



Defending speech-enabled LLMs against adversarial jailbreak threats

Antonios Alexos^{†1}, Raghveer Peri², Sai Muralidhar Jayanthi², Metehan Cekic², Srikanth Vishnubhotla², Kyu J. Han^{*2}, Srikanth Ronanki²

¹University of California, Irvine

²Amazon Web Services, AI labs

aalexos@uci.edu, {raghperi, saimucja, mcekic, srikvish, kyujhan, ronanks}@amazon.com

Abstract

The additional modality (such as speech) in multimodal large language models (LLM) increases their vulnerability to adversarial jailbreak attacks. Adversarial training (AT) techniques have shown great promise as defenses in traditional adversarial robustness literature. But they are less explored as countermeasures in speech-enabled LLMs due to the limited availability of training data and computational complexity. In this work, we develop AT techniques tailored to speech LLMs using a combination of synthesized harmful and benign queries. We experiment with different training data configurations, and evaluate the methods on strong white-box adversarial attacks. We demonstrate through extensive ablations on 2 models of different sizes that using just 4hrs of harmful speech queries for AT (with 150 hours of benign speech) can provide significant gains compared to vanilla safety fine-tuning, improving safety by 45%-300% relative depending on the model.

Index Terms: Speech large language models, adversarial training, jailbreak attacks, LLM safety, responsible AI

1. Introduction

The recent expansive proliferation of multi-modal large language models (LLMs) mandates careful considerations of their safety and protections against their misuse. These models are more susceptible to adversarial jailbreak threats compared to their text-only counterparts owing to the additional modality for attack [1]. This increased vulnerability has prompted researchers to explore various defensive strategies. Previous works have developed pre-processing defenses against adversarial attacks on vision and speech language models [1, 2]. However, these defenses work *reactively* during inference, and can negatively impact the utility of the system. On the other hand, *pro-active* defenses such as adversarial training (AT) have found limited success in the generative LLM regime due to their computational complexity [3, 4]. It is particularly challenging to incorporate these techniques in the speech domain due to the unavailability of high-quality data for AT. In this work, we present adversarial training techniques to successfully safeguard speech LLMs against perturbation-based adversarial jailbreak attacks by leveraging synthetic speech data.

We propose training strategies using a combination of adversarial training and instruction fine-tuning steps to effectively leverage adversarially perturbed samples, while retaining the helpfulness of the systems, and simultaneously maintaining computational tractability. We generate synthetic speech data from 4 different text sources using state-of-the-art text-to-speech (TTS) models to perform adversarial training.

* Author was affiliated with Amazon at time of work

† Work done during internship

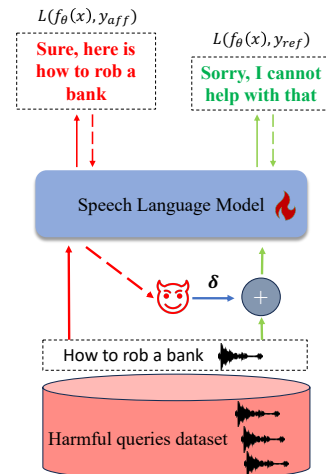


Figure 1: Proposed adversarial training defense against jailbreak threats. We use an attacker-in-the-loop (affirmative responses y_{aff} to harmful queries) and train model on perturbed speech to generate refusal responses y_{ref} .

We evaluate the proposed methods against strong adversarial jailbreak threats using white-box threat models, where we assume the attackers have full access (including weights) to the models. We show that the proposed techniques outperform the safety of speech LLMs compared to vanilla safety fine-tuning in the strongest attack scenario by 45%-300% relative depending on the backend LLM. To the best of our knowledge, this is the first work to successfully tailor AT techniques to defend speech LLMs against adversarial jailbreaks.

Our main contributions are as follows:

1. We demonstrate through experiments on 2 models of different sizes that adversarial training can generally improve robustness of speech LLMs against strong adversarial white-box threats.
2. We experiment with different training data settings using synthetically generated speech samples, and show that using just 4 hrs of speech containing harmful questions and 150 hrs of instruction fine-tuning data can significantly improve safety. We also show that data size should be carefully chosen depending on the size of the model.
3. We perform extensive ablations with a range of different attack strength configurations, and show that adversarial training significantly enhances the safety of speech LLMs compared to vanilla safety fine-tuning across all threat scenarios.

2. Methodology

2.1. Preliminaries and Related Work

2.1.1. 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 [5, 6]. 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 [7, 8, 9, 10]. Jailbreak attacks exploit LLM vulnerabilities to override safety mechanisms, commonly through prompt injection, where crafted prompts induce harmful content generation [11, 12].

2.1.2. Adversarial Attacks

An alternative to prompt-based jailbreaking is adversarial attacks, which perturb input data to induce harmful responses from the model [13, 14]. Multi-modal LLMs, processing data from various modalities like images and audio, are particularly vulnerable to these attacks [15, 2]. Adversarial attacks are classified as white-box, where the attacker has full access to the model, and black-box, with no access [16]. This work focuses on the projected gradient descent (PGD) method [3], a white-box attack that iteratively perturbs inputs to achieve adversarial goals as follows:

$$x^{i+1} = x^i - \alpha \cdot \text{sgn} \left(\nabla_x \mathcal{L}_{\mathcal{P}\mathcal{G}\mathcal{D}} \left(f_{\theta} \left(x^i + \delta \right), y_{aff} \right) \right) \quad (1)$$

where sgn is the sign operator, α is the step size, x^i is the input at the i^{th} iteration, and x^0 is the original signal. We intentionally exclude the projection operator of the original PGD method, since we observed that even without projection constraints the attack on speech LLMs is barely perceivable (signal-to-noise ratio > 50dB). In jailbreak scenarios, the perturbations aim to compel the model to produce affirmative responses (y_{aff}) to harmful queries. We use cross-entropy loss between the generated output and an affirmative response to optimize the adversarial perturbation.

2.1.3. Defenses

Defenses against adversarial jailbreak attacks include system prompts [17], self-reminders [18], and in-context safe demonstrations [19]. Additionally, textual or multi-modal unlearning has been proposed [20]. However, these methods often require crafting model-specific prompts and may not generalize well. In classical adversarial robustness literature, adversarial training [3] and its derivatives [21] 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 gray-box and black-box attacks [22], robust prompt optimization incorporating adversaries into the defense objective [23], and continuously optimizing safety prompts as trainable embeddings [24]. To our knowledge, no methods target speech-enabled LLMs against adversarial jailbreaks; text-based defenses do not extend to adversarial audio inputs, so we propose adversarial training as the first defense specifically tailored for speech LLMs.

2.2. Adversarial Training for Speech LLMs

As noted in Section 2.1 and shown in Fig. 1, adversarial training involves utilizing PGD-perturbed samples during training as shown below:

$$\min_{\mathbb{E}_{(x, y_{ref}) \in \mathcal{D}}} \left[\min_{\delta \in \mathcal{S}} \mathcal{L}_{\mathcal{P}\mathcal{G}\mathcal{D}}(\theta, x + \delta, y_{aff}) \right] \quad (2)$$

where $\mathcal{L}_{\mathcal{P}\mathcal{G}\mathcal{D}}$ represents loss on the adversarially perturbed samples to generate affirmative responses (y_{aff}), y_{ref} are the refusal answers to harmful questions and δ is the adversarial perturbation. As further elaborated in Algorithm 1, the training consists of alternating between 1 AT epoch for every adv_{epoch} epochs of regular training. As shown in Eq. 2, in the AT epochs the model is fine-tuned to provide refusal responses to adversarially perturbed samples of harmful queries (x_q^{adv}). In the regular training epochs, the model is fine-tuned to provide helpful responses (y_{help}) to benign queries. The amount of *helpful* data used during adversarial training (\mathcal{D}_{help}) and the number of regular training epochs (adv_{epoch}) for each AT epoch are parameters that can be adjusted based on the model and safety/utility requirements.

Algorithm 1 Speech LLM Adversarial Training

Require: Training data $\mathcal{D}_{harm} = \{(\mathbf{x}_q, y_{aff}, y_{ref})\}_{i=1}^{N_1}$, $\mathcal{D}_{help} = \{(\mathbf{x}_q, y_{help})\}_{i=1}^{N_2}$, neural network model $f_{\theta}(\cdot)$ with parameters $\theta \in \mathbb{R}^d$, loss function $\mathcal{L}(\cdot, \cdot)$, PGD adversarial attack $\mathcal{A}(\cdot, \cdot)$, attack stepsize α , batch size B , learning rate η , rate of alternating AT epochs adv_{epoch}

- 1: **for** $t = 1, 2, \dots, T$ epochs **do**
- 2: **if** $t \bmod adv_{epoch} == 0$ **then** \triangleright Perform adversarial training
- 3: **for** $i = 1, 2, \dots, \lfloor N_1/B \rfloor$ **do** \triangleright Batch iterations
- 4: Sample a minibatch \mathcal{B} from \mathcal{D}_{harm}
- 5: $\mathbf{x}_q^{adv} \leftarrow \mathcal{A}(\mathbf{x}_q, y_{aff}, f_{\theta}, \alpha, N_{iter})$ \triangleright Generate adversarial example using Eq. 1 for N_{iter} iterations
- 6: $\mathcal{L} = \frac{1}{B} \sum_{(\mathbf{x}_q, y_{ref}) \in \mathcal{B}} \mathcal{L}(f_{\theta}(\mathbf{x}_q^{adv}), y_{ref})$ \triangleright Compute loss to refuse harmful queries
- 7: Update model parameters $\theta \leftarrow \theta - \eta \nabla_{\theta} \mathcal{L}$
- 8: **end for**
- 9: **else** \triangleright Perform regular training
- 10: **for** $i = 1, 2, \dots, \lfloor N_2/B \rfloor$ **do** \triangleright Batch iterations
- 11: Sample a minibatch \mathcal{B} from \mathcal{D}_{help}
- 12: $\mathcal{L} = \frac{1}{B} \sum_{(\mathbf{x}_q, y_{help}) \in \mathcal{B}} \mathcal{L}(f_{\theta}(\mathbf{x}_q), y_{help})$ \triangleright Compute general loss on benign queries
- 13: Update model parameters $\theta \leftarrow \theta - \eta \nabla_{\theta} \mathcal{L}$
- 14: **end for**
- 15: **end if**
- 16: **end for**

3. Experimental setup

3.1. Model

The Speech LLM 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

Table 1: Datasets for training and adversarial attacks

Name	Task	# samples	# hrs
Librispeech	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

FLAN-T5 family of models [25]: FLAN-T5-large (800M) and FLAN-T5-XL (3B).

3.2. Data

Our training data can be categorized into 3 based on the task as shown in Table 1: **Automatic Speech Recognition (ASR) pretraining**, **Q&A training** (D_{help} in Algorithm 1), and **adversarial training** (D_{harm} in Algorithm 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, [26, 27] using an in-house TTS system. We created 2 more subsets of this *Help-all* dataset with different number of samples randomly chosen, where *Help-50K* contains 50K samples and *Help-25K* contains 25K samples.

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 [28, 29, 13, 30] for this purpose. We synthesized audio corresponding to the text queries using XXTTS [31] and Bark-TTS¹, using multiple speakers (ranging from 2 to 26 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”[2]. For the refusal target responses, we use responses from the text sources directly. To the best of our knowledge, a public dataset with real human-read speech samples of jailbreak questions is not publicly available. Therefore, for evaluations, we reserved 66 samples from the synthesized harmful queries unseen by the models during training.

3.3. Setup

3.3.1. Training

We pretrained the Speech LLM on ASR task over 26 epochs using the full Librispeech dataset [32] 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 (**Base-l** and **Base-xl**). We then performed adversarial training for an additional 100 epochs, alternating between 1 epoch of adversarial training (using *Harm-train* data) for every $adv_{epoch}=1$ epoch² of Q&A training with helpful data. These models are reported as **AT-l** and **AT-xl**. 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 fine-tuned the model for safety alignment for 100 epochs following the Base model, using the *Harm-train* data (**ST-l** and **ST-xl**). Details of the training setup for these models are provided in Table 2. We perform LoRA

¹<https://github.com/suno-ai/bark>

²We experimented with few different values for adv_{epoch} and found this setting to work marginally better on a small validation set

Table 2: Details of the model training setups used in this paper

Name	Downstream LLM	Safety tuning	Adversarial training
Base-l	FLAN-T5-large	✗	✗
Base-xl	FLAN-T5-xl	✗	✗
ST-l	FLAN-T5-large	✓	✗
ST-xl	FLAN-T5-xl	✓	✗
AT-l	FLAN-T5-large	✗	✓
AT-xl	FLAN-T5-xl	✗	✓

fine-tuning for ST and AT methods with rank=16 and alpha=10, supplemented by an 8 24 GB-GPU cluster. 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, though the presented techniques can easily be extended to other speech LLM frameworks including joint speech-text input models.

3.3.2. Evaluation

We attack the model using the white-box PGD attack, and evaluate along two dimensions: safety and relevance. The safety rate (SR) determines the harmfulness of the responses, $SR = (N - N_{unsafe})/N$, where N_{unsafe} is the number of unsafe responses. The relevance rate (RR) assesses whether the Speech LLM’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. In addition, we also evaluate the effort needed by the attacker to produce successful jailbreaks, as is common in adversarial robustness literature [33]. We compute the average number of iterations to produce a successful attack for each model and attack configuration. A larger number of iterations indicates a more effective defense.

Due to the laborious and time-consuming task of obtaining human evaluations, it has become common practice in jailbreaking research to leverage LLMs as judges [34]. The safety and relevance evaluation was conducted using a public LLM (with temperature=0) as judge (similar to [2]), which grades the answers based on the questions, given a predefined prompt. We also collected human labels for safety and relevance from 4 annotators on 100 randomly chosen question-response pairs (25 pairs/annotator). The LLM evaluations obtained a 65% recall for unsafe responses comparing against the human labels. This suggests though the LLM evaluations are not perfect, they provide a reasonable estimate of the safety and relevance of the responses. Especially since the same LLM prompts were used to evaluate all models, they offer a fair comparison between the models.

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). This likely arises from the better modeling capacity of larger models to learn from adversarial inputs. Furthermore, the attack performance plateaus at 100 iterations for most models, indicating that our evaluations were performed on a converged attacker [33]. From Fig. 3, we can observe the effect of attack step size (α in Eq.1) on the safety of the models. Here, we again

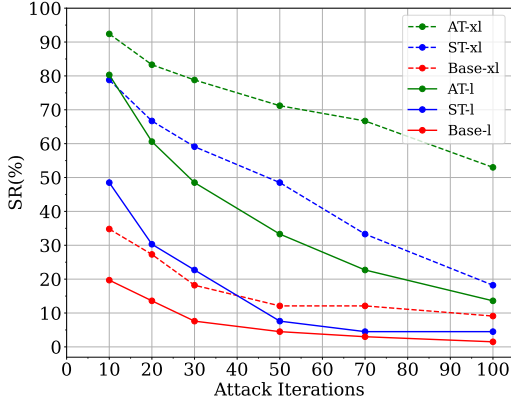


Figure 2: %Safety rate (\uparrow) versus attack iterations ($\alpha=0.001$). Proposed method (AT) outperforms baselines across the board for both model sizes with more pronounced difference under stronger attacks.

observe that the proposed model outperforms the other models by a large margin in all cases, including in the strongest attack scenario of $\alpha=0.001$. We can also observe that the baseline models' safety drops rapidly for strong attacks ($\alpha=0.001$) compared to no-attack scenario ($\alpha=0$). The ST-xl model shows a relative degradation of 50% (97.0% \rightarrow 48.5%), whereas the proposed methods show much better robustness with a relative degradation of only 28.8% (100.0% \rightarrow 71.2%).

In Table 3, we present an evaluation of models across several metrics mentioned in Section 3.3, along with 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, RSR, and require more iterations for successful adversarial attacks. This indicates the proposed defense mechanism increases the attack budget, while also reducing attack success at a fixed budget. Another observation is that for the larger model (-xl), using more amount of helpful data is beneficial, while for the smaller model (-l) larger mix of helpful data potentially leads to over-fitting. Therefore, it is important to consider the model size when determining the mix of harmful and helpful samples during adversarial training.

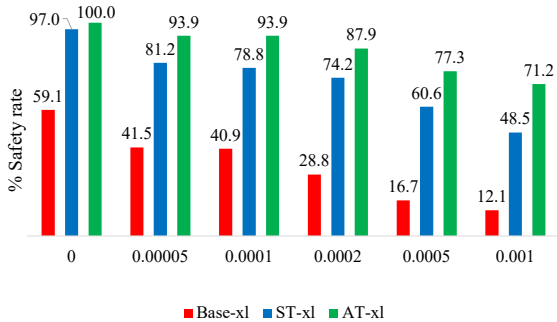


Figure 3: %Safety rate (\uparrow) versus α (attack iter.: 50). Proposed method (AT) outperforms baselines in all scenarios, particularly with stronger attacks (higher α)

Finally, we evaluated the safety of the different models when presented with speech containing unperturbed harmful questions and report the results in Fig. 4. This can be considered the baseline safety rate in the regular operating scenario of no adversarial attack. First, it can be seen that the speech-enabled LLMs

Table 3: Performance (safety and relevance) under various training data conditions ($\alpha=0.001$, iterations:50).

Model	%SR(\uparrow)	%RR(\uparrow)	Avg. iter. \uparrow	Helpful data
Base-l	4.5	53	5	Help-all
ST-l	7.6	69	13	Help-all
AT-l	25.8	75	18	Help-25K
AT-l	33.3	79	19	Help-50K
AT-l	19.7	81	14	Help-all
Base-xl	12.1	78	8	Help-all
ST-xl	48.5	83	17	Help-all
AT-xl	18.2	77	19	Help-25K
AT-xl	60.6	84	23	Help-50K
AT-xl	71.2	86	19	Help-all

have a similar or even higher safety rate than their text-only counterparts (fine-tuned for safety on the text-only portion of Help-all dataset). This shows that the additional speech modality did not degrade the original safety offered by the text-only models. We observe that for all methods, the models with the larger LLM (-xl) performs better than the smaller model (-l). This highlights the benefit of a larger model in improving the safety. Furthermore, the proposed methods (AT) perform better than the baselines (Base and ST) of corresponding model sizes. This shows that the adversarial training technique did not introduce any unnecessary performance regressions in the absence of adversarial attacks, and even improves the safety.

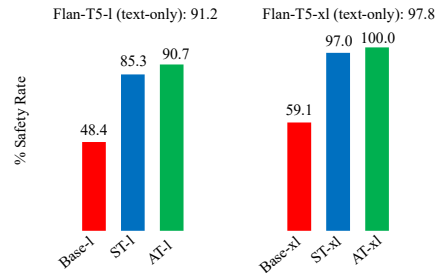


Figure 4: %Safety rate (\uparrow) of the different models without any attack. This shows that the proposed methods do not introduce any undesired degradation of safety on data in the absence of adversarial perturbations.

5. Conclusion

We proposed adversarial training as a defense mechanism against adversarial jailbreak threats on speech LLMs. We used synthesized speech samples from a collection of instruction-tuning and safety-alignment text datasets to train our models. We performed extensive experiments with different attack configurations and showed that the proposed methods significantly outperform vanilla safety alignment techniques in all considered threat scenarios. Furthermore, we showed that our methods increase the attacker's budget of iterations to produce successful jailbreaks, further highlighting their effectiveness. These methods can also be combined with other well-known countermeasures with provable safety guarantees to further improve robustness. This paper thus serves as a foundation for future studies in speech-based adversarial robustness, offering insights on data generation, model pretraining, and practical deployment.

6. References

- [1] X. Qi, K. Huang, A. Panda, P. Henderson, M. Wang, and P. Mittal, "Visual adversarial examples jailbreak aligned large language models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 19, 2024, pp. 21 527–21 536.
- [2] R. Peri, S. M. Jayanthi, S. Ronanki, A. Bhatia, K. Mundnich, S. Dingliwal, N. Das, Z. Hou, G. Huybrechts, S. Vishnubhotla, D. Garcia-Romero, S. Srinivasan, K. J. Han, and K. Kirchhoff, "Speechguard: Exploring the adversarial robustness of multimodal large language models," 2024. [Online]. Available: <https://arxiv.org/abs/2405.08317>
- [3] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," 2019. [Online]. Available: <https://arxiv.org/abs/1706.06083>
- [4] L. Yu, V. Do, K. Hamardzumyan, and N. Cancedda, "Robust llm safeguarding via refusal feature adversarial training," *arXiv preprint arXiv:2409.20089*, 2024.
- [5] S. Hua, S. Jin, and S. Jiang, "The limitations and ethical considerations of chatgpt," *Data intelligence*, vol. 6, no. 1, pp. 201–239, 2024.
- [6] A. Kumar, S. Singh, S. V. Murty, and S. Ragupathy, "The ethics of interaction: Mitigating security threats in llms," *arXiv preprint arXiv:2401.12273*, 2024.
- [7] A. Askill, Y. Bai, A. Chen, D. Drain, D. Ganguli, T. Henighan, A. Jones, N. Joseph, B. Mann, N. DasSarma, N. Elhage, Z. Hatfield-Dodds, D. Hernandez, J. Kernion, K. Ndousse, C. Olsson, D. Amodei, T. Brown, J. Clark, S. McCandlish, C. Olah, and J. Kaplan, "A general language assistant as a laboratory for alignment," 2021. [Online]. Available: <https://arxiv.org/abs/2112.00861>
- [8] X. Shen, Z. Chen, M. Backes, Y. Shen, and Y. Zhang, "'do anything now': Characterizing and evaluating in-the-wild jailbreak prompts on large language models," 2024. [Online]. Available: <https://arxiv.org/abs/2308.03825>
- [9] A. Wei, N. Haghtalab, and J. Steinhardt, "Jailbroken: How does llm safety training fail?" 2023. [Online]. Available: <https://arxiv.org/abs/2307.02483>
- [10] D. Hendrycks, C. Burns, S. Basart, A. Critch, J. Li, D. Song, and J. Steinhardt, "Aligning AI with shared human values," *CoRR*, vol. abs/2008.02275, 2020. [Online]. Available: <https://arxiv.org/abs/2008.02275>
- [11] Y. Liu, G. Deng, Y. Li, K. Wang, Z. Wang, X. Wang, T. Zhang, Y. Liu, H. Wang, Y. Zheng, and Y. Liu, "Prompt injection attack against llm-integrated applications," 2024. [Online]. Available: <https://arxiv.org/abs/2306.05499>
- [12] X. Liu, Z. Yu, Y. Zhang, N. Zhang, and C. Xiao, "Automatic and universal prompt injection attacks against large language models," 2024. [Online]. Available: <https://arxiv.org/abs/2403.04957>
- [13] A. Zou, Z. Wang, N. Carlini, M. Nasr, J. Z. Kolter, and M. Fredrikson, "Universal and transferable adversarial attacks on aligned language models," 2023. [Online]. Available: <https://arxiv.org/abs/2307.15043>
- [14] C. Guo, A. Sablayrolles, H. Jégou, and D. Kiela, "Gradient-based adversarial attacks against text transformers," 2021. [Online]. Available: <https://arxiv.org/abs/2104.13733>
- [15] X. Qi, K. Huang, A. Panda, P. Henderson, M. Wang, and P. Mittal, "Visual adversarial examples jailbreak aligned large language models," 2023. [Online]. Available: <https://arxiv.org/abs/2306.13213>
- [16] C. Guo, J. R. Gardner, Y. You, A. G. Wilson, and K. Q. Weinberger, "Simple black-box adversarial attacks," 2019. [Online]. Available: <https://arxiv.org/abs/1905.07121>
- [17] H. Jin, L. Hu, X. Li, P. Zhang, C. Chen, J. Zhuang, and H. Wang, "Jailbreakzoo: Survey, landscapes, and horizons in jailbreaking large language and vision-language models," *arXiv preprint arXiv:2407.01599*, 2024.
- [18] Y. Xie, J. Yi, J. Shao, J. Curl, L. Lyu, Q. Chen, X. Xie, and F. Wu, "Defending chatgpt against jailbreak attack via self-reminders," *Nature Machine Intelligence*, vol. 5, no. 12, pp. 1486–1496, 2023.
- [19] Z. Wei, Y. Wang, and Y. Wang, "Jailbreak and guard aligned language models with only few in-context demonstrations," *arXiv preprint arXiv:2310.06387*, 2023.
- [20] T. Chakraborty, E. Shayegani, Z. Cai, N. Abu-Ghazaleh, M. S. Asif, Y. Dong, A. K. Roy-Chowdhury, and C. Song, "Cross-modal safety alignment: Is textual unlearning all you need?" *arXiv preprint arXiv:2406.02575*, 2024.
- [21] R. Rade and S.-M. Moosavi-Dezfooli, "Helper-based adversarial training: Reducing excessive margin to achieve a better accuracy vs. robustness trade-off," in *ICML 2021 Workshop on Adversarial Machine Learning*, 2021.
- [22] Y. Mo, Y. Wang, Z. Wei, and Y. Wang, "Studious bob fight back against jailbreaking via prompt adversarial tuning," *arXiv preprint arXiv:2402.06255*, 2024.
- [23] A. Zhou, B. Li, and H. Wang, "Robust prompt optimization for defending language models against jailbreaking attacks," *arXiv preprint arXiv:2401.17263*, 2024.
- [24] C. Zheng, F. Yin, H. Zhou, F. Meng, J. Zhou, K.-W. Chang, M. Huang, and N. Peng, "On prompt-driven safeguarding for large language models," in *Forty-first International Conference on Machine Learning*, 2024.
- [25] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma *et al.*, "Scaling instruction-finetuned language models," *Journal of Machine Learning Research*, vol. 25, no. 70, pp. 1–53, 2024.
- [26] T. Sun, X. Zhang, Z. He, P. Li, Q. Cheng, X. Liu, H. Yan, Y. Shao, Q. Tang, S. Zhang, X. Zhao, K. Chen, Y. Zheng, Z. Zhou, R. Li, J. Zhan, Y. Zhou, L. Li, X. Yang, L. Wu, Z. Yin, X. Huang, Y.-G. Jiang, and X. Qiu, "Moss: An open conversational large language model," *Machine Intelligence Research*, vol. 21, no. 5, pp. 888–905, 2024.
- [27] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, and T. B. Hashimoto, "Stanford alpaca: An instruction-following llama model," https://github.com/tatsu-lab/stanford_alpaca, 2023.
- [28] M. Mazeika, L. Phan, X. Yin, A. Zou, Z. Wang, N. Mu, E. Sakhaee, N. Li, S. Basart, B. Li, D. Forsyth, and D. Hendrycks, "Harmbench: A standardized evaluation framework for automated red teaming and robust refusal," 2024. [Online]. Available: <https://arxiv.org/abs/2402.04249>
- [29] Z. Zhang, L. Lei, L. Wu, R. Sun, Y. Huang, C. Long, X. Liu, X. Lei, J. Tang, and M. Huang, "Safetybench: Evaluating the safety of large language models," 2024. [Online]. Available: <https://arxiv.org/abs/2309.07045>
- [30] P. Chao, E. Debenedetti, A. Robey, M. Andriushchenko, F. Croce, V. Sehwag, E. Dobriban, N. Flammarion, G. J. Pappas, F. Tramèr, H. Hassani, and E. Wong, "Jailbreakbench: An open robustness benchmark for jailbreaking large language models," 2024.
- [31] E. Casanova, K. Davis, E. Gölge, G. Gökner, I. Gulea, L. Hart, A. Aljafari, J. Meyer, R. Morais, S. Olayemi *et al.*, "Xtts: a massively multilingual zero-shot text-to-speech model," *arXiv preprint arXiv:2406.04904*, 2024.
- [32] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.
- [33] N. Carlini, A. Athalye, N. Papernot, W. Brendel, J. Rauber, D. Tsipras, I. Goodfellow, A. Madry, and A. Kurakin, "On evaluating adversarial robustness," *arXiv preprint arXiv:1902.06705*, 2019.
- [34] P. Chao, A. Robey, E. Dobriban, H. Hassani, G. J. Pappas, and E. Wong, "Jailbreaking black box large language models in twenty queries," *arXiv preprint arXiv:2310.08419*, 2023.