text encodings and use the result as the image-wise logits for adversarial generation.

3.4.3 VQA

We evaluate LMM robustness on the ScienceQA, which contains 21k multimodal multiple-choice questions [23].

4. Experimental Results and Analysis

We show our experimental results and analysis in the following sections. We report both LMMs' accuracy as well as the image encoder's accuracy on the task that was used to generate adversaries. We adopt the notations Pre, Post_N and Post_S to refer to accuracy for pre-attack, post-attack under normal setting, and post-attack under strong setting, respectively.

4.1. Are LMMs Robust Against Adversarial Visual Input?

To investigate the impact of adversarial visual inputs on LMMs, our initial analysis focuses on the caption retrieval task. This task serves as a measure of the LMMs' overall comprehension of visual inputs. The results of this analysis, conducted on COCO 2014val, are presented in Table 1. Under the third section, the data distinctly illustrates a significant decrease in post-attack accuracy across all three LMMs when subjected to both PGD and APGD attacks, under both normal and strong settings.

4.2. Evaluating LMMs' VQA Performance

In this section, we detail the experimental outcomes of the LMMs in VQA tasks under adversarial visual attacks. The primary results are summarized in Table 2. Our results indicate a noteworthy deviation from what we have observed

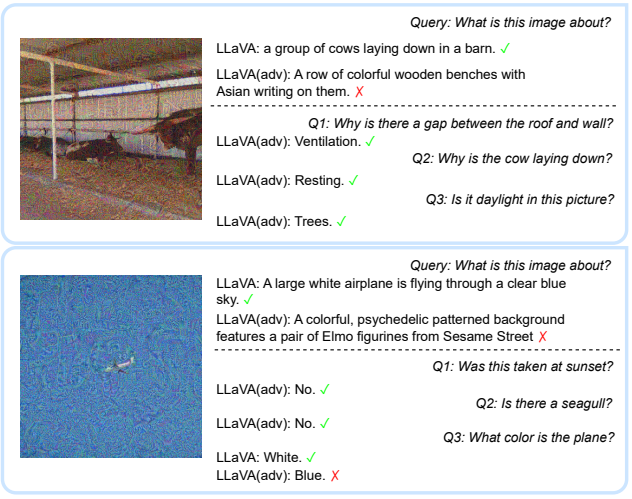

Figure 2. Two sample adversarial images from COCO 2014val, generated under APGD Post_S. "LLaVA" and "LLaVA(adv)" refer to LLaVA's responses using the clean Pre-attack and post-attack image, respectively. Above the dotted line in each cell, we query LLaVA for the general description; below the dotted line are questions taken from VQA V2 dataset.

about the caption retrieval task in Sec. 4.1, which did not show that LMMs possess any robustness against visual adversaries. Based on the results from Table 2, all three LMMs being evaluated exhibit considerable resilience in various VQA datasets, despite the significant decrease in adversarial accuracy of their corresponding visual encoders, as shown under the "Visual Encoder Accuracy" columns. For instance, with the ScienceQA dataset, the Post_N "Visual Encoder Accuracy" plummeted below 1% for all three types of attacks, and for both the CLIP and BLIP visual encoders. However, the accuracy of all three LMMs decreased by less than 7% compared to their pre-attack accuracy.

What could be the cause of such discrepancies in LMMs' robustness between the VQA and caption retrieval tasks? We make two conjectures:

1. The robustness of LMMs depends on whether the query is about what is being attacked. Since the attack target for generating visual adversarial samples is what is being described in the image description, then intuitively those aspects not mentioned in the description shall be less affected by the attack.
2. Additional contexts (e.g., contexts in ScienceQA's questions) aid in LMMs' robustness.

We will experimentally support the two claims in the following sections.

4.3. Visual Adversarial Attacks are not Universal to LMMs

In this section, we present an empirical analysis demonstrating that while LMMs are not inherently resilient to visual

Model	Dataset	Attack	VQA Acc (%)			Visual Encoder Acc (%)		
			Pre	$Post_N$	$Post_S$	Pre	$Post_N$	$Post_S$
LLaVA	ScienceQA(image)	PGD	71.59	68.77 (-3)	64.75 (-9)	42.92	10.08 (-77)	0.92 (-98)
LLaVA	ScienceQA(image)	APGD	71.59	69.81 (-2)	68.22 (-4)	42.83	5.68 (-87)	0.06 (-99)
LLaVA	ScienceQA(image)	CW	71.59	71.69 (+0.1)	71.34 (-0.3)	42.95	12.76 (-70)	0.03 (-99)

Table 2. Results on VQA datasets. We attack CLIP visual encoder to generate adversarial examples for LLaVA . Adversarial examples are used as input image along with question as input text. "VQA Accuracy" refers to the performance of each LMM; "Visual Encoder Accuracy" refers to the accuracy of the visual encoder on image-to-text retrieval, which is used for generating visual adversaries for VQA. Numbers in parenthesis show % change w.r.t. the Pre-attack accuracy.

adversarial attacks, as evidenced by their performance in caption retrieval tasks, they are capable of delivering correct responses when the query's focus differs from the target of the attack. To illustrate this, we take the Visual Question Answering (VQA) V2 dataset as a case study. Here, we generate adversarial images using the text label "a photo of <class>", with the attack primarily aimed at the central object of the image. We observe that the adversarial attack's effectiveness is heightened when the query, during evaluation, pertains to the same target – the principal object in the image. Conversely, the attack's impact diminishes when the query relates to different aspects of the image. In Figure 2, we show LLaVA's responses to queries on two adversarial images under APGD-S. When querying about the general description of the image, it is clear that LLaVA's post-attack answers are completely deviated from what the image is about.

4.4. Adding Context Improves LMM Robustness

To examine the effect of context on LMMs' robustness, we reuse the image classification task. We first ask LLaVA to generate a general one-sentence description for each class. We then insert the generated description corresponding to the correct object into the prompt for querying the LMMs about the main object in the image. Besides the additional context, everything else is kept the same. Results are shown in Table 3. We observe that after adding a short sentence of context, the post-attack accuracy for all three LMM models increase by a large margin. In particular, the accuracy drop for BLIP2/InstructBLIP under PGD/APGD reduce to only 20%, as opposed to an average of 60% drop without context.

Model	Attack	Pre@1	$Post_N$@1	$Post_S$@1
LMM Acc (%)				
LLaVA	PGD	87.51	48.25(-45)	22.58(-74)
LLaVA	APGD	87.51	52.06(-41)	8.11(-91)
LLaVA	CW	87.51	80.64(-8)	77.1(-12)
BLIP2-T5	PGD	86.47	28.64(-67)	2.98(-97)
BLIP2-T5	APGD	86.47	31.39(-67)	2.37(-97)
BLIP2-T5	CW	86.47	70.11(-19)	58.85(-32)
InstructBLIP	PGD	89.89	21.09(-77)	3.66(-96)
InstructBLIP	APGD	89.89	22.35(-75)	2.18(-98)
InstructBLIP	CW	89.89	37.81(-58)	31.91(-64)
LMM with Context Acc (%)				
LLaVA	PGD	93.74	73.62(-21)	57.06(-39)
LLaVA	APGD	93.74	72.61(-23)	37.65(-60)
LLaVA	CW	93.74	91.76(-2)	90.2(-4)
BLIP2-T5	PGD	97.67	87.54(-10)	94.92(-3)
BLIP2-T5	APGD	97.67	87.29(-11)	98.43(+1)
BLIP2-T5	CW	97.67	94.97(-3)	92.76(-5)
InstructBLIP	PGD	88.94	66.92(-25)	71.61(-19)
InstructBLIP	APGD	88.94	68.74(-23)	89.22(-0)
InstructBLIP	CW	88.94	84.92(-5)	82.51(-7)

Table 3. Top-1 image classification result on COCO 2014val. The first table section shows visual encoder accuracy, referring to CLIP/EVA-CLIP's accuracy on classification; second section shows LMMs' accuracy; third section show LMMs' accuracy, after the context is added to the query. Numbers in parenthesis show % change w.r.t. the Pre-attack accuracy.

5. Conclusion

In this study, we systematically evaluate the susceptibility of LMMs to visual adversarial inputs across a diverse array of tasks and datasets. Our findings suggests LMMs are highly vulnerable to visual adversarial attacks, even when such adversaries are crafted with respect to the visual model alone. On the other hand, we find that LMMs are "robust" when the query and attack target does not match. Such characteristics indicates the traditional task-specific adversarial generation techniques are not universally effective against current LMM, and points to the need for further research into new adversarial attack strategies, particularly in the context of zero-shot inference. Finally, we find adding context about the querying object improves LMMs' visual robustness. We therefore propose a strategy to decompose questions into multiple existence questions associated with the corresponding context, which achieved notable improvements in robustness on COCO and Imagenet classification.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine 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. In *Advances in Neural Information Processing Systems*, 2022. 1

[2] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, 2018. 2

[3] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 1

[4] 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*, page 387–402. Springer Berlin Heidelberg, 2013. 2

[5] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57, 2017. 2

[6] 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?, 2023. 2

[7] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic, 2023. 1

[8] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server, 2015. 2

[9] 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, 2023. 1

[10] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022. 2

[11] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *Proceedings of the 37th International Conference on Machine Learning*, pages 2206–2216. PMLR, 2020. 2

[12] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 1, 2

[13] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li. Boosting adversarial attacks with momentum. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9185–9193, Los Alamitos, CA, USA, 2018. IEEE Computer Society. 2

[14] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015. 2

[15] Young Kyun Jang, Geonmo Gu, Byungsoo Ko, Isaac Kang, and Nam Ik Cho. Deep hash distillation for image retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 2

[16] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*, 2023. 1

[17] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 1, 2

[18] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1, 2

[19] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 2

[20] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 1

[21] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *International Conference on Learning Representations*, 2017. 2

[22] Yi Liu, Gelei Deng, Zhengzi Xu, Yuekang Li, Yaowen Zheng, Ying Zhang, Lida Zhao, Tianwei Zhang, and Yang Liu. Jailbreaking chatgpt via prompt engineering: An empirical study, 2023. 2

[23] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022. 3

[24] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. 2

[25] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples, 2016. 2

[26] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. page 506–519, New York, NY, USA, 2017. Association for Computing Machinery. 2

[27] Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Peter Henderson, Mengdi Wang, and Prateek Mittal. Visual adversarial examples jailbreak aligned large language models, 2023. 2

[28] Abhinav Rao, Sachin Vashistha, Atharva Naik, Somak Aditya, and Monojit Choudhury. Tricking llms into disobedience: Understanding, analyzing, and preventing jailbreaks, 2023. 2

[29] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23:828–841, 2017. 2

[30] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014. 2

[31] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail?, 2023. 2

[32] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2022. 1

[33] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models, 2023. 1, 2

[34] Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023. 2