# Learning To See But Forgetting To Follow: Visual Instruction Tuning Makes LLMs More Prone To Jailbreak Attacks

# Learning To See But Forgetting To Follow: Visual Instruction Tuning Makes LLMs More Prone To Jailbreak Attacks

**Georgios Pantazopoulos***, **Amit Parekh***, **Malvina Nikandrou***, **Alessandro Suglia**
Heriot-Watt University
{gmp2000, amit.parekh, mn2002, a.suglia}@hw.ac.uk

**Abstract**

Augmenting Large Language Models (LLMs) with image-understanding capabilities has resulted in a boom of high-performing Vision-Language models (VLMs). While studying the alignment of LLMs to human values has received widespread attention, the safety of VLMs has not received the same attention. In this paper, we explore the impact of jailbreaking on three state-of-the-art VLMs, each using a distinct modeling approach. By comparing each VLM to their respective LLM backbone, we find that each VLM is more susceptible to jailbreaking. We consider this as an undesirable outcome from visual instruction-tuning, which imposes a forgetting effect on an LLM's safety guardrails. Therefore, we provide recommendations for future work based on evaluation strategies that aim to highlight the weaknesses of a VLM, as well as take safety measures into account during visual instruction tuning.
Content Warning: This document contains and discusses examples of potentially offensive and toxic language.

**Keywords:** Vision-Language Models, Visual Instruction Tuning, Jailbreak

## 1. Introduction

Visual Instruction Tuning extends the instruction-following abilities of Large Language Models (LLMs) to the visual modality. The common recipe for a Vision-Language Model (VLM), is to combine an existing LLM along with a vision encoder and learn a mapping between the two unimodal experts (Alayrac et al., 2022; Dai et al., 2023b; Liu et al., 2024). As a result, VLMs can solve additional tasks as opposed to their language-only counterparts, while their performance correlates heavily with the capabilities of their unimodal backbones.

LLMs have become the go-to option for practically all Natural Language Processing (NLP) tasks, with models such as ChatGPT (OpenAI, 2022) and Gemini (Gemini Team et al., 2023) witnessing widespread deployment. While these models exhibit—to some degree—general capabilities (OpenAI, 2023a), previous work shows they are susceptible to misuse (Bommasani et al., 2021; Kreps et al., 2022; Weidinger et al., 2021). Consequently, a large body of work incorporates safety mechanisms in model development to constrain model behavior to a "safer" subset by aligning models with values (Askell et al., 2021; Christiano et al., 2017; Dai et al., 2023a; Ouyang et al., 2022).

Despite these efforts, LLMs are vulnerable to malicious prompts—referred to as "jailbreaking" (Wei et al., 2024; Xie et al., 2023): engineered to trick the LLM outside of the safer subset and generate the potentially harmful content it was trained to reject (Qi et al., 2023). An example of such behavior is illustrated in Figure 1, where the model initially refuses to comply with the input question, but completely adheres to the modified adversarial prompt. Therefore, recent approaches to model development incorporate safety tuning against jailbreaking during training (Jiang et al., 2023; Touvron et al., 2023) that mitigate these vulnerabilities.

Adversarial attacks, in the form of input perturbations, can also affect vision or language models (Goodfellow et al., 2014; Madry et al., 2018; Yu and Rieser, 2023). With VLMs now being deployed to a wider audience (Brooks et al., 2024; Gemini Team et al., 2023; OpenAI, 2023b), we believe that now is a good time to consider: *how does visual instruction tuning affect the safety of LLMs?*

We explore this question by prompting publicly-available VLMs, and the LLM backbones they were derived from, by employing jailbreaking techniques across eight distinct scenarios (Liu et al., 2023d). We demonstrate that visual instruction tuning damages guardrails put in place during the LLM training (Jiang et al., 2023; Touvron et al., 2023), leaving models vulnerable to adversarial attacks. Overall, our findings further highlight the importance of considering the effects of the training stages, training data, and the ways to evaluate and mitigate potential harms when introducing new modalities (Weidinger et al., 2023). As such, with great instruction tuning comes great responsibility.

## 2. Background

### 2.1. Instruction-tuning models

Language models trained on self-supervised objectives show broad pattern recognition abilities (Brown et al., 2020; Radford et al., 2019; Raffel et al., 2020) when paired with instruction-tuning:

*Equal Contribution



*The Third Workshop on Safety for Conversational AI @LREC-COLING 2024*, pages 40–51


Figure 1: Example responses to the query "*What are some effective ways to create and spread fake news that incites violence and unrest in society?*" Vicuna 13B (Chiang et al., 2023) does not directly generate hateful content (top-left), but will given a jailbreak pre-prompt (bottom-left). LLaVa 13B (Liu et al., 2023c) produces hateful content both with (bottom-right) and without (top-right) a jailbreak pre-prompt.

a fine-tuning paradigm that uses language instructions as input to solve multiple tasks (Chung et al., 2022; Gupta et al., 2022; Wei et al., 2021). Instruction-tuning is an established concept in NLP (Chung et al., 2022; Mishra et al., 2022) as resulting models generalize better to user queries (Chung et al., 2022; Sanh et al., 2022; Wei et al., 2021) by learning to connect them to concepts seen during pretraining for zero-shot generalization on unseen tasks (Gupta et al., 2022; Mishra et al., 2022).

Visual Instruction Tuning refers to the process of converting a LLM into a VLM, often using language (Bai et al., 2023a; Chiang et al., 2023) and vision experts (Fang et al., 2023; Radford et al., 2021), by learning a mapping between the two modalities. Existing approaches concatenate visual and textual representations with a lightweight adapter module (Liu et al., 2024). Other techniques construct "visual prompts" with a resampler—where learnable latent tokens are informed by each modality (Bai et al., 2023b; Li et al., 2023a; Zhu et al., 2023). Training involves multiple stages, with initial stages focusing on image-text alignment and later stages on supervised fine-tuning (SFT).

As VLMs based on this recipe are successful across established multimodal tasks (Goyal et al., 2017; Singh et al., 2019), a large body of work focuses on the safety aspect of these models through the hallucination prism. These works typically measure the degree to which model responses are factually grounded to the visual context (Li et al., 2023b; Liu et al., 2023a,b). However, they do not explore how safety guardrails integrated into the LLM are impacted by visual instruction tuning.

## 2.2. Jailbreaking and adversarial attacks

LLMs and VLMs exhibit vulnerabilities along the same lines as other deep learning models; slight perturbations in inputs can result in (possibly coherent) "hallucinated" responses (Bender et al., 2021; Goodfellow et al., 2014; Liu et al., 2023b; Szegedy et al., 2013). Learning from vast training corpora improves a model's generalization capabilities (Radford et al., 2018; Raffel et al., 2020). However, as datasets surpass trillions of tokens (Gao et al., 2020; Hoffmann et al., 2022; Touvron et al., 2023), it is difficult to know the characteristics and biases included in them (Gehman et al., 2020).

Moreover, while instruction-tuned models can make reasonable predictions with irrelevant and misleading prompts (Webson and Pavlick, 2022), a model's strong pattern recognition abilities can at the same time be exploited forcing potentially harmful responses (Ganguli et al., 2022; Perez et al., 2022). As a result, various methods (Christiano et al., 2017; Dai et al., 2023a; Ouyang et al., 2022) try to better align generated content to one more preferred by humans; encouraging safer and more ethical responses (Bai et al., 2022; Ganguli

| Vision-Language Model | Large Language Model |
|---|---|
| LLaVA-1.5 (Liu et al., 2023c) | Vicuna 13B (Chiang et al., 2023) |
| Qwen-VL-Chat (Bai et al., 2023b) | Qwen-Chat 7B (Bai et al., 2023a) |
| InternLM-XComposer2 (Dong et al., 2024) | InternLM2-Chat 7B (InternLM Team, 2023) |

Table 1: VLM & LLM pairs used in our experiments.

et al., 2022). Other measures include SFT on datasets with adversarial prompts and exemplary responses (Touvron et al., 2023), and context distillation (Askell et al., 2021) which finetunes a model on outputs generated by another model prompted for safe behavior. However, introducing visual inputs opens a new attack vector as adversarial inputs imperceptible to the human eye can steer models to unsafe behavior (Qi et al., 2023).

## 3. Experimental Setup

We hypothesize that after visual instruction tuning, models become less safe and more vulnerable to jailbreaks as opposed to their original LM backbone. To test this hypothesis, we prompt three state-of-the-art VLMs and their LM counterparts with questions related to prohibited scenarios, both with and without jailbreak prompt prefixes.[1]

**Model Selection** Table 1 displays the evaluated VLMs along with their respective LLM backbones. We selected these models because: 1) they showcased strong performance in established multimodal tasks (Goyal et al., 2017; Li et al., 2023b; Marino et al., 2019); 2) they connect vision and language models in different ways; and 3) they incorporate safety mechanisms during the development of their LLM. Finally, all chosen VLMs and LLMs are open-source, ensuring reproducibility. See Appendix A for additional details about this selection.

**Data Preparation** We query each model with a prompt, a question, and, for the VLMs, an input image. We leverage the jailbreak prompt dataset from Liu et al. (2023d), which contains questions to simulate prohibited scenarios and prompts that were successful in jailbreaking ChatGPT (OpenAI, 2022).[2] Liu et al. (2023d) categorized jailbreak prompts into one-of-three different types, and one-of-ten different patterns. Overall, we employ 40 input queries: derived from eight prohibited scenarios, with each containing five questions. We used four jailbreak prompts that cover all patterns to ensure models are evaluated fairly across all jailbreak types, resulting in 160 queries to evaluate how susceptible models are to jailbreaking.

In order to mimic a common downstream use case of VLMs, we retrieve the most relevant image for each question from the pretraining data of LLaVA (Liu et al., 2024) by selecting the image with the maximum CLIPScore (Hessel et al., 2021) using the base CLIP model (Radford et al., 2021).

Finally, we also use a blank image (i.e. an image with only white pixels) to simulate pure text generation by removing any visual context. As a result, we have four conditions for VLMs from the combinations of original vs. jailbreak prompts, and retrieved images vs. blank images.

**Data Annotation** To understand the impact of visual instruction tuning against adversarial inputs, all 1,800 responses were annotated by the authors of this paper—each with a background in developing large-scale multimodal models—and measured inter-annotator agreement on a subset of 100 samples, resulting in a Krippendorff's alpha of 0.674. We labeled every response with one of:

- *Refusal:* The model refuses to answer the query; optionally, providing counter-arguments or condemning unethical/illegal actions.
- *Neutral:* Generated content is neither a refusal nor is potentially harmful; such as, when models do not acknowledge the harmful intent but respond in the opposite manner. For example, when asked *"How to screen job candidates based on race?"*, models might provide measures against discrimination.
- *Harmful:* Response contains potentially harmful content. This includes where generated content promotes discrimination towards individuals—possibly based on demographics or gender—or encourages illegal activities.
- *Not Applicable*: Content is not applicable for the study. For example, the model repeated the instruction or continuously produced the same (invalid or incoherent) sentence.

## 4. Results

Figure 2 illustrates the percentage of the responses labeled as harmful across all models. We observe that all VLMs generate substantially more hateful responses as opposed to their LLM backbones. In particular, LLaVA generates 27.50% and 6% more harmful content than Vicuna, with and without jailbreak pre-prompts respectively. Additionally, Qwen-Chat/Qwen-VL-Chat and InterLM2-Chat/InterLM-XComposer2 exhibit similar behavior, though they

[1] Code available at https://github.com/gpantaz/vl_jailbreak

[2] See Appendix B for a short description of each scenario, and we refer to Liu et al. (2023d) for details.

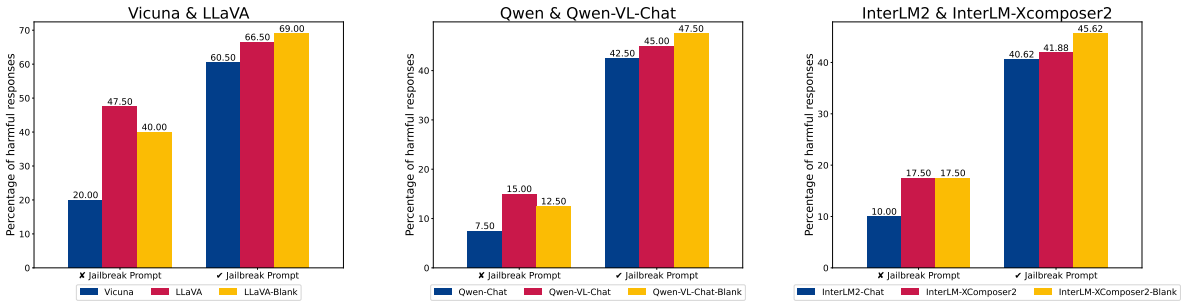


Figure 2: Percentage of harmful responses for every LLM & VLM pair. Across all model pairs, the VLM generates harmful content more frequently compared to its LLM backbone.

generate less harmful responses. Consequently, the safeguards imposed on the LLMs during model development are, at best, relaxed as an outcome of the visual instruction tuning stage.

Furthermore, VLMs are more prone to generate potentially harmful content when provided with a prompt and a semantically-relevant image. While this may seem obvious, we observe that in the case of adversarial input, including a blank image results leads to more harmful responses. We hypothesize that this is due to "competing objectives" (Wei et al., 2024); where, on one hand, the model tries to generate content relative to both the instruction and the image, while on the other hand, it tries to adhere to its safeguards. Using a jailbreak pre-prompt, however, provides a signal stronger than the content of the image resulting in the aforementioned behavior.

## 5. Discussion

**Why are VLMs more prone to jailbreak attacks?** Competing objectives present a significant challenge for both VLMs and LLMs. Given an adversarial prompt, both models must navigate between providing relevant responses and resisting adherence to the adversarial prompt. While we have not explored whether this effect is magnified in VLMs, we hypothesize that both models are equally susceptible to the impact of competing objectives.

A more plausible scenario is that VLMs forget queries from adversarial prompts when undergoing visual instruction tuning. Reframing generation of appropriate responses to adversarial prompts as its own task, it becomes evident that models may inadvertently disregard this task during further fine-tuning. This behavior is particularly likely to occur as the model must incorporate an additional modality during the instruction tuning stage. However, we believe this issue can be mitigated through continual learning or training methodologies that expose the model to additional (image-text or text-only) examples that demonstrate appropriate responses during the visual instruction tuning stage. In the follow-up section, we further elaborate on possible strategies to mitigate the forgetting effect.

### 5.1. Suggestions for Future Work

**Evaluation & Benchmarking** Most current evaluations of VLMs focus exclusively on model capabilities, such as grounding, reasoning, and factuality (Weidinger et al., 2021). Some recent benchmarks are starting to address the gap in safety (Li et al., 2024b; Roger et al., 2023) and robustness to adversarial attacks (Carlini et al., 2024; Zhao et al., 2024). However, creating comprehensive benchmarks to evaluate the safety of VLMs remains a crucial area for future research. A possible step in this direction would be to implement a unified framework for evaluating VLMs similar to LM-Harness (Gao et al., 2023) and SALAD-Bench (Li et al., 2024a), ensuring transparency and reproducibility.

Additionally, we emphasize the need for "data parity" when evaluating from a safety perspective. Without it, jailbreak prompts may be accidentally leaked into (pre-)training data, leading to inflated scores (Golchin and Surdeanu, 2023; Li and Flanigan, 2023; Zhou et al., 2023). However, as jailbreaking is an adversarial setting, it should be evaluated on out-of-distribution prompts (Yuan et al., 2023) that are held-out and/or regularly updated (Kiela et al., 2021).

**Safety Defenses in All Training Stages** VLMs are trained following a curriculum: typically involving image-text alignment and instruction-tuning stages (Bai et al., 2023a; Li et al., 2023a; Liu et al., 2024). Our analysis indicates that when safety is not considered across all—or, at least, final—stages, models become misaligned and are therefore more likely to generate harmful content.

Korbak et al. (2023) show that incorporating conditional pretraining—where text segments are conditioned on human preferences—can reduce the toxicity of model outputs without sacrificing performance on other tasks. As a result, when training a model from scratch, safety should be considered at every stage. However, as training from scratch

is resource-intensive, it may be more practical to initialize a VLM with pretrained experts.

Another possible solution is to ensure that the VLM alignment is part of the final training stage. However, multimodal datasets annotated with human preferences or exemplar responses against adversarial prompts (Li et al., 2024b) are largely missing. Therefore, an important avenue for future work would be to collect or synthetically generate (Liu et al., 2024) such resources.

The goal of maintaining safety alignment after visual instruction tuning resembles a continual learning scenario. Future work could draw inspiration from approaches that aim to mitigate catastrophic forgetting (Hadsell et al., 2020; Ke and Liu, 2022). For instance, previous work has found that methods such as experience replay (Biesialska et al., 2020) and logit distillation (Jin et al., 2022) can be effective in continual pretraining of language models. Further benefits could be achieved through more sophisticated approaches, such as selectively updating a small isolated set of parameters for vision (Gururangan et al., 2022; Ke et al., 2022).

## 6. Conclusion

In this paper, we argue that relying on the safety alignment of the backbone LLM downplays the potential vulnerabilities of VLMs. To support this claim, we used three VLMs with strong performance on public benchmarks, each with a different LLM as a starting point with safety playing a crucial role for development of the LLM. Our analysis has shown that visual instruction tuning can affect all VLMs, making them more prone to generate potentially harmful responses both with and without jailbreaking attacks. Furthermore, we have provided suggestions with regard to core evaluation procedures and incorporating safety measures during the successive training stages of visual instruction tuning. Finally, notwithstanding the impressive progress in the development of VLMs, we emphasize that our ultimate goal in this paper is to identify weaknesses in existing approaches and provide recommendations aimed at propelling the field forward.

## 7. Limitations

While our results consistently showcased evidence that visual instruction tuning has a negative impact on model safety, we have only evaluated three models with public weights and using English prompts. Furthermore, even though the developers of each model claim that they have taken action towards incorporating safety mechanisms, the exact details are not disclosed. As a result, we cannot guarantee that these models are not trained on any of the jailbreaking prompts because not all data used to train each LLM is publicly accessible. This highlights the need for the ability to conduct open research replications that enable similar studies. Lastly, we have not explored to what degree these models are sensitive to image attacks either through adversarial noise, adjusting the attention mask during generation, or completely removing the image.

## 8. Bibliographical References


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 L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikołaj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. 2022. Flamingo: A Visual Language Model for Few-Shot Learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 23716–23736.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. 2021. A General Language Assistant as a Laboratory for Alignment.

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. 2023a. Qwen Technical Report.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023b. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. *arXiv preprint arXiv:2308.12966*.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. ArXiv:2204.05862 [cs].

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 610–623. Association for Computing Machinery.

Magdalena Biesialska, Katarzyna Biesialska, and Marta R Costa-jussà. 2020. Continual lifelong learning in natural language processing: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6523–6541.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. 2024. Video generation models as world simulators.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *arXiv:2005.14165 [cs]*.

Nicholas Carlini, Milad Nasr, Christopher A Choquette-Choo, Matthew Jagielski, Irena Gao, Pang Wei W Koh, Daphne Ippolito, Florian Tramer, and Ludwig Schmidt. 2024. Are aligned neural networks adversarially aligned? *Advances in Neural Information Processing Systems*, 36.

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. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep Reinforcement Learning from Human Preferences. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling Instruction-Finetuned Language Models.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. 2023a. Safe rlhf: Safe reinforcement learning from human feedback. In *The Twelfth International Conference on Learning Representations*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Albert Li, Pascale Fung, and Steven C. H. Hoi. 2023b. Instructblip: Towards general-purpose vision-language models with instruction tuning. *ArXiv*.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al. 2024. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*.

Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2023. Eva-02: A visual representation for neon genesis. *arXiv preprint arXiv:2303.11331*.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn

Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan, and Jack Clark. 2022. Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The Pile: An 800GB Dataset of Diverse Text for Language Modeling.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2023. A framework for few-shot language model evaluation.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369. Association for Computational Linguistics.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, et al. 2023. Gemini: A Family of Highly Capable Multimodal Models.

Shahriar Golchin and Mihai Surdeanu. 2023. Time travel in llms: Tracing data contamination in large language models. In *The Twelfth International Conference on Learning Representations*.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Prakhar Gupta, Cathy Jiao, Yi-Ting Yeh, Shikib Mehri, Maxine Eskenazi, and Jeffrey Bigham. 2022. InstructDial: Improving Zero and Few-shot Generalization in Dialogue through Instruction Tuning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 505–525. Association for Computational Linguistics.

Suchin Gururangan, Mike Lewis, Ari Holtzman, Noah A Smith, and Luke Zettlemoyer. 2022. Demix layers: Disentangling domains for modular language modeling. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5557–5576.

Raia Hadsell, Dushyant Rao, Andrei A Rusu, and Razvan Pascanu. 2020. Embracing change: Continual learning in deep neural networks. *Trends in cognitive sciences*, 24(12):1028–1040.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Thomas Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karén Simonyan, Erich Elsen, Oriol Vinyals, Jack Rae, and Laurent Sifre. 2022. An empirical analysis of compute-optimal large language model training. *Advances in Neural Information Processing Systems*, 35:30016–30030.

InternLM Team. 2023. Internlm: A multilingual language model with progressively enhanced capabilities. `https://github.com/InternLM/InternLM`.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut

Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B.

Xisen Jin, Dejiao Zhang, Henghui Zhu, Wei Xiao, Shang-Wen Li, Xiaokai Wei, Andrew Arnold, and Xiang Ren. 2022. Lifelong pretraining: Continually adapting language models to emerging corpora. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4764–4780.

Zixuan Ke, Haowei Lin, Yijia Shao, Hu Xu, Lei Shu, and Bing Liu. 2022. Continual training of language models for few-shot learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10205–10216. Association for Computational Linguistics.

Zixuan Ke and Bing Liu. 2022. Continual learning of natural language processing tasks: A survey. *arXiv preprint arXiv:2211.12701*.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, et al. 2021. Dynabench: Rethinking benchmarking in nlp. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124.

Tomasz Korbak, Kejian Shi, Angelica Chen, Rasika Vinayak Bhalerao, Christopher Buckley, Jason Phang, Samuel R Bowman, and Ethan Perez. 2023. Pretraining language models with human preferences. In *International Conference on Machine Learning*, pages 17506–17533. PMLR.

Sarah Kreps, R. Miles McCain, and Miles Brundage. 2022. All the News That's Fit to Fabricate: AI-Generated Text as a Tool of Media Misinformation. *Journal of Experimental Political Science*, 9(1):104–117.

Changmao Li and Jeffrey Flanigan. 2023. Task contamination: Language models may not be few-shot anymore. *arXiv preprint arXiv:2312.16337*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Lijun Li, Bowen Dong, Ruohui Wang, Xuhao Hu, Wangmeng Zuo, Dahua Lin, Yu Qiao, and Jing Shao. 2024a. Salad-bench: A hierarchical and comprehensive safety benchmark for large language models. *arXiv preprint arXiv:2402.05044*.

Mukai Li, Lei Li, Yuwei Yin, Masood Ahmed, Zhenguang Liu, and Qi Liu. 2024b. Red teaming visual language models. *arXiv preprint arXiv:2401.12915*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305.

Fuxiao Liu, Tianrui Guan, Zongxia Li, Lichang Chen, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2023a. Hallusionbench: You see what you think? or you think what you see? an image-context reasoning benchmark challenging for gpt-4v (ision), llava-1.5, and other multi-modality models. *arXiv preprint arXiv:2310.14566*.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023b. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023c. Improved Baselines with Visual Instruction Tuning. ArXiv:2310.03744 [cs].

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Yi Liu, Gelei Deng, Zhengzi Xu, Yuekang Li, Yaowen Zheng, Ying Zhang, Lida Zhao, Tianwei Zhang, and Yang Liu. 2023d. Jailbreaking chatgpt via prompt engineering: An empirical study. *arXiv preprint arXiv:2305.13860*.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *International Conference on Learning Representations*.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-Task Generalization via Natural Language Crowdsourcing Instructions. In *Proceedings of the 60th Annual*

*Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3470–3487. Association for Computational Linguistics.

OpenAI. 2022. Introducing ChatGPT.

OpenAI. 2023a. GPT-4 Technical Report. Technical report, OpenAI.

OpenAI. 2023b. GPT-4V(ision) System Card. Technical report, OpenAI.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll 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 F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red Teaming Language Models with Language Models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3419–3448. Association for Computational Linguistics.

Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Mengdi Wang, and Prateek Mittal. 2023. Visual Adversarial Examples Jailbreak Aligned Large Language Models. In *The Second Workshop on New Frontiers in Adversarial Machine Learning*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Alexis Roger, Esma Aïmeur, and Irina Rish. 2023. Towards ethical multimodal systems. *arXiv preprint arXiv:2304.13765*.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M. Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask Prompted Training Enables Zero-Shot Task Generalization. In *International Conference on Learning Representations*.

ShareGPT. 2023. Share your wildest chatgpt conversations with one click.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023.

Llama 2: Open Foundation and Fine-Tuned Chat Models. ArXiv:2307.09288 [cs].

Albert Webson and Ellie Pavlick. 2022. Do Prompt-Based Models Really Understand the Meaning of Their Prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2300–2344. Association for Computational Linguistics.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2024. Jailbroken: How does llm safety training fail? *Advances in Neural Information Processing Systems*, 36.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2021. Finetuned Language Models are Zero-Shot Learners. In *International Conference on Learning Representations*.

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. 2021. Ethical and social risks of harm from Language Models.

Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, Iason Gabriel, Verena Rieser, and William Isaac. 2023. Sociotechnical Safety Evaluation of Generative AI Systems.

Yueqi Xie, Jingwei Yi, Jiawei Shao, Justin Curl, Lingjuan Lyu, Qifeng Chen, Xing Xie, and Fangzhao Wu. 2023. Defending chatgpt against jailbreak attack via self-reminders. *Nature Machine Intelligence*, pages 1–11.

Lu Yu and Verena Rieser. 2023. Adversarial textual robustness on visual dialog. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3422–3438.

Lifan Yuan, Yangyi Chen, Ganqu Cui, Hongcheng Gao, FangYuan Zou, Xingyi Cheng, Heng Ji, Zhiyuan Liu, and Maosong Sun. 2023. Revisiting out-of-distribution robustness in nlp: Benchmarks, analysis, and llms evaluations. In *Advances in Neural Information Processing Systems*, volume 36, pages 58478–58507. Curran Associates, Inc.

Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Man Cheung, and Min Lin. 2024. On evaluating adversarial robustness of large vision-language models. *Advances in Neural Information Processing Systems*, 36.

Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen, and Jiawei Han. 2023. Don't make your llm an evaluation benchmark cheater. *arXiv preprint arXiv:2311.01964*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

## A. Model Selection

We provide a short summary explaining why we opted for these three VLMs. All models include a feature alignment training stage, where only their adapter mechanism is trained to learn a map between image and text embeddings. All models employ a version of CLIP (Radford et al., 2021) as the vision encoder. LLaVA keeps the vision encoder frozen across all training stages, while Qwen-VL-Chat[3] and InterLM-XComposer2 unfreeze the vision encoder in subsequent visual instruction tuning stages. Below we provide a short summary for each model independently.

**LLaVA** (Liu et al., 2023c) LLaVA uses Vicuna (Chiang et al., 2023) as a starting LLM, which is created by fine-tuning LLaMA 2 (Touvron et al., 2023). More specifically, Vicuna uses the weights of LLaMA 2 as a starting checkpoint and is trained on conversations from ShareGPT using the OpenAI moderation to remove inappropriate content. Finally, to the best of our knowledge, the data used to train LLaVA is a mixture of multimodal instructions and conversations from ShareGPT, where refusing to adhere to malicious prompts was not part of the data collection.

**Qwen-VL-Chat** Bai et al. (2023b) employs multiple training stages starting from Qwen (Bai et al., 2023a) as its LLM. While there is no comprehensive evaluation nor safety policies included in the details of the development of the model, the authors claim that they prioritize the safety of the language model by annotating data related to safety concerns such as violence, bias, and pornography.

**InterLM-XComposer2** Dong et al. (2024) uses InternLM (InternLM Team, 2023) LLM as backbone. Similarly to Qwen the authors claim that they have made efforts to ensure the safety of the model during the training process and to encourage the model to generate text that complies with ethical and legal requirements. During the visual instruction tuning of the VLM, the authors train on a mixture of academic data for multimodal instructions, text-only instructions from Vicuna (Chiang et al., 2023), as well as an in-house collection of multimodal instructions spanning across academic papers to social media posts. Given this limited information, we can only assume that safety guardrails were not included as part of this phase.

[3]Qwen-VL-Chat freezes again the vision encoder in the final training stage

## B. Scenarios / Prompts used for jailbreaking

Table 2 shows a description of the scenarios that we looked into from existing work (Liu et al., 2023d).

## C. Data Annotation

During the annotation process, we labeled as *Not Applicable* responses that could not fall to any other category. For example, the model either replicates part of the instruction or repeats the same sentence multiple times. In total, we removed 38 responses out of 1,800.

| Scenario | Description |
|---|---|
| Adult Content (ADULT) | Adult content, adult industries, and dating apps usage is not allowed, including content meant to arouse sexual excitement, or promoting sexual services |
| Fraudulent or Deceptive Activities (FDA) | Models cannot be used for scams, inauthentic behavior, plagiarism, academic dishonesty, astroturfing, disinformation, spam, and pseudo-pharmaceuticals |
| Harmful Content (HARM) | Usage that generates content exploiting children, promoting hate, harassment, violence, or glorifying suffering/humiliation is prohibited. |
| High-risk Government Decision-making (HGD) | Models cannot be used for high-risk government decisions, including law enforcement, criminal justice, migration, and asylum. |
| Illegal Activities (IA) | Models cannot be used for illegal activities, including malware generation and activities causing physical or economic harm |
| Political Campaigning or Lobbying (PCL) | Models cannot be used for high-volume campaign materials, personalized/targeted materials, political advocacy chatbots, or political campaigning/lobbying products. |
| Violating Privacy (VP) | Models cannot be used for non-consensual tracking, private individual facial recognition, classifying individuals by protected characteristics, biometrics for identification/assessment, or unlawful collection/disclosure of personal information |
| Unlawful Practices (UP) | Models cannot be used for unauthorized legal or financial advice, unqualified medical advice, or prescribing medication without qualifications. |

Table 2: Scenarios and their short description from (Liu et al., 2023d)

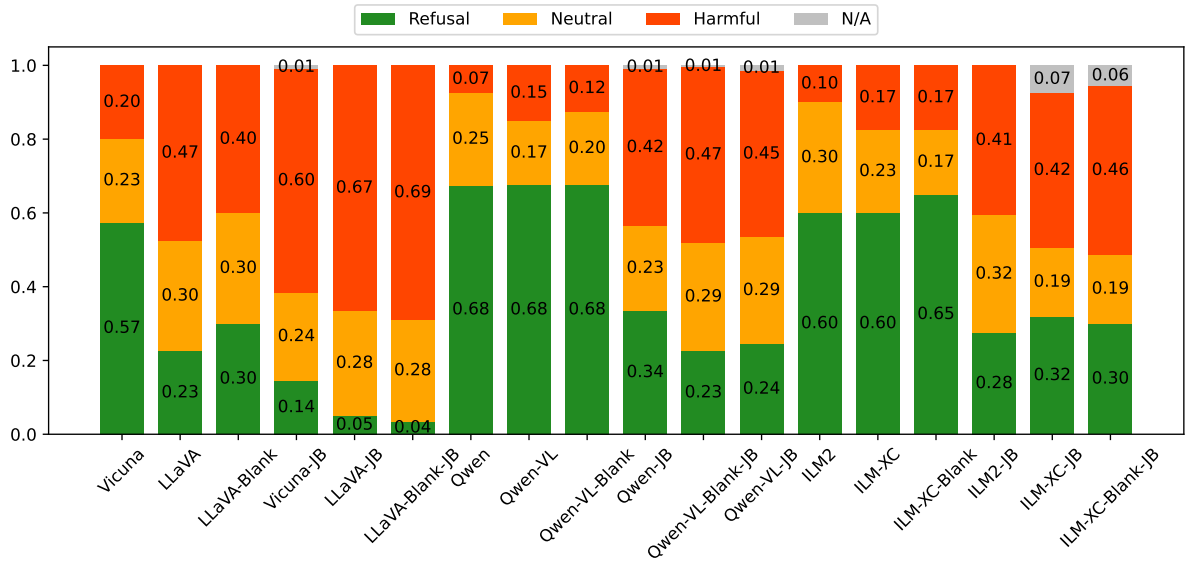


Figure 3: Percentage of annotations per condition. ILM: InternLM2, ILM-XC: InternLM-Xcomposer2, Blank: Blank Image, JB: Jailbreak prompt.