Takeshi Goto\textsuperscript{1}, Kensuke Ono\textsuperscript{1}, and Akira Morita\textsuperscript{1}

\textsuperscript{1}Affiliation not available

March 06, 2024
Evaluating the Cybersecurity Robustness of Commercial LLMs against Adversarial Prompts: A PromptBench Analysis

Takeshi Goto, Kensuke Ono, and Akira Morita

Abstract—This study presents a comprehensive evaluation of the cybersecurity robustness of five leading Large Language Models (LLMs) - ChatGPT-4, Google Gemini, Anthropic Claude, Meta Llama, and Mistral 8x7B - against adversarial prompts using the PromptBench benchmark. Through a dual approach of quantitative and qualitative analysis, the research explores each model’s performance, resilience, and vulnerabilities. Quantitative metrics such as accuracy, precision, recall, and F1 scores offer a statistical comparison across models, while qualitative insights reveal distinct patterns of response and susceptibility to various adversarial strategies. The findings highlight significant variations in model robustness, underlining the importance of a complex approach to enhancing LLM security. This study not only sheds light on current limitations but also emphasizes the need for advancing evaluation methodologies and model development practices to mitigate potential threats and ensure the safe deployment of LLMs in sensitive and critical applications.

Index Terms—Adversarial Attacks, Evaluation Benchmarks, Large Language Models, Model Robustness, Natural Language Processing, Security Vulnerabilities

I. INTRODUCTION

Recent developments in Large Language Models (LLMs) have ushered in transformative changes across numerous domains, from natural language processing to complex decision-making systems [1], [2]. Such advancements have not only showcased the functional superiority of LLMs but have also raised pertinent questions regarding their security and reliability [2]. Despite the exponential growth in their capabilities, the cybersecurity robustness of these models has received relatively less attention, a gap that this study aims to address. The robustness of LLMs against adversarial attacks is paramount for their safe and reliable deployment in sensitive and critical applications.

Adversarial attacks against LLMs represent a significant threat vector, exploiting model vulnerabilities to induce erroneous outputs or reveal sensitive information. Previous studies have predominantly focused on evaluating the functional performance of LLMs, with less emphasis on their resilience to such attacks [1], [3]–[5]. This oversight is concerning given the increasing integration of LLMs into security-sensitive environments. Current benchmarks, while extensive in scope, often fall short in systematically probing the cybersecurity defenses of these models. This discrepancy highlights the need for benchmarks like PromptBench, specifically designed to assess and enhance the security posture of LLMs.

PromptBench emerges as a pivotal benchmark in addressing the aforementioned challenges, offering a structured framework to evaluate the adversarial robustness of LLMs [6]. This study leverages PromptBench to conduct a comprehensive analysis of five commercial LLMs, including ChatGPT-4, Google Gemini, Anthropic Claude, Meta Llama, and the open-source “Mistral 8x7B” model. By employing PromptBench’s diverse set of adversarial prompts, this research aims to uncover the varying degrees of cybersecurity resilience across these models. The inclusion of “Mistral 8x7B”, a model structured around a Mixture of Experts (MoE) with 8 “experts” holding 7 billion parameters each, adds a unique dimension to our analysis, potentially offering insights into the security implications of MoE architectures in LLMs.

This study makes the following major contributions:

1) It provides a comprehensive evaluation of five leading Large Language Models (LLMs) - ChatGPT-4, Google Gemini, Anthropic Claude, Meta Llama, and Mistral 8x7B - using the PromptBench benchmark to assess their robustness against adversarial prompts.

2) It introduces a dual analytical framework combining quantitative metrics such as accuracy, precision, recall, and F1 score with qualitative analysis to uncover models’ response patterns and vulnerabilities.

3) It highlights the importance of integrating cybersecurity considerations throughout the LLM development cycle to mitigate vulnerabilities exposed by adversarial prompts.

4) It suggests directions for future research, emphasizing the development of more sophisticated adversarial prompts and robust evaluation methods to enhance LLM security and reliability.

The rest of this study is organized as follows: Section 2 explores previous studies focused on the robustness of LLMs, types of adversarial attacks targeting these models, and evaluation benchmarks designed to assess their vulnerabilities and performance. Section 3 outlines the methodological framework adopted for leveraging PromptBench to evaluate the robustness of various LLMs against adversarial prompts. Section 4 presents the findings from the evaluation of LLMs using the PromptBench benchmark, detailing both quantitative and qualitative analyses to provide a comprehensive understanding of the models’ performance and robustness against adversarial prompts. Section 5 looks into the critical aspects of the study, exploring the implications, limitations, and future directions without merely echoing the results. Section 6 summarizes the main findings of the study, highlighting their significance and
implications for the field of LLMs and cybersecurity.

II. RELATED WORK

This section explores previous studies focused on the robustness of LLMs, types of adversarial attacks targeting these models, and evaluation benchmarks designed to assess their vulnerabilities and performance.

A. Adversarial Attacks on LLMs

Several studies have highlighted different facets of adversarial attacks targeting LLMs, underscoring the spectrum of vulnerabilities these models face. Some studies identified a method for creating textual inputs that lead LLMs to generate incorrect or biased responses, emphasizing the threat to model integrity [7]–[11]. Others explored how slight, imperceptible modifications to input data can drastically alter model outputs, revealing the sensitivity of LLMs to input manipulation [12]–[15]. Further investigations demonstrated the potential for adversarial examples to extract private information from models, thus raising significant privacy concerns [3], [7], [16]–[18]. The development of adversarial attacks that exploit the model’s reliance on contextual cues to induce misclassifications points to critical weaknesses in their understanding capabilities [19]–[22]. Research into the use of syntactically correct but semantically misleading prompts uncovered further avenues for misleading LLMs [9], [23]–[26]. The effectiveness of adversarial attacks in bypassing content filters designed to prevent the generation of harmful outputs was also documented, highlighting challenges in ensuring safe outputs [3], [7], [27]. Lastly, studies have shown that even state-of-the-art LLMs are not immune to carefully designed adversarial inputs, underscoring the ongoing arms race between model development and adversarial attack techniques [7], [28], [29].

B. Benchmarks for LLM Evaluation

The literature reveals a rich ecosystem of benchmarks designed to test various aspects of LLM performance and robustness. Some focused on measuring the general linguistic and reasoning abilities of LLMs, providing a comprehensive assessment of their capabilities [30]–[32]. Other benchmarks were developed to specifically evaluate models’ performance on tasks requiring deep semantic understanding, testing the limits of LLM comprehension [1], [32]–[35]. The introduction of datasets aimed at probing the ethical reasoning and bias of LLMs marked an important step towards more responsible AI [33], [36], [37]. Research into benchmarks assessing the factual accuracy of LLM outputs in generating knowledge-based responses was crucial in highlighting the challenges of ensuring information reliability [38], [39]. The creation of PromptBench, with its focus on evaluating LLMs’ resilience to adversarial prompts, fills a critical gap in the evaluation landscape, offering insights into models’ security robustness [6]. Comparative studies leveraging PromptBench have illuminated significant variations in how different LLMs handle security challenges, reinforcing the importance of specialized benchmarks in the development of more secure AI systems [6]. Furthermore, the continuous evolution of benchmarks to include novel adversarial techniques reflects the dynamic nature of the field, underscoring the need for ongoing research and model assessment [33], [39], [40]–[42].

C. Robustness and Security in LLMs

Research on enhancing the robustness and security of LLMs against adversarial attacks has yielded various strategies, including the development of defensive mechanisms to detect and mitigate such attacks [43], [44]. Investigations into the use of adversarial training to improve model resilience have shown promise in reducing vulnerability to crafted inputs [17], [28], [45], [46]. The exploration of encryption techniques for safeguarding model data against extraction attacks has highlighted the potential for stronger security protocols [4], [5], [7], [47], [48]. Efforts to design LLMs with inherent resistance to adversarial manipulation suggest a proactive approach to model security [3], [7], [15], [24], [49]. Studies focusing on the audit and certification of LLMs for security vulnerabilities offer a pathway to more trustworthy AI deployment [4], [9], [50]–[52]. The impact of dataset sanitization on improving model robustness against attacks derived from biased or malicious inputs has been thoroughly evaluated [47], [53], [54]. Lastly, the exploration of transfer learning as a means to enhance LLM security without compromising performance provides a balanced approach to model development [55], [56].

D. Ethical Considerations in LLM Deployment

The deployment of LLMs raises substantial ethical considerations, particularly in relation to the potential for adversarial attacks to exacerbate biases or enable misinformation [8], [9], [24]. Research has been conducted on the ethical frameworks necessary for responsible LLM deployment, emphasizing the importance of transparency and accountability [55], [57]–[59]. Studies investigating the propagation of bias through adversarial attacks have called for more rigorous bias mitigation strategies in LLM development [24], [47], [55], [60], [61]. The risk of misinformation being amplified by LLMs susceptible to adversarial inputs has prompted discussions on the need for robust fact-checking mechanisms [9], [47]. Ethical guidelines proposed for the deployment of LLMs in sensitive applications seek to balance innovation with societal impact [62], [63]. The exploration of user privacy in the context of LLMs highlights the need for privacy-preserving technologies in model design [18], [64]. Finally, the consideration of LLMs’ environmental impact has been mentioned, particularly about its extensive electricity and hardware usage [23], [55], [65].

III. METHODOLOGY

This section outlines the methodological framework adopted for leveraging PromptBench to evaluate the robustness of various Large Language Models (LLMs) against adversarial prompts. It details the selection of LLMs, the characteristics of the PromptBench benchmark, and the metrics used to assess model performance and robustness.
A. LLMs Under Study

The LLMs selected for evaluation in this study represent a cross-section of the current landscape of commercial and open-source models. To facilitate a detailed comparison, Table I presents the distinctive features of each model.

Each model was chosen for its distinct approach to language modeling, providing a comprehensive view of the field’s current capabilities and challenges. ChatGPT-4, renowned for its conversational prowess, offers an advanced understanding of dialogue and sophisticated human interaction, making it a leading candidate for applications requiring high-quality, natural language generation. Google Gemini, leveraging Google’s vast data resources, stands out for its contextual comprehension and sophisticated text understanding, indicative of Google’s long-standing expertise in search and information retrieval. Anthropic’s Claude emphasizes ethical AI development, integrating safety features designed to mitigate harmful outputs, which underscores the increasing importance of responsible AI in technology deployment. Meta Llama showcases Meta AI’s focus on creating versatile models capable of understanding and generating text in multiple languages, reflecting the global nature of digital communication. Finally, Mistral 8x7B represents a significant stride in open-source AI, with its MoE architecture offering a novel approach to scalability and specialization in LLMs.

The diverse architectures and unique features of these models form the basis for a comprehensive evaluation of their robustness against adversarial prompts, highlighting their respective strengths and areas for improvement in the context of cybersecurity.

B. PromptBench Benchmark

PromptBench is a comprehensive benchmark designed to evaluate the robustness of LLMs against a variety of adversarial prompts. It encompasses a wide array of prompts that challenge models on multiple fronts, including semantic understanding, logic reasoning, and resistance to misleading or ambiguous inputs. The benchmark’s design reflects the need to simulate realistic scenarios where models might be exploited to generate harmful or incorrect outputs. By systematically presenting these models with adversarial prompts, PromptBench aims to uncover vulnerabilities and assess how well each LLM maintains its performance integrity under potentially malicious usage.

To provide a clearer understanding of PromptBench’s features and its applicability to this study, Table II details its key characteristics.

As illustrated, PromptBench’s structured approach to evaluating adversarial robustness makes it an ideal tool for this study’s goal of assessing the cybersecurity resilience of leading LLMs. Its comprehensive methodology not only sheds light on each model’s vulnerabilities but also provides a benchmark against which improvements in model design and training can be measured. This rigorous evaluation framework ensures that our study’s findings will contribute valuable insights into the current state of LLM security and guide future efforts to enhance the robustness of these models against adversarial threats.

C. Evaluation Metrics

The robustness of LLMs against adversarial prompts is critically assessed using a comprehensive set of quantitative and qualitative metrics. This multifaceted approach is designed to provide a holistic understanding of each model’s resilience and areas for improvement.

- **Accuracy**: Measures the proportion of correct responses to the total number of prompts. This metric is fundamental for assessing the basic performance of LLMs in generating contextually appropriate responses.
- **Precision**: Evaluates the ratio of correctly positive identified prompts to the total identified as positive. Precision is crucial for understanding how models manage to avoid false positives in adversarially crafted inputs.
- **Recall**: Assesses the ratio of correctly identified positive prompts to the total actual positive prompts. This metric highlights the models’ capability to capture all relevant adversarial challenges without missing any.
- **F1 Score**: Provides a balance between precision and recall, offering a single metric to evaluate the overall performance of the models against adversarial prompts. The F1 score is particularly useful for comparing model performance when there are disparities between precision and recall.
- **Response Consistency**: Measures the stability of model responses under varying adversarial conditions. This metric is indicative of the models’ resilience to adversarial inputs, reflecting their ability to maintain consistent performance.
- **Vulnerability Analysis**: A qualitative assessment of the models’ weaknesses when faced with specific types of adversarial prompts. This analysis offers deep insights into potential areas of improvement and the nature of the models’ security vulnerabilities.

Each metric has been carefully selected to shed light on different aspects of model performance and robustness. Together, they provide a comprehensive evaluation framework that not only assesses how well each LLM performs in the face of adversarial challenges but also identifies specific vulnerabilities and areas for improvement. This approach ensures a sophisticated understanding of LLM robustness, highlighting areas of strength and concern crucial for advancing the field of cybersecurity in LLM applications.

IV. RESULTS

This section presents the findings from the evaluation of LLMs using the PromptBench benchmark, detailing both quantitative and qualitative analyses to provide a comprehensive understanding of the models’ performance and robustness against adversarial prompts.

A. Quantitative Analysis

The quantitative analysis uncovers significant variations in the performance of the evaluated LLMs, as shown in the
Developing Entity Transformer-based Employs a quantitative approach to assessing robustness, using metrics such as accuracy, precision, and recall to evaluate Multilingual processing, diverse task handling Includes a diverse set of prompts designed to test various aspects of model performance, such as semantic understanding and Mixture of Experts (MoE) architecture, scalable problem-solving

Anthropic Advanced conversational abilities, extensive knowledge base

Google Transformer-based Particularly apt for evaluating the cybersecurity resilience of LLMs, offering insights into potential vulnerabilities and the Open Source Ethical considerations, safety features

Meta Llama Transformer-based Like Google Gemini, while excelling in contextual understanding, show reduced performance consistency in ethically ambiguous situations. The varying degrees of vulnerability across models to specific types of adversarial inputs underscore the importance of targeted model training and evaluation. Models such as Meta Llama exhibit significant variability in response to multilingual and culturally sophisticated prompts, suggesting a need for more diverse and comprehensive training datasets. Meanwhile, Mistral 8x7B’s specialized focus on technical domains highlights the trade-offs between domain specialization and general knowledge robustness.

Through this qualitative analysis, we gain valuable insights into the strengths and weaknesses of each LLM, informing future enhancements to model design, training methodologies, and adversarial prompt development strategies. The findings emphasize the necessity for a multifaceted approach to improving LLM robustness, balancing between enhancing technical proficiency and ensuring ethical and culturally aware responses. This comprehensive analysis, combining both quantitative metrics and qualitative insights, offers a sophisticated view of the current state of LLM robustness against adversarial attacks. The findings underscore the importance of ongoing improvements in model architecture, training methods, and evaluation benchmarks to enhance the security and reliability of LLMs in real-world applications.

V. Critical Discussion

This section looks into the critical aspects of the study, exploring the implications, limitations, and future directions without merely echoing the results. It aims to contextualize the findings within broader research and practical applications.

A. Model Vulnerabilities and Ethical Considerations

This study highlighted notable vulnerabilities in the evaluated LLMs, underscoring the ethical obligations of developers to address these weaknesses. The variance in model performance against adversarial prompts not only reflects technical challenges but also raises ethical questions regarding the deployment of LLMs in sensitive contexts. For instance, the propensity of some models to replicate biases or succumb to misleading information necessitates a more rigorous approach to training and evaluation. This responsibility extends beyond merely enhancing model accuracy; it encompasses the need to

<table>
<thead>
<tr>
<th>Model</th>
<th>Developing Entity</th>
<th>Unique Features</th>
<th>Architecture</th>
</tr>
</thead>
<tbody>
<tr>
<td>ChatGPT-4</td>
<td>OpenAI</td>
<td>Advanced conversational abilities, extensive knowledge base</td>
<td>Transformer-based</td>
</tr>
<tr>
<td>Google Gemini</td>
<td>Google</td>
<td>Sophisticated understanding of context and nuance</td>
<td>Transformer-based</td>
</tr>
<tr>
<td>Anthropic Claude</td>
<td>Anthropic</td>
<td>Ethical considerations, safety features</td>
<td>Transformer-based</td>
</tr>
<tr>
<td>Meta Llama</td>
<td>Meta AI</td>
<td>Multilingual processing, diverse task handling</td>
<td>Transformer-based</td>
</tr>
<tr>
<td>Mistral 8x7B</td>
<td>Open Source</td>
<td>Mixture of Experts (MoE) architecture, scalable problem-solving</td>
<td>MoE</td>
</tr>
</tbody>
</table>

TABLE II: Key Features of the PromptBench Benchmark

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Design Principles</td>
<td>Emphasizes real-world applicability by simulating scenarios likely to be encountered in deployment, focusing on adversarial robustness.</td>
</tr>
<tr>
<td>Adversarial Prompts</td>
<td>Includes a diverse set of prompts designed to test various aspects of model performance, such as semantic understanding and logic reasoning.</td>
</tr>
<tr>
<td>Methodology</td>
<td>Employs a quantitative approach to assessing robustness, using metrics such as accuracy, precision, and recall to evaluate performance integrity under adversarial conditions.</td>
</tr>
<tr>
<td>Suitability</td>
<td>Particularly apt for evaluating the cybersecurity resilience of LLMs, offering insights into potential vulnerabilities and the effectiveness of defense mechanisms.</td>
</tr>
</tbody>
</table>

This table illustrates the diverse approaches and capacities of each LLM to maintain performance integrity and respond adaptively to adversarial challenges. For example, Anthropic Claude’s design priorities on safety and ethical considerations are evident in its high response consistency and lower vulnerability to ethical dilemmas. Conversely, models like Google Gemini, while excelling in contextual understanding, show reduced performance consistency in ethically ambiguous situations. The varying degrees of vulnerability across models to specific types of adversarial inputs underscore the importance of targeted model training and evaluation. Models such as Meta Llama exhibit significant variability in response to multilingual and culturally sophisticated prompts, suggesting a need for more diverse and comprehensive training datasets. Meanwhile, Mistral 8x7B’s specialized focus on technical domains highlights the trade-offs between domain specialization and general knowledge robustness.

This complex performance landscape underscores the importance of a balanced approach to model development and evaluation, emphasizing not just a single metric but a comprehensive suite of metrics to gauge an LLM’s robustness against adversarial challenges. The variations in performance metrics across models highlight the distinct design and training methodologies that influence their ability to interpret and respond to complex and potentially deceptive inputs. Through this detailed quantitative analysis, we gain valuable insights into the models’ strengths and weaknesses, informing future directions for enhancing LLM security and reliability.

B. Qualitative Analysis

The qualitative analysis focused on the complex response patterns and failure modes of the evaluated LLMs against adversarial prompts. This analysis revealed the models’ differing capabilities to understand and counteract adversarially designed inputs, highlighting their resilience or vulnerability in various scenarios.

This table demonstrates the diverse approaches and capacities of each LLM to maintain performance integrity and respond adaptively to adversarial challenges. For example, Anthropic Claude’s design priorities on safety and ethical considerations are evident in its high response consistency and lower vulnerability to ethical dilemmas. Conversely, models like Google Gemini, while excelling in contextual understanding, show reduced performance consistency in ethically ambiguous situations. The varying degrees of vulnerability across models to specific types of adversarial inputs underscore the importance of targeted model training and evaluation. Models such as Meta Llama exhibit significant variability in response to multilingual and culturally sophisticated prompts, suggesting a need for more diverse and comprehensive training datasets. Meanwhile, Mistral 8x7B’s specialized focus on technical domains highlights the trade-offs between domain specialization and general knowledge robustness.

Through this qualitative analysis, we gain valuable insights into the strengths and weaknesses of each LLM, informing future enhancements to model design, training methodologies, and adversarial prompt development strategies. The findings emphasize the necessity for a multifaceted approach to improving LLM robustness, balancing between enhancing technical proficiency and ensuring ethical and culturally aware responses. This comprehensive analysis, combining both quantitative metrics and qualitative insights, offers a sophisticated view of the current state of LLM robustness against adversarial attacks. The findings underscore the importance of ongoing improvements in model architecture, training methods, and evaluation benchmarks to enhance the security and reliability of LLMs in real-world applications.

V. Critical Discussion

This section looks into the critical aspects of the study, exploring the implications, limitations, and future directions without merely echoing the results. It aims to contextualize the findings within broader research and practical applications.

A. Model Vulnerabilities and Ethical Considerations

This study highlighted notable vulnerabilities in the evaluated LLMs, underscoring the ethical obligations of developers to address these weaknesses. The variance in model performance against adversarial prompts not only reflects technical challenges but also raises ethical questions regarding the deployment of LLMs in sensitive contexts. For instance, the propensity of some models to replicate biases or succumb to misleading information necessitates a more rigorous approach to training and evaluation. This responsibility extends beyond merely enhancing model accuracy; it encompasses the need to
safeguard against the propagation of misinformation and the perpetuation of harmful stereotypes.

B. Implications for Model Training and Design

The insights garnered from this analysis offer substantial implications for model training and architectural design. The differential performance of models across various metrics suggests that a one-size-fits-all approach to LLM training is insufficient. Instead, tailored training strategies that account for specific adversarial vulnerabilities are crucial. This includes not only diversifying training datasets but also integrating adversarial examples directly into the training process. Moreover, the study underscores the potential benefits of exploring novel architectural innovations, such as the MoE approach employed by Mistral 8x7B, to enhance model robustness.

C. Future Directions in Adversarial Prompt Development

The development of adversarial prompts represents a rapidly evolving field, with this study contributing to its expansion. The complex understanding of model vulnerabilities uncovered here paves the way for more sophisticated adversarial prompt designs. Future research should focus on creating prompts that more closely mimic the subtleties of real-world misinformation tactics, thereby providing a more accurate gauge of LLM robustness. Additionally, there is a growing need for collaborative efforts to standardize adversarial prompt development, facilitating comparative evaluations across studies and models.

D. Limitations and Scope for Improvement

While this study provides valuable insights into the robustness of LLMs against adversarial prompts, it is not without its limitations. One notable constraint is the reliance on a predefined set of adversarial prompts, which may not fully encapsulate the breadth of potential real-world attacks. Furthermore, the qualitative nature of some evaluations introduces subjectivity into the analysis. Future studies should aim to broaden the scope of adversarial scenarios examined and explore more objective methods for qualitative assessment. Additionally, investigating the impact of continuous model training and updates on performance against adversarial prompts will be crucial for understanding the dynamic nature of LLM robustness.

VI. Conclusion

This study performed on a comprehensive evaluation of five leading LLMs, including ChatGPT-4, Google Gemini,
Anthropic Claude, Meta Llama, and Mistral 8x7B, using the PromptBench benchmark to assess their robustness against adversarial prompts. The quantitative and qualitative analyses revealed significant variations in the performance of these models, both in terms of metrics such as accuracy, precision, recall, and F1 score, and in their response consistency and vulnerability to specific adversarial tactics. These findings underscore the critical need for ongoing advancements in model design, training methodologies, and evaluation frameworks to enhance the security and reliability of LLMs in real-world applications.

Moreover, the study highlights the intricate balance between model performance and security. While some models demonstrated exceptional capability in certain areas, they exhibited vulnerabilities in others, suggesting that no model is universally superior across all aspects of cybersecurity resilience. This underscores the importance of a multifaceted approach to LLM development, where security considerations are integrated throughout the design and training process. Future research should focus on developing more sophisticated adversarial prompts and robust evaluation methods to further understand and mitigate the vulnerabilities of LLMs, ensuring their safe and ethical application across diverse domains.

REFERENCES


