# Defending speech-enabled LLMs against adversarial jailbreak threats

**Anonymous ACL submission**


## Abstract

Speech Language Models (SLMs), which integrate speech input modalities with Large Language Models (LLMs), are emerging as powerful tools in natural language processing. However, like many other deep learning models, SLMs are vulnerable to adversarial jailbreak attacks. In this paper, we report on adversarial defense techniques to enhance the robustness of SLMs against such attacks and enhance their safety alignment. We develop and implement SLM-specific adversarial training using synthesized speech instructions for both harmful and benign samples. Our method significantly outperforms unaligned and safety-aligned models in safety metrics against adversarial attacks in different attack configurations. The adversarially trained model demonstrates an almost three-fold improvement in safety responses. A comprehensive ablation study further validates the effectiveness of our approach as a countermeasure against attacks, paving the way for safer SLM deployment in real-world applications.


## 1 Introduction

Multi-modal large language models (LLM) that can ingest different input modalities such as speech, text and images, have gained increasing popularity owing to their versatile capabilities (Liu et al., 2024a; Yin et al., 2023; Das et al., 2024). However, broad usage of such technologies mandates careful considerations of their safety and to ensure guardrails against their misuse. Multi-modal LLMs are particularly susceptible to adversarial jailbreak threats owing to the additional modality for attack (Qi et al., 2024). Previous works have developed pre-processing defenses against adversarial attacks on vision and speech language models (Qi et al., 2024; Peri et al., 2024). However, these defenses work reactively during inference, and can negatively impact on the utility of the system. On the other hand, limited studies (Zhou et al., 2024; Zheng et al., 2024) have explored proactive defenses such as adversarial training (AT) (Madry et al., 2019) in the generative LLM regime, although adversarial training is a de facto standard in adversarial robustness.

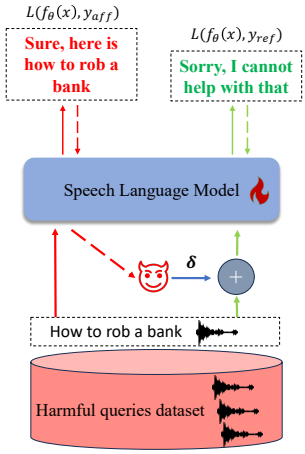


Figure 1: Proposed adversarial training defense against jailbreak threats. We use an attacker-in-the-loop to robustify the model.

In this work, we explore adversarial training in the context of speech language models, focusing on scenarios where adversaries perturb the speech input. Specifically, we propose training strategies using a combination of adversarial training and instruction fine-tuning steps during optimization to effectively leverage the adversarial perturbations, while retaining the helpfulness of the systems. We use synthesized speech data from five different datasets generated with state-of-the-art text-to-speech (TTS) models to perform adversarial training. In this work, we limit our analysis to spoken question-answer (Q&A) setup, where the input question is entered in spoken form and the model produces text response.

Our main contributions are as follows:

1. We demonstrate that adversarial training can enhance the robustness of SLMs against ad-

versarial white-box attacks, while also maintaining or improving their utility.

2. We explore different adversarial training configurations and data sizes using synthetically generated speech samples for effective robustness while being computationally tractable.

3. To the best of our knowledge, ours is the first work to show that adversarial training significantly enhances the safety of an SLM, surpassing that of safety alignment across different attack configurations.

## 2 Methodology

### 2.1 Preliminaries and Related Work

**Safety Alignment and jailbreak attacks.** Modern LLMs achieve their diverse capabilities through training on vast datasets sourced from the internet, which may include offensive content, harmful information, and discriminatory biases. This compromises their alignment with human values and societal ethics (Hua et al., 2024; Kumar et al., 2024). To enhance LLM safety, methods like reinforcement learning from human feedback (RLHF) and direct preference optimization (DPO) have been proposed, although they require substantial labeled human preference data, which is often scarce. Conversely, safety training, involving harmful queries and refusal responses, can be automated (Hendrycks et al., 2020; Askell et al., 2021; Ouyang et al., 2022; Shen et al., 2024; Wei et al., 2023a). Jailbreak attacks exploit LLM vulnerabilities to override safety mechanisms, commonly through prompt injection, where crafted prompts induce harmful content generation (Liu et al., 2024c,b; Piet et al., 2024).

**Adversarial Attacks.** An alternative to prompt-based jailbreaking is adversarial attacks, which perturb input data to induce harmful responses from the model (Zou et al., 2023; Guo et al., 2021). Multi-modal LLMs, processing data from various modalities like images and audio, are particularly vulnerable to these attacks (Qi et al., 2023; Peri et al., 2024). Adversarial attacks are classified as white-box, where the attacker has full access to the model, and black-box, with no access (Guo et al., 2019). This work focuses on the projected gradient descent (PGD) method, a white-box attack that iteratively perturbs inputs to achieve adversarial goals (Madry et al., 2019) as follows:

$$x^{i+1} = x^i - \alpha \cdot \operatorname{sgn}\left(\nabla_x \mathcal{L}\left(\mathcal{F}\left(x^i + \delta\right), y_{aff}\right)\right) \tag{1}$$

where sgn is the sign operator, $\alpha$ is the step size, $x^i$ is the input at the $i^{th}$ iteration, and $x^0$ is the original signal. In jailbreak scenarios, the perturbations aim to compel the model to produce affirmative responses ($y_{aff}$) to harmful queries.

**Defenses.** Defenses against adversarial jailbreak attacks include system prompts (Jin et al., 2024), self-reminders (Xie et al., 2023), and in-context safe demonstrations (Wei et al., 2023b). Additionally, textual or multi-modal unlearning has been proposed (Chakraborty et al., 2024). However, these methods often require crafting model-specific prompts and may not generalize across well. In classical adversarial robustness literature, adversarial training (Madry et al., 2019) and its derivatives (Rade and Moosavi-Dezfooli, 2021) have become standard for enhancing model robustness against adversarial perturbations. Adversarial training is effective against various threats by incorporating adversarial examples during training. Recent defense techniques inspired by adversarial training include training prompt controls alongside user prompts to combat grey-box and black-box attacks (Mo et al., 2024), robust prompt optimization incorporating adversaries into the defense objective (Zhou et al., 2024), and continuously optimizing safety prompts as trainable embeddings (Zheng et al., 2024). In this work, we introduce the first adversarial training method specifically designed to defend against adversarial attacks for SLMs.

### 2.2 Adversarial Training for SLMs

As noted in Section 2.1, adversarial training involves utilizing PGD-perturbed samples during training as shown below:

$$\min \mathbb{E}_{(x,y_{ref})\in\mathcal{D}}[\min_{\delta\in\mathcal{S}} \mathcal{L}(\theta, x+\delta, y_{aff})] \tag{2}$$

where $\mathcal{L}$ represents the loss of the adversarially perturbed samples to generate affirmative responses ($y_{aff}$), $\theta$ are the parameters of the model, $y_{ref}$ are the refusal answers to harmful questions and $\delta$ is the adversarial perturbation. As shown in Eq. 2, the adversarial defense mechanism consists of two loops: the inner loop represents the attacker trained using affirmative responses to harmful prompts, while the outer loop represents our model, trained to refuse responding to harmful queries. This is

similar to the work on text-only LLMs proposed by Mazeika et al. (Mazeika et al., 2024b), where they employ an "away loss" and a "toward loss" akin to our refusal and affirmative responses. Note that we incorporate both adversarially perturbed instances and "clean" data (helpful Q&A) to train the SLM to retain performance on benign queries.

# 3 Experimental setup

## 3.1 Model

The SLM comprises of an audio encoder, a Conv-1D downsampling module, and an LLM model. The audio encoder is a 24-layer conformer model with embedding size of 768 and 8 attention heads that extracts acoustic features from speech. The Conv-1D downsampling module is employed to reduce the frame rate of the audio features through a learnable module, to mitigate the length discrepancy between the audio features and the input text tokens. This module consists of two successive blocks of 1-D convolutions (kernel size: 3). For the downstream LLM, we experiment with two different candidates from the FLAN-T5 family of models (Chung et al., 2024): FLAN-T5-large (800M) and FLAN-T5-XL (3B).

## 3.2 Data

Our training data can be categorized into 3 based on the task: **ASR pretraining**, **Q&A training**, and **adversarial training** as shown in Table 1. For the spoken Q&A task with speech instruction and textual response pairs, we synthesized speech corresponding to publicly available text-to-text instruction tuning datasets, (Sun et al., 2024; Taori et al., 2023) using an in-house text-to-speech (TTS) system. We created 2 more subsets of this *Help-all* dataset with different number of samples randomly chosen.

As shown in Figure 1, adversarial training requires affirmative responses as well as the refusal responses to spoken queries. We combined the harmful text queries and safe responses (refusal to respond to harmful queries) from 4 sources (Mazeika et al., 2024a; Zhang et al., 2024; Zou et al., 2023; Chao et al., 2024) for this purpose. We synthesized audio corresponding to the text queries using XXTTS (Casanova et al., 2024) and Bark-TTS[1], using multiple speakers for each sample. For the affirmative target responses, we generated text using automatic methods by sampling from different affirmations such as "Sure, I can help you with that". For the refusal target responses, we use the responses from the text sources directly. For evaluations, we reserved 66 samples from the synthesized harmful queries dataset.

Table 1: Datasets for training and adversarial attacks

| Name | Task | # samples | # hrs |
|---|---|---|---|
| Libri | ASR pre-training | 280K | 960 |
| Help-all | Q&A training | 160K | 150 |
| Help-50K | | 50K | 50 |
| Help-25K | | 25K | 25 |
| Harm-train | Adversarial training | 6.6K | 4 |
| Eval | Adversarial attacks | 66 | 0.04 |

## 3.3 Setup

**Training** We pretrained the SLM on ASR task over 26 epochs using the *Libri* dataset mentioned in Table 1. Subsequently, we trained two models for 200 epochs using the *Help-all* data for the Q&A task using the FLAN-T5-large and FLAN-T5-xl downstream LLMs. These models are referred to as **Base-l** and **Base-xl** in the paper. We then performed adversarial training for an additional 100 epochs, alternating between 1 epoch of adversarial training (using *Harm-train* data) and 1 epoch of Q&A training with helpful data[2]. These models are reported as **AT-l** and **AT-xl** depending on the downstream LLM. During adversarial training, we used an attack step size of 0.001 with 30 and 50 attack iterations for the AT-l and AT-xl models respectively. As a baseline, we trained model for safety alignment for 100 epochs following the Base model, using the *Help-all* data. These models are called **ST-l** and **ST-xl** in the rest of the paper.

**Evaluation** We attack the model using the white-box PGD attack, and evaluate along two dimensions: safety and relevance. The relevance rate (RR) assesses whether the SLM's response is pertinent to the question, computed as $RR = N_r/N$, where $N_r$ is the number of relevant responses out of $N$ total questions. The safety rate (SR) determines the harmfulness of the responses, $SR = N_s/N$, where $N_s$ is the number of safe responses. The safety and relevance evaluation was conducted using Claude model as judge (similar to (Peri et al., 2024)), which grades the answers based on the questions, given a predefined

[1] https://github.com/suno-ai/bark

[2] We also experimented with different ratio of adversarial training epochs, but found this to setting to work slightly better

prompt. In addition, we also evaluate the effort needed by the attacker to produce successful jailbreaks through the average number of iterations to achieve successful attack. A larger number of iterations indicates a more robust attack.

## 4 Results and Discussion

From Fig. 2, we can observe that the proposed adversarial training method significantly outperforms the base model (without safety training) and the safety trained model (ST) at all attack iterations. However, as the iterations increase, the attack becomes more potent, causing the safety metric to decrease for all models. Furthermore, we can observe that models with larger downstream LLM (-xl) are more robust than the smaller models (-l). From Fig. 3, we can observe the effect of attack step size ($\alpha$ in Eq.1) on the safety of the models. Here, we again observe that the proposed model outperforms the other models by a large margin in all cases.

In Table 2, we present an evaluation of models across several metrics mentioned in Section 3.3, and also the amount of helpful data used during training. The results show that the adversarially trained model outperforms all other methods across all metrics. Specifically, the AT models achieve a higher SR and require more iterations for the adversarial attack to succeed. This indicates the proposed defense mechanism increases the attack budget. Another observation is that using more samples of helpful data improves the relevance metric, which suggests a trade-off between computational budget and performance.

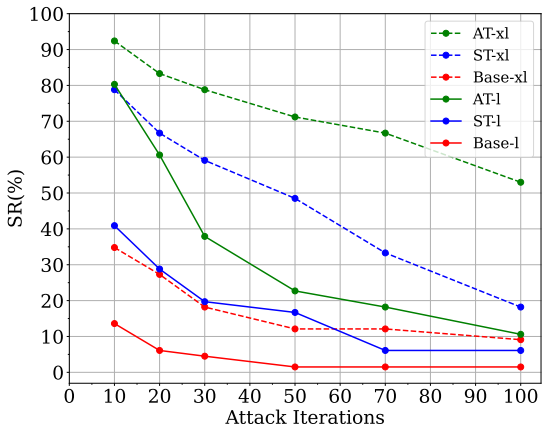


Figure 2: Safety versus attack iterations ($\alpha$=0.001). Proposed method (AT) outperforms baselines across the board for both model sizes.

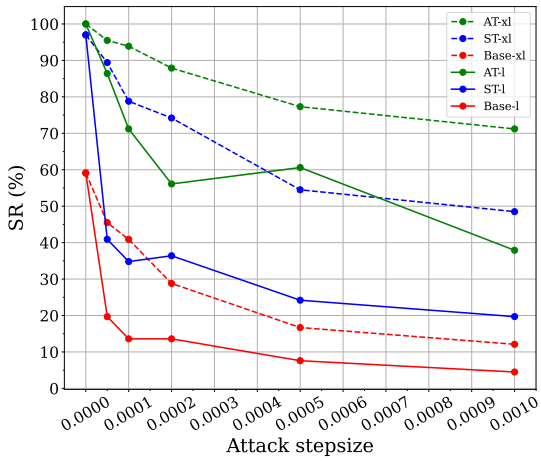


Figure 3: Safety versus attack step size (iterations: 30(l), 50(xl). Proposed method (AT) outperforms baselines in all scenarios.

Table 2: Performance of the proposed method in different data regimes ($\alpha$=0.001, iterations:30(l), 50(xl)).

| Model | %SR | %RR | Avg. iter | Helpful data |
|---|---|---|---|---|
| Base-l | 4.5 | 53.0 | 3 | Help-all |
| ST-l | 19.7 | 69.0 | 12 | Help-all |
| AT-l | 40.9 | 75.0 | 28 | Help-25K |
| AT-l | 53.3 | 79.0 | 21 | Help-50K |
| AT-l | 43.3 | 81.0 | 21 | Help-all |
| Base-xl | 12.1 | 78.0 | 8 | Help-all |
| ST-xl | 48.5 | 83.0 | 17 | Help-all |
| AT-xl | 18.2 | 77.0 | 19 | Help-25K |
| AT-xl | 60.6 | 84.0 | 23 | Help-50K |
| AT-xl | 71.2 | 86.0 | 19 | Help-all |

## 5 Conclusion

In this work, we proposed adversarial training as a defense mechanism against adversarial jailbreak threats on speech LLMs. We used synthesized speech samples on a collection of instruction-tuning and safety-alignment datasets to train our models. We performed extensive experiments in different attack configurations, and showed that the proposed methods significantly outperform vanilla safety alignment techniques. Furthermore, we show that our methods increase the attackers required budget of iterations to produce successful jailbreaks. Future work would focus on improving the computational requirements of adversarial training.

## Limitations

In this work, we avail a preference model as judge to assess safety of SLMs. However, we acknowledge that such a judge may not always align with

human judgement, which might lead to a minor margin of error in our safety annotations, which we plan to address in future work. Furthermore, our work provides a limited exploration into SLMs trained with safety-aligned text LLMs, although SLMs themselves are trained with safety-aligned spoken data. However, given that our approach has already shown efficacies of such models, we leave the thorough exploration to future work. Lastly, concerns of misuse by malicious practitioners prevent us from releasing the training datasets and models, limiting replication by other researchers.

## Ethics Statement

All speech datasets we use have anonymous speakers. We do not have any access to nor try to create any PII (Personal Identifiable Information) of speakers, and our model neither identifies speakers nor uses speaker embeddings. Furthermore, we obtained necessary consent from all the participants of our data collection efforts following approval by an internal review board.

While we acknowledge the ethical risks associated with jailbreaking techniques, this work represents a valuable contribution towards a deeper understanding of speech-language model capabilities and limitations. Our aim is to enable further research that improves model robustness, leading to safer and more beneficial applications. By responsibly investigating methods to circumvent restrictions, we shed light on potential vulnerabilities that could be exploited by malicious attackers if left unaddressed. Critically, our work also proposes and evaluates countermeasures to mitigate such jailbreaking attacks. While we encourage ethical debate on such emerging issues, we believe the merits of responsible disclosure, proactive security improvements, and developing defensive techniques outweigh any potential risks associated with our narrow jailbreaking experiments under controlled conditions. Overall, our work aims to make progress towards more secure and robust multimodal speech-language models.

## References


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. *Preprint*, arXiv:2112.00861.

Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Göknar, Iulian Gulea, Logan Hart, Aya Aljafari, Joshua Meyer, Reuben Morais, Samuel Olayemi, et al. 2024. Xtts: a massively multilingual zero-shot text-to-speech model. *arXiv preprint arXiv:2406.04904*.

Trishna Chakraborty, Erfan Shayegani, Zikui Cai, Nael Abu-Ghazaleh, M Salman Asif, Yue Dong, Amit K Roy-Chowdhury, and Chengyu Song. 2024. Cross-modal safety alignment: Is textual unlearning all you need? *arXiv preprint arXiv:2406.02575*.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramèr, Hamed Hassani, and Eric Wong. 2024. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. *Preprint*, arXiv:2404.01318.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Nilaksh Das, Saket Dingliwal, Srikanth Ronanki, Rohit Paturi, Zhaocheng Huang, Prashant Mathur, Jie Yuan, Dhanush Bekal, Xing Niu, Sai Muralidhar Jayanthi, Xilai Li, Karel Mundnich, Monica Sunkara, Sundararajan Srinivasan, Kyu J Han, and Katrin Kirchhoff. 2024. Speechverse: A large-scale generalizable audio language model. *Preprint*, arXiv:2405.08295.

Chuan Guo, Jacob R. Gardner, Yurong You, Andrew Gordon Wilson, and Kilian Q. Weinberger. 2019. Simple black-box adversarial attacks. *Preprint*, arXiv:1905.07121.

Chuan Guo, Alexandre Sablayrolles, Hervé Jégou, and Douwe Kiela. 2021. Gradient-based adversarial attacks against text transformers. *Preprint*, arXiv:2104.13733.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2020. Aligning AI with shared human values. *CoRR*, abs/2008.02275.

Shangying Hua, Shuangci Jin, and Shengyi Jiang. 2024. The limitations and ethical considerations of chatgpt. *Data intelligence*, 6(1):201–239.

Haibo Jin, Leyang Hu, Xinuo Li, Peiyan Zhang, Chonghan Chen, Jun Zhuang, and Haohan Wang. 2024. Jailbreakzoo: Survey, landscapes, and horizons in jailbreaking large language and vision-language models. *arXiv preprint arXiv:2407.01599*.

Ashutosh Kumar, Sagarika Singh, Shiv Vignesh Murty, and Swathy Ragupathy. 2024. The ethics of interaction: Mitigating security threats in llms. *arXiv preprint arXiv:2401.12273*.

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

Xiaogeng Liu, Zhiyuan Yu, Yizhe Zhang, Ning Zhang, and Chaowei Xiao. 2024b. Automatic and universal prompt injection attacks against large language models. *Preprint*, arXiv:2403.04957.

Yi Liu, Gelei Deng, Yuekang Li, Kailong Wang, Zihao Wang, Xiaofeng Wang, Tianwei Zhang, Yepang Liu, Haoyu Wang, Yan Zheng, and Yang Liu. 2024c. Prompt injection attack against llm-integrated applications. *Preprint*, arXiv:2306.05499.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2019. Towards deep learning models resistant to adversarial attacks. *Preprint*, arXiv:1706.06083.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. 2024a. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *Preprint*, arXiv:2402.04249.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, et al. 2024b. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*.

Yichuan Mo, Yuji Wang, Zeming Wei, and Yisen Wang. 2024. Studious bob fight back against jailbreaking via prompt adversarial tuning. *arXiv preprint arXiv:2402.06255*.

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

Raghuveer Peri, Sai Muralidhar Jayanthi, Srikanth Ronanki, Anshu Bhatia, Karel Mundnich, Saket Dingliwal, Nilaksh Das, Zejiang Hou, Goeric Huybrechts, Srikanth Vishnubhotla, Daniel Garcia-Romero, Sundararajan Srinivasan, Kyu J Han, and Katrin Kirchhoff. 2024. Speechguard: Exploring the adversarial robustness of multimodal large language models. *Preprint*, arXiv:2405.08317.

Julien Piet, Maha Alrashed, Chawin Sitawarin, Sizhe Chen, Zeming Wei, Elizabeth Sun, Basel Alomair, and David Wagner. 2024. Jatmo: Prompt injection defense by task-specific finetuning. In *European Symposium on Research in Computer Security*, pages 105–124. Springer.

Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Peter Henderson, Mengdi Wang, and Prateek Mittal. 2023. Visual adversarial examples jailbreak aligned large language models. *Preprint*, arXiv:2306.13213.

Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Peter Henderson, Mengdi Wang, and Prateek Mittal. 2024. Visual adversarial examples jailbreak aligned large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 21527–21536.

Rahul Rade and Seyed-Mohsen Moosavi-Dezfooli. 2021. Helper-based adversarial training: Reducing excessive margin to achieve a better accuracy vs. robustness trade-off. In *ICML 2021 Workshop on Adversarial Machine Learning*.

Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. 2024. "do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. *Preprint*, arXiv:2308.03825.

Tianxiang Sun, Xiaotian Zhang, Zhengfu He, Peng Li, Qinyuan Cheng, Xiangyang Liu, Hang Yan, Yunfan Shao, Qiong Tang, Shiduo Zhang, Xingjian Zhao, Ke Chen, Yining Zheng, Zhejian Zhou, Ruixiao Li, Jun Zhan, Yunhua Zhou, Linyang Li, Xiaogui Yang, Lingling Wu, Zhangyue Yin, Xuanjing Huang, Yu-Gang Jiang, and Xipeng Qiu. 2024. Moss: An open conversational large language model. *Machine Intelligence Research*, 21(5):888–905.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2023a. Jailbroken: How does llm safety training fail? *Preprint*, arXiv:2307.02483.

Zeming Wei, Yifei Wang, and Yisen Wang. 2023b. Jailbreak and guard aligned language models with only few in-context demonstrations. *arXiv preprint arXiv:2310.06387*.

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*, 5(12):1486–1496.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*.

Zhexin Zhang, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. 2024. Safety-bench: Evaluating the safety of large language models. *Preprint*, arXiv:2309.07045.

Chujie Zheng, Fan Yin, Hao Zhou, Fandong Meng, Jie Zhou, Kai-Wei Chang, Minlie Huang, and Nanyun Peng. 2024. On prompt-driven safeguarding for large language models. In *Forty-first International Conference on Machine Learning*.

Andy Zhou, Bo Li, and Haohan Wang. 2024. Robust prompt optimization for defending language models against jailbreaking attacks. *arXiv preprint arXiv:2401.17263*.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *Preprint*, arXiv:2307.15043.