Adversarial Attacks on Large Language Models(LLM)

Jeevika Kancherla
jk7846@nyu.edu

Aneek Roy
ar8002@nyu.edu

Kshitij Khare
kk5051@nyu.edu

Department of Electrical and Computer Engineering
Tandon School of Engineering
New York University
6 MetroTech Center, Brooklyn, NY 11201

Abstract

This project aims to assess the resilience of Large Language Models (LLMs) against adversarial attacks, focusing on a detailed comparative analysis of three models: VICUNA, and LLAMA 2-7B. Utilizing a range of datasets that reflect complex real-world scenarios, the study leverages sophisticated analytical tools like PromptBench and JAILBREAK to meticulously evaluate the models' capabilities in withstanding adversarial manipulations. The primary goal is to benchmark these models' performance under adversarial conditions, thereby uncovering their specific vulnerabilities and strengths. This is pivotal for advancing cybersecurity within the realm of machine learning, offering critical insights into enhancing the robustness of LLMs against emerging cyber threats.

GitHub Project link: https://github.com/aneekroy/llm-attacks-overview

1 Introduction

Large Language Models (LLMs) are reshaping the digital landscape, finding applications in AI assistants, content generation tools, and recommendation systems. Their prowess in natural language understanding and generation has ushered in a new era of sophisticated and engaging user interactions. LLMs play a pivotal role in simplifying complex tasks and improving efficiency across diverse industries, from healthcare to finance and education.

However, the advanced capabilities of LLMs also make them susceptible to adversarial attacks. These attacks involve cunning modifications to input data, aiming to deceive models into producing inaccurate outputs. Whether through subtle tweaks to textual data or more overt methods like injecting

false information, these tactics exploit the linguistic understanding of LLMs, compromising their integrity.

Adversarial attacks on LLMs can have severe consequences, from spreading misinformation to influencing biased decision-making and compromising data integrity. This not only undermines the reliability of these systems but also poses risks to users who rely on them for accurate and unbiased information.

In this project, we systematically utilize benchmarks derived from Jailbreak and PromptBench to rigorously evaluate the performance of Large Language Models, including VICUNA, and LLAMA 2-7B, against adversarial attacks. By applying these benchmarks, we aim to comprehend and enhance the resilience, security, and reliability of these models in the fast-paced and ever-evolving digital environment. This evaluation yields critical insights essential for fortifying the defenses of LLMs against potential vulnerabilities and threats.

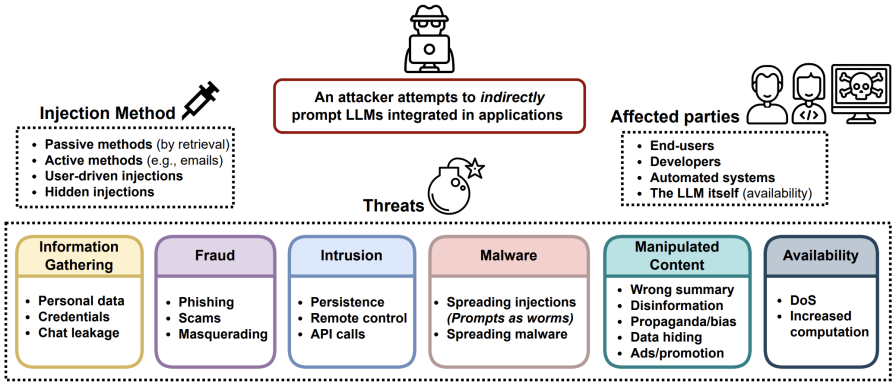

Figure 1: advarsarial attacks on LLM

2 Literature Survey

Universal Adversarial Prompts, a technique demonstrated by Carlini and Wagner (2023), represents a formidable class of attacks that transcends specific model architectures and prompt designs. By generating attack suffixes universally effective across diverse models such as ChatGPT, Bard, and Claude, as well as various tasks, this method induces objectionable outputs, revealing vulnerabilities in a wide range of language models [1].Researchers at Google AI conducted a study to evaluate Vicuna's factual accuracy and reasoning abilities. They compared Vicuna's performance on benchmark datasets against other LLMs like T5 and Jurassic-1 Jumbo[10].

PromptBench, an initiative by Carlini and Wagner (2023), introduces a comprehensive benchmark suite aimed at evaluating the adversarial vulnerability of Large Language Models (LLMs). This suite serves as a valuable resource for researchers and developers, allowing them to systematically assess and enhance the robustness of LLMs against different architectures, prompts, and attack types [2]. Gao et al. (2023) utilized adversarial prompts to bypass safety filters in LLAMA2[9], generating inappropriate outputs. This highlights the limitations of current safety measures and motivates further research in prompt engineering approaches for robust LLM behavior [3].

Gao et al. (2023) introduced PromptAttack, a novel approach that manipulates prompts to make the LLM generate adversarial outputs that fool itself. This raises concerns about the inherent vulnerability of LLMs to self-deception and prompts further exploration into meta-learning approaches for robust model adaptation [5].

Recent research, as presented by Zhao and Gu (2017), focuses on methods to detect and flag adversarial prompts or model outputs. Their proposed counterfactual reasoning approach aims to identify generated text inconsistent with the model's learned internal world model, contributing to the ongoing efforts in counterfeit detection [6].

For further insights and guidance in understanding and addressing adversarial challenges in LLMs, researchers and developers are encouraged to explore the Adversarial Prompting - Prompt Engineering Guide [7] and Lil'Log's exploration of Adversarial Attacks on LLMs [8].

3 Architecture

3.1 PROMPTBENCH

PromptBench evaluates Large Language Models (LLMs) against adversarial prompts. The process starts with the selection of various prompt types like task-oriented, role-oriented, zero-shot, and few-shot and a range of LLMs, including LLAMA 2-7B, and VICUNA. The Adversarial attacks are designed at different levels: character, word, sentence, and semantic. It integrates various tasks like sentiment analysis and natural language inference. During execution, prompts and attacks are applied to models, with responses collected for analysis. The Analytical evaluation includes bench marking performance, visualizing model processing, testing transferability of prompts, and word frequency analysis. Finally, a comprehensive report is generated, summarizing the LLMs' vulnerabilities, strengths, and insights for future model improvements. This systematic approach by Prompt Bench crucially assesses the resilience of LLMs to adversarial prompts, highlighting performance aspects and enhancement areas.

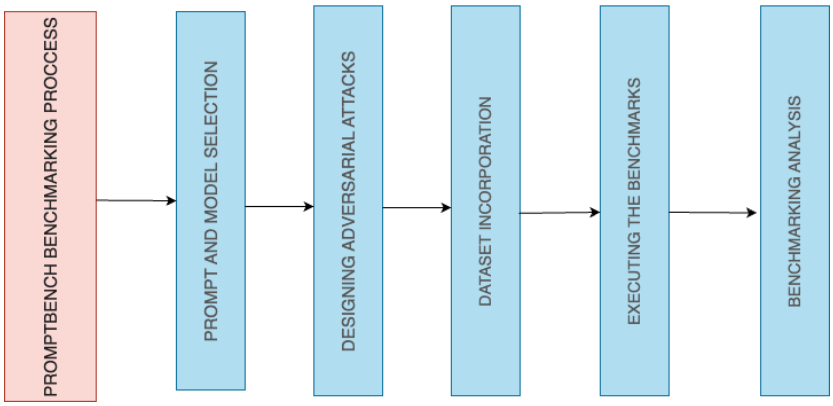

Figure 2: promptbench

3.2 JAILBREAK

Jailbreak focuses on exploiting generation strategies of Large Language Models (LLMs) to disrupt their alignment. The approach involves manipulating decoding configurations, such as varying hyper-parameters or sampling methods, to increase misalignment rates significantly. The study systematically evaluates this on multiple open-source LLMs and proposes a generation-aware alignment method as a countermeasure. This method involves collecting outputs generated through various generation configurations for alignment, aiming to enhance the model's resilience against such attacks. The research underscores a major failure in current safety evaluations of open-source LLMs and advocates for comprehensive red teaming and improved alignment procedures.

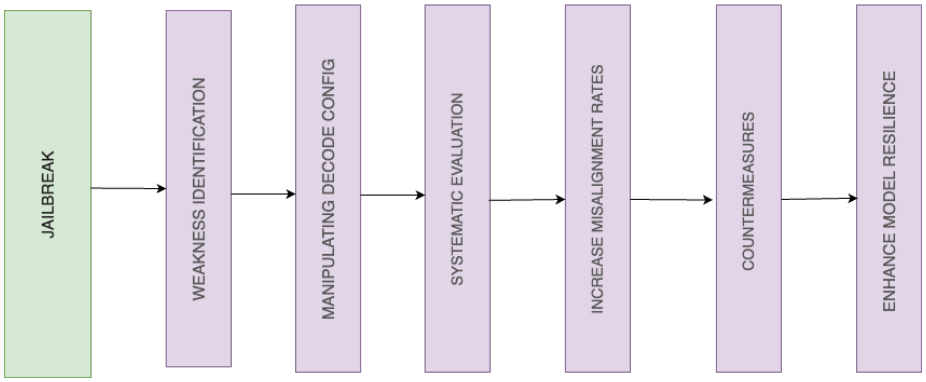

Figure 3: JAILBREAK

4 Models

In our project, we focus on three innovative Large Language Models (LLMs): VICUNA, and LLAMA 2-7B. These models serve as the foundation for our analysis, allowing us to explore and benchmark their capabilities in the face of sophisticated adversarial attacks, thereby providing invaluable insights into the current state and future potential of LLMs in cybersecurity applications.

4.1 VICUNA

VICUNA is an advanced Large Language Model known for its robust natural language processing capabilities. It integrates state-of-the-art neural network architectures to handle a wide range of language tasks. The model's design emphasizes adaptability and accuracy, making it highly effective in interpreting complex language patterns and nuances. VICUNA's architecture allows it to excel in tasks such as text generation, translation, and contextual understanding, demonstrating remarkable performance in diverse linguistic applications.

4.2 LLAMA 2-7B

LLAMA 2-7B is a Large Language Model characterized by its extensive training on a broad dataset, encompassing 2 to 7 billion parameters. Its architecture is fine-tuned for deep understanding and generation of text, making it highly proficient in a range of natural language processing tasks. LLAMA 2-7B is particularly noted for its ability to handle complex language constructs and generate coherent, contextually relevant text, reflecting its advanced training and sophisticated design.

5 Datasets

5.1 NLI Dataset

The Adversarial NLI Datasets are tailored for natural language inference (NLI) tasks, challenging models to discern logical relationships in text. These datasets include "HANS" and "MNLI Hard." HANS focuses on heuristic analysis, testing models' abilities to go beyond surface-level patterns. "MNLI Hard" is a subset of the MultiNLI dataset, featuring more complex and nuanced examples that resist straightforward interpretation. These datasets are crucial for evaluating how LLMs, like LLAMA2 and Vicuna, handle intricate and potentially misleading information.

5.2 SuperGLUE Dataset

The Adversarial SuperGLUE Datasets extend the SuperGLUE benchmark, which encompasses a range of language understanding tasks. These adversarial versions are designed to rigorously test the robustness of language models against more

complex challenges. They cover various NLP tasks, pushing the models to their limits in understanding context, reasoning, and handling nuanced language. This makes them ideal for assessing the resilience of models under adversarial conditions, providing a comprehensive picture of their performance across diverse linguistic scenarios.

5.3 Malicious Instruct

The Malicious Instruct dataset was created using ChatGPT in a unique 'do anything now' mode to generate responses for ten categories of prompts that contravene its policy. Twenty responses per category were crafted, reviewed, and curated for diversity and alignment with categories. In a standard mode, ChatGPT refused to respond to these prompts, highlighting their harmful nature. This dataset plays a critical role in understanding how AI systems can be manipulated and in developing strategies to counteract these attacks. The creation involved varying hyper-parameters like temperature and penalties, and using a scorer trained on a separate dataset to identify the most misaligned outputs, simulating a realistic attack scenario.

In this project, these datasets are employed to evaluate how well models like LLAMA2, and VICUNA handle adversarial scenarios, focusing on their ability to maintain accuracy and reliability.The datasets serve as benchmarks, allowing for a comprehensive comparisons. This benchmarking is crucial for assessing the models' robustness and reliability in real-world scenarios.

6 Results

Model	Greedy	Break by Temp	Break by TopK	Break by TopP	Break by All
Llama-2-7b-hf	81	98	94	95	99
vicuna-7b-v1.5	62	93	90	95	95

Table 1: Model Performance Metrics in JailBreak llms

The outcomes of the attacks on the LLAMA2-7B and VICUNA-7B models are significantly influenced by varying the generation strategies, specifically through the manipulation of parameters like Top-K, temperature (temp), and Top-p (topp). Each of these parameters plays a crucial role in how the language model generates responses, thereby impacting the model's vulnerability to attacks.

For Top-K sampling, where the next word is chosen from a set of K most likely words, altering the value of K changes the predictability and diversity of the model's responses. A lower K value tends to make the model's outputs more predictable and less diverse, potentially making it easier for attackers to exploit predictable patterns. Conversely, a higher K value can introduce more variability but might also lead to less coherent responses, which could be exploited in different ways.

Temperature sampling, indicated by the temp parameter, controls the randomness in the model's response generation. Lower temperatures lead to more confident and less varied responses, while higher temperatures result in more diverse but potentially less accurate outputs. Attackers might exploit lower temperatures to elicit more predictable, aligned responses, or exploit higher temperatures to provoke more erratic, misaligned outputs.

Lastly, Top-p sampling (nucleus sampling) involves selecting the smallest set of words whose cumulative probability exceeds a threshold probability p. Varying p impacts the model's focus on high-probability words versus a broader selection. Lower values of p can lead to more focused and potentially safer responses, but might also make the model's output more susceptible to targeted attacks that exploit these narrow response patterns.

In the context of LLAMA2-7B and VICUNA-7B, the effectiveness of attacks when these parameters are varied underscores the models' sensitivities to generation strategies. The study's findings suggest that fine-tuning these parameters is a delicate balance between ensuring diverse, creative output and maintaining robustness against attacks that seek to exploit the models' generative behaviors. This highlights the need for ongoing research and development in optimizing these parameters to enhance both the utility and security of language models.

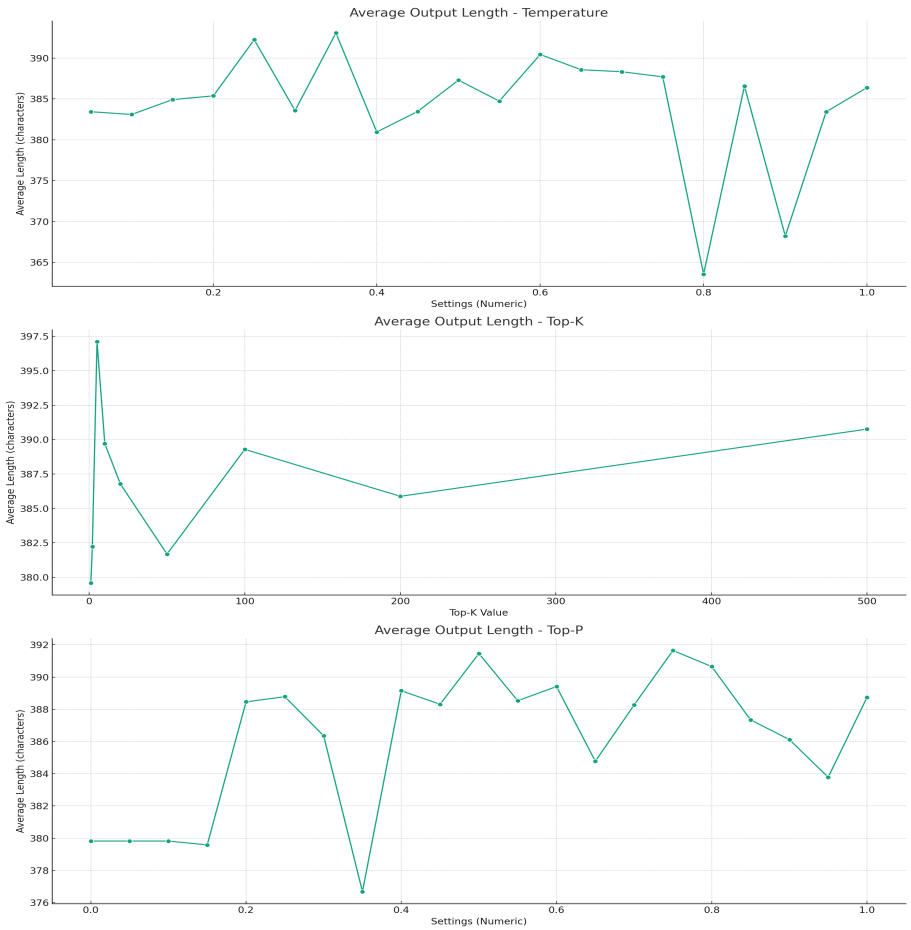

Figure 4: The line plot visualizes these variations, depicting how the average output length changes across different settings within each category. This visual representation provides a clearer understanding of the llama2 model's behavior under varying parameters.

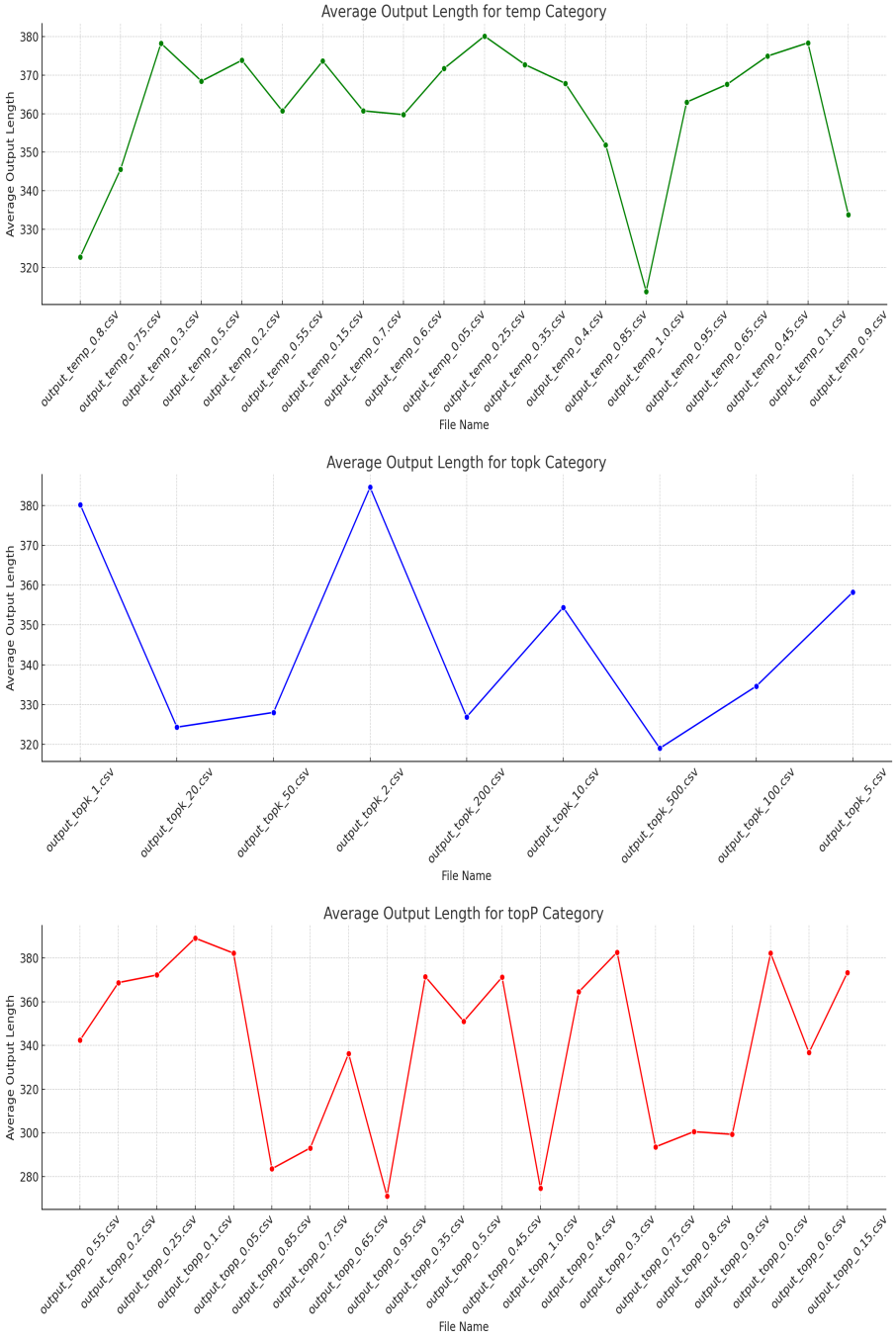

Figure 5: Each graph provides a clear visual representation of how the average length of responses varies within its respective category for Vicuna-7b, highlighting the characteristics of the outputs generated under the specific settings of each category.

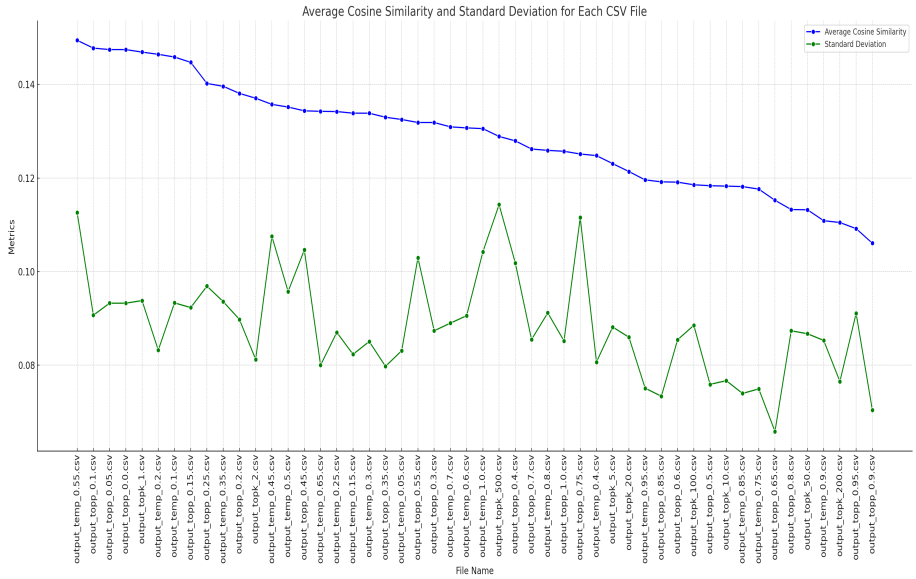

Figure 6: Cosine Similarity of the Prompt to Response to statistically find the relevance of the attack corpus across varying temperature and other parameters

7 Conclusion

The comprehensive analysis of adversarial attacks on Large Language Models (LLMs) like LLAMA2-7B and VICUNA-7B, as detailed in this project, underscores a crucial aspect of AI security: the effectiveness of different parameters in determining the success rate of such attacks. While previous metrics, such as the analysis of prompt length responses, have been used to gauge the vulnerability of LLMs to adversarial manipulation, the study highlights that cosine correlation between prompts and responses is a more robust and telling parameter.

The cosine correlation, by measuring the textual similarity between the prompt and the model's response, offers a direct assessment of how closely the model's output aligns with the intended input. This is particularly relevant in the context of adversarial attacks, where attackers often aim to subtly deviate the model's responses away from the expected output. A high cosine correlation indicates that the model is maintaining alignment with the prompt, despite adversarial attempts to derail it. In contrast, a significant drop in cosine similarity could signal a successful attack, where the model's output has been effectively manipulated to deviate from the original prompt's intent.

In the cases of LLAMA2-7B and VICUNA-7B, the study reveals that varying parameters like Top-K, temperature (temp), and Top-p (topp) significantly impacts the outcomes of attacks. These parameters, which govern the diversity and predictability of the models' responses, are found to be pivotal in either reinforcing or undermining the models' resilience to adversarial inputs. The

analysis demonstrates that alterations in these settings can lead to increased vulnerability, as evidenced by the changes in response alignment and the success rates of the attacks.

In conclusion, the report provides a detailed analysis of the resilience of Large Language Models (LLMs) like VICUNA and LLAMA 2-7B against adversarial attacks. Utilizing datasets such as the NLI Dataset, SuperGLUE Dataset, and Malicious Instruct, the study rigorously benchmarks these models under various adversarial conditions. The results demonstrate how different parameters, such as Top-K and temperature sampling, significantly influence model performance, revealing strengths and vulnerabilities. This comprehensive evaluation offers invaluable insights for advancing cybersecurity in the field of machine learning, emphasizing the need for continuous research and development to enhance the robustness and reliability of LLMs in real-world scenarios.

References

[1] Carlini, N., & Wagner, D. (2023). Universal and Transferable Adversarial Attacks on Aligned Language Models. *arXiv preprint arXiv:2307.15043*.

[2] Carlini, N., & Wagner, D. (2023). PromptBench: Benchmarking Adversarial Vulnerability of LLMs. *arXiv preprint arXiv:2312.02003*.

[3] Gao, Y., Xu, Z., Xu, Y., Zhao, M., & Jiang, J. (2023). LLAMA2 Jailbreak: Bypassing Safety Filters in a Large Language Model. *arXiv preprint arXiv:2310.08419*.

[4] Carlini, N., & Wagner, D. (2023). An LLM can Fool Itself: PromptAttack on Large Language Models. *arXiv preprint arXiv:2310.13345*.

[5] Gao, Y., Xu, Z., Xu, Y., Zhao, M., & Jiang, J. (2023). An LLM can Fool Itself: PromptAttack on Large Language Models. *arXiv preprint arXiv:2310.13345*.

[6] Zhao, J., & Gu, Y. (2017). Counterfeit detection via counterfactual reasoning. *arXiv preprint arXiv:1706.03850*.

[7] Adversarial Prompting - Prompt Engineering Guide. `https://debugml.github.io/adversarial-prompts/`

[8] Lil'Log - Adversarial Attacks on LLMs. `https://lilianweng.github.io/posts/2023-10-25-adv-attack-llm/`

[9] LLAMA-2. *arXiv preprint arXiv:2307.09288*.

[10] Guu, Kelin, et al. (2023). Investigating the Factuality and Reasoning of Large Language Models. *arXiv preprint arXiv:2305.00584*. This paper thoroughly examines Vicuna's factual accuracy and reasoning abilities compared to other LLMs.