Evaluation of Domain-Specific Prompt Engineering Attacks on Large Language Models

Charly Ashcroft[1] and Kahari Whitaker[1]

[1]Treatance

August 01, 2024

Abstract

The rapid integration of artificial intelligence into critical domains such as healthcare, finance, and legal services has necessitated a closer examination of the robustness and reliability of advanced language models. Adversarial prompt engineering presents a novel and significant method to systematically evaluate and exploit vulnerabilities within these models, highlighting the imperative for enhanced defensive strategies. A comprehensive evaluation was conducted on Claude and Gemini models, employing domain-specific adversarial prompts to test their performance across various sectors. The results indicated significant degradation in accuracy, reliability, and response time under adversarial conditions, revealing context-dependent vulnerabilities that compromise model integrity. Detailed statistical analyses and visualizations illustrated the substantial impact of adversarial inputs, providing robust evidence of the necessity for improved mitigation techniques. Patterns of susceptibility were identified, suggesting the need for tailored defensive approaches for different domains. The study contributes valuable insights into the inherent weaknesses of advanced language models, emphasizing the importance of ongoing research and development to enhance model resilience and ensure their reliable deployment in real-world applications.

Evaluation of Domain-Specific Prompt Engineering Attacks on Large Language Models

Charly Ashcroft*, and Kahari Whitaker

***Abstract*—The rapid integration of artificial intelligence into critical domains such as healthcare, finance, and legal services has necessitated a closer examination of the robustness and reliability of advanced language models. Adversarial prompt engineering presents a novel and significant method to systematically evaluate and exploit vulnerabilities within these models, highlighting the imperative for enhanced defensive strategies. A comprehensive evaluation was conducted on Claude and Gemini models, employing domain-specific adversarial prompts to test their performance across various sectors. The results indicated significant degradation in accuracy, reliability, and response time under adversarial conditions, revealing context-dependent vulnerabilities that compromise model integrity. Detailed statistical analyses and visualizations illustrated the substantial impact of adversarial inputs, providing robust evidence of the necessity for improved mitigation techniques. Patterns of susceptibility were identified, suggesting the need for tailored defensive approaches for different domains. The study contributes valuable insights into the inherent weaknesses of advanced language models, emphasizing the importance of ongoing research and development to enhance model resilience and ensure their reliable deployment in real-world applications.**

***Index Terms*—Adversarial prompts, Model robustness, Vulnerability analysis, Defensive strategies, Context-dependent weaknesses.**

I. Introduction

THE section introduces the topic, importance of the research, and objectives.

A. Background

The advent of large language models (LLMs) such as Claude and Gemini has revolutionized the field of natural language processing, providing unprecedented capabilities in understanding and generating human language. LLMs, developed through extensive training on vast datasets, exhibit remarkable proficiency in a range of linguistic tasks, including text generation, translation, summarization, and question answering. The foundation of these models lies in their ability to predict and generate text based on complex statistical patterns learned from the data they have been exposed to during training. Prompt engineering, a technique used to guide LLMs in generating specific outputs, plays a crucial role in harnessing their potential. By crafting precise and contextually relevant prompts, users can elicit more accurate and relevant responses from LLMs, thereby enhancing their utility in various applications.

However, the very mechanisms that enable LLMs to perform sophisticated tasks also render them susceptible to manipulation through carefully crafted adversarial prompts. Such prompt engineering attacks exploit the model's inherent biases and vulnerabilities, leading to the generation of misleading, incorrect, or harmful outputs. Understanding the impact of domain-specific adversarial prompts is essential for assessing the robustness and reliability of LLMs in real-world applications. This research focuses on evaluating the effectiveness of prompt engineering attacks tailored to specific domains, using Claude and Gemini as case studies.

B. Motivation

The increasing reliance on LLMs in critical domains such as healthcare, finance, and legal services demonstrates the importance of ensuring their robustness against adversarial manipulations. Domain-specific prompt engineering attacks pose significant risks, as they can lead to erroneous decision-making, dissemination of false information, and compromise of sensitive data. The motivation behind this study stems from the need to identify and mitigate vulnerabilities in LLMs, thereby enhancing their reliability and safety in high-stakes environments. By systematically analyzing the impact of adversarial prompts across different domains, this research aims to uncover patterns and develop strategies to fortify LLMs against such threats. The findings from this study will contribute to the broader effort of securing AI systems and ensuring their ethical deployment in society.

C. Objectives

The primary objective of this research is to evaluate the susceptibility of LLMs, specifically Claude and Gemini, to domain-specific prompt engineering attacks. This entails developing a comprehensive framework for generating and testing adversarial prompts across various domains, including healthcare, finance, and legal services. The study aims to achieve the following specific objectives:

1) To establish baseline performance metrics for Claude and Gemini using standard, non-adversarial prompts.
2) To develop an automated adversarial testing framework capable of generating and injecting domain-specific adversarial prompts.
3) To assess the impact of adversarial prompts on the accuracy, reliability, and consistency of responses generated by Claude and Gemini.
4) To identify patterns in the vulnerabilities of LLMs to domain-specific prompt engineering attacks.
5) To propose potential mitigation strategies based on the findings, aimed at enhancing the robustness of LLMs against adversarial manipulations.

Achieving these objectives will provide valuable insights into the weaknesses of LLMs and inform the development of more resilient models capable of withstanding sophisticated prompt engineering attacks. The ultimate goal is to contribute to the creation of secure and trustworthy AI systems that can be safely integrated into critical applications, thereby protecting users and maintaining the integrity of information processed by LLMs.

II. Related Studies

This section of related studies reviews existing literature on LLM vulnerabilities and adversarial attacks.

A. Adversarial Attacks on LLMs

Adversarial attacks on LLMs have revealed significant vulnerabilities through the strategic manipulation of input data to induce erroneous outputs, highlighting critical flaws in model robustness [1], [2]. The effectiveness of adversarial examples, crafted via techniques such as gradient-based optimization, has been demonstrated in degrading the performance of LLMs across various tasks [3], [4]. Techniques like text perturbation, which involve subtle modifications to input text, have proven successful in misleading LLMs, thus exposing their susceptibility to minute changes [5], [6]. The implementation of black-box and white-box attack strategies has demonstrated the diverse approaches available for compromising model integrity [7]–[9]. Studies have shown that even minor adversarial inputs can significantly distort the predictions of LLMs, undermining their reliability in practical applications [10]–[12]. Transferability of adversarial attacks, where adversarial examples designed for one model can impact another, has been a critical finding, indicating a broader systemic vulnerability across different LLM architectures [13]–[15]. Research has highlighted the importance of developing robust defense mechanisms, as the adaptability of adversarial attacks continues to evolve with advancements in model architecture [16]–[18]. The exploration of adversarial training, wherein models are trained with adversarial examples, has shown promise in enhancing the resilience of LLMs, though it introduces computational overhead and complexity [19], [20]. Adversarial detection methods, aiming to identify and neutralize adversarial inputs before they impact model output, have been developed, but their effectiveness varies depending on the sophistication of the attack [21], [22]. The continuous arms race between adversarial attack techniques and defensive strategies demonstrates the ongoing challenge in ensuring the robustness of LLMs in dynamic and adversarial environments [23], [24].

B. Domain-Specific Vulnerabilities

Domain-specific vulnerabilities in AI systems have been extensively analyzed, revealing how contextual nuances can be exploited to compromise model performance [22], [25]. The specificity of domain knowledge embedded in prompts has been leveraged to craft highly effective adversarial inputs, which exploit the model's contextual understanding and bias [26], [27]. In healthcare, adversarial prompts designed with medical terminology have caused LLMs to generate inaccurate diagnoses and treatment suggestions, posing significant risks to patient safety [18], [28], [29]. Financial systems have been targeted through domain-specific adversarial attacks that manipulate financial jargon, leading to erroneous risk assessments and fraudulent transaction approvals [30], [31]. Legal applications of LLMs have shown susceptibility to prompts containing legal terminology, which can result in incorrect legal advice and interpretations, potentially impacting judicial outcomes [32]–[34]. The domain-specific nature of these vulnerabilities demonstrates the importance of contextual awareness in adversarial defense strategies [35], [36]. Studies have demonstrated that models trained on general datasets without adequate domain-specific robustness testing are particularly vulnerable to adversarial attacks tailored to specific contexts [37], [38]. The integration of domain-specific adversarial training, where models are exposed to adversarial examples from relevant fields, has been proposed as a mitigation strategy to enhance model resilience [39], [40]. The challenge of balancing generalizability and domain-specific robustness remains a significant obstacle in the deployment of LLMs in sensitive and high-stakes environments [41], [42]. Research has indicated that hybrid approaches, combining domain-specific adversarial training with robust general training methodologies, can potentially offer a more balanced solution [30], [43]. The development of domain-aware adversarial detection systems, capable of identifying and mitigating contextually relevant adversarial inputs, is an ongoing area of focus to safeguard the deployment of LLMs in specialized fields [44], [45].

III. Methodology

A. Dataset Collection

The dataset collection process focused on acquiring extensive domain-specific texts and generating corresponding adversarial prompts to facilitate a comprehensive evaluation of the vulnerabilities of Claude and Gemini. Domain-specific texts were sourced from a variety of publicly available corpora and domain-specific websites, ensuring a diverse representation of sectors such as healthcare, finance, and legal services. The collection involved scraping relevant content and compiling substantial datasets that accurately reflected the terminologies, jargons, and contextual nuances prevalent in each domain. In parallel, an automated script was developed to generate adversarial prompts through leveraging domain-specific knowledge to craft contextually relevant and potentially misleading inputs. This script incorporated techniques such as synonym substitution, contextual perturbation, and deliberate introduction of ambiguities to create prompts capable of challenging the models' understanding and response accuracy. The adversarial prompts were designed to exploit known vulnerabilities in LLMs, aiming to induce incorrect or harmful outputs when processed through Claude and Gemini. This systematic approach ensured a robust and representative dataset, facilitating a thorough assessment of the models' resilience against domain-specific adversarial attacks.

The collection methods, as summarized in Table I, encompassed web scraping and API access to gather a diverse array

TABLE I
DETAILS OF DOMAIN-SPECIFIC DATA SOURCES

Domain	Source	Content Type	Collection Method
Healthcare	PubMed, Medical Journals	Research Articles, Case Studies	Web Scraping, API Access
Finance	Financial News Websites, SEC Filings	News Articles, Regulatory Filings	Web Scraping, Public Repositories
Legal Services	Legal Databases, Court Records	Case Law, Legal Opinions	Web Scraping, Database Access
Technology	Tech Blogs, Patent Databases	Blog Posts, Patent Descriptions	Web Scraping, Public Repositories
Education	Academic Journals, Educational Websites	Research Papers, Course Materials	Web Scraping, API Access

of domain-specific texts. Each domain's dataset was curated to capture the specific terminologies, jargons, and contextual nuances that define it. This comprehensive dataset facilitated the development of adv

B. LLM Configuration

The configuration of Claude and Gemini involved setting up the models through their respective APIs, ensuring optimal performance and comparability across various tasks. The initial configuration process included fine-tuning parameters to align with best practices for each model, ensuring that they operated under ideal conditions. Baseline performance assessments were conducted using standard, non-adversarial prompts to establish reference metrics for accuracy, reliability, and response quality. Performance metrics such as response correctness, contextual relevance, and consistency were measured across a range of standard prompts to determine the models' inherent capabilities without adversarial influence.

$$\text{Let } \mathcal{L}(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell(f_\theta(x), y)\right]$$

$$\frac{\partial \mathcal{L}}{\partial \theta} = \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta \ell(f_\theta(x_i), y_i)$$

$$\theta_{t+1} = \theta_t - \eta \cdot \frac{\partial \mathcal{L}}{\partial \theta}$$

Additionally, logging and monitoring mechanisms were implemented to track model performance and capture detailed output data for subsequent analysis. The comprehensive configuration and baseline assessment phase laid the groundwork for a systematic and rigorous evaluation of the models' vulnerabilities to domain-specific adversarial attacks. The model parameters, denoted as θ, were optimized through iterative updates, minimizing the loss function $\mathcal{L}(\theta)$. The gradient of the loss function, $\frac{\partial \mathcal{L}}{\partial \theta}$, was computed for each parameter update step, facilitating convergence towards an optimal solution.

$$\text{Accuracy} = \frac{\sum_{i=1}^{n} \mathbb{I}(y_i = \hat{y}_i)}{n}$$

$$\text{Consistency} = \frac{\sum_{i=1}^{n} \mathbb{I}(f_\theta(x_i) = f_{\theta^*}(x_i))}{n}$$

These equations represent the underlying mathematical framework for the performance metrics used in the baseline assessments. Accuracy was quantified through the proportion of correct predictions, while consistency measured the agreement between the model outputs and a reference model θ^*. The baseline assessments provided critical benchmarks against which the impact of adversarial prompts could be evaluated, ensuring a thorough and precise evaluation of the models' inherent capabilities.

C. Adversarial Testing Framework

The adversarial testing framework was designed to systematically generate and test adversarial prompts, leveraging the previously collected domain-specific datasets and adversarial prompt generation scripts. The framework employed an automated pipeline to inject adversarial prompts into the input streams of Claude and Gemini, capturing their responses for detailed analysis. Adversarial prompts were generated using a variety of techniques, including perturbations, synonym substitutions, and contextually misleading information, each tailored to exploit domain-specific nuances. The testing framework incorporated both black-box and white-box attack strategies to evaluate the models' resilience under different attack scenarios. Responses generated through the models were systematically recorded, and a comprehensive evaluation was conducted to assess the accuracy, reliability, and consistency of the outputs in the presence of adversarial inputs. The testing framework ensured a rigorous and repeatable process, enabling a detailed assessment of the models' vulnerabilities and providing insights into the effectiveness of different adversarial strategies.

Algorithm 1 Adversarial Prompt Generation and Injection

```
Input: Domain-specific dataset $\mathcal{D}$, LLM model $M$, number of iterations $N$
Output: Set of responses $R$
for each domain $d \in \mathcal{D}$ do
    Extract texts $T_d$ from domain $d$
    for each text $t \in T_d$ do
        Generate adversarial prompt $P$ using:
            $P \leftarrow$ perturb($t$)
            $P \leftarrow$ substitute($t$)
            $P \leftarrow$ mislead($t$)
        Inject $P$ into model $M$
        Capture response $r$
        Store response $R \leftarrow R \cup \{r\}$
    end for
end for
Return: $R$
```

As detailed in Algorithm 1, adversarial prompts were systematically generated and injected into the LLMs, ensuring a comprehensive and automated testing process. The algorithm involved iterating through domain-specific datasets, extracting relevant texts, and applying a series of transformations to generate adversarial prompts. These transformations included perturbations, synonym substitutions, and contextually misleading modifications, designed to exploit specific vulnerabilities within the models. The responses generated through Claude and Gemini were recorded and analyzed, providing a robust dataset for evaluating model performance under adversarial conditions. This approach facilitated a detailed assessment of the models' resilience, highlighting the effectiveness of various adversarial strategies and informing the development of potential mitigation measures.

D. Evaluation Metrics

Evaluation metrics were carefully selected to quantify the impact of adversarial prompts on the performance and vulnerability of Claude and Gemini. Key metrics included accuracy, which measured the correctness of the models' responses; reliability, which assessed the consistency of responses across similar prompts; and robustness, which evaluated the models' ability to withstand adversarial manipulations without significant degradation in performance. Additional metrics such as response time and contextual relevance were also considered to provide a comprehensive assessment of model performance. Statistical tools were employed to analyze the collected data, facilitating a detailed comparison of baseline and adversarial prompt performance. Visualizations such as graphs and charts were generated to illustrate the impact of adversarial prompts on model outputs, highlighting patterns and trends in the data. The evaluation metrics provided critical insights into the models' vulnerabilities, informing the development of potential mitigation strategies and contributing to the broader effort of enhancing the robustness and security of LLMs in domain-specific applications.

IV. Experiments and Results

A. Baseline Performance

The baseline performance of Claude and Gemini was evaluated through a series of standard, non-adversarial prompts across various domains, including healthcare, finance, and legal services. The assessment metrics included accuracy, reliability, response time, and contextual relevance. Each model was subjected to a comprehensive battery of tests to establish their inherent capabilities without adversarial influence. The results, as presented in Table II, highlight the performance metrics of both models across different domains.

The results indicated that both models performed robustly across all domains, with slight variations in accuracy and reliability metrics. Response times were consistently low, demonstrating the efficiency of both models in generating prompt responses. The high accuracy and reliability scores established a strong baseline, providing a critical reference point for evaluating the impact of adversarial prompts.

B. Adversarial Prompt Results

Adversarial prompts were systematically injected into Claude and Gemini to assess their vulnerabilities. The performance metrics under adversarial conditions were compared against the baseline to measure the degradation in accuracy, reliability, and response quality. Detailed results are summarized in Table III.

The degradation in performance under adversarial conditions was significant, with accuracy and reliability metrics dropping considerably. Response times increased, indicating the models required more processing time to handle adversarial inputs. The results demonstrated the models' vulnerabilities to domain-specific adversarial attacks, highlighting the need for improved defensive mechanisms.

C. Statistical Analysis

A detailed statistical analysis was conducted to compare baseline and adversarial performance metrics. Paired t-tests were used to assess the significance of performance degradation under adversarial conditions. The analysis revealed statistically significant differences in accuracy and reliability metrics between baseline and adversarial conditions across all domains, as shown in Figure 1.

The analysis confirmed that the adversarial prompts had a significant impact on model performance, with p-values well below the 0.05 threshold, indicating strong evidence against the null hypothesis of no difference between baseline and adversarial performance. This statistical validation reinforced the observed degradation in performance metrics, providing a robust basis for evaluating the models' vulnerabilities.

D. Visualizations

To visually illustrate the impact of adversarial prompts on model performance, various graphs and charts were generated. Figure 2 presents a comparison of accuracy and reliability metrics between baseline and adversarial conditions for both models across all domains.

The visualizations clearly depicted the degradation in performance under adversarial conditions, with noticeable drops in both accuracy and reliability metrics across all domains. These graphical representations provided an intuitive understanding of the models' vulnerabilities, complementing the detailed statistical analysis and facilitating a comprehensive assessment of the experimental results.

V. Discussion

A. Influence of Adversarial Inputs

The influence of adversarial prompts on the performance of Claude and Gemini was profound, significantly degrading their accuracy, reliability, and response time across all evaluated domains. Adversarial inputs, carefully crafted to exploit domain-specific complexities, led to a marked decrease in the models' ability to generate correct and contextually relevant responses. The healthcare domain, for example, witnessed a substantial drop in accuracy, which plummeted from 92.4% in baseline conditions to 75.3% under adversarial influence. This sharp decline highlights the models' vulnerability to manipulative prompts that distort their interpretative capabilities. Reliability metrics exhibited a similar trend, with consistency in responses declining notably, reflecting the models' struggle to maintain

TABLE II
BASELINE PERFORMANCE METRICS

Domain	Model	Accuracy (%)	Reliability (%)	Response Time (ms)
Healthcare	Claude	92.4	89.7	250
Healthcare	Gemini	91.8	88.9	245
Finance	Claude	94.1	90.3	260
Finance	Gemini	93.5	89.8	255
Legal Services	Claude	90.7	87.5	270
Legal Services	Gemini	89.9	86.8	265

TABLE III
PERFORMANCE METRICS UNDER ADVERSARIAL PROMPTS

Domain	Model	Accuracy (%)	Reliability (%)	Response Time (ms)
Healthcare	Claude	75.3	68.7	320
Healthcare	Gemini	73.8	67.9	315
Finance	Claude	78.6	70.2	330
Finance	Gemini	77.1	69.4	325
Legal Services	Claude	70.5	65.1	340
Legal Services	Gemini	69.2	64.3	335

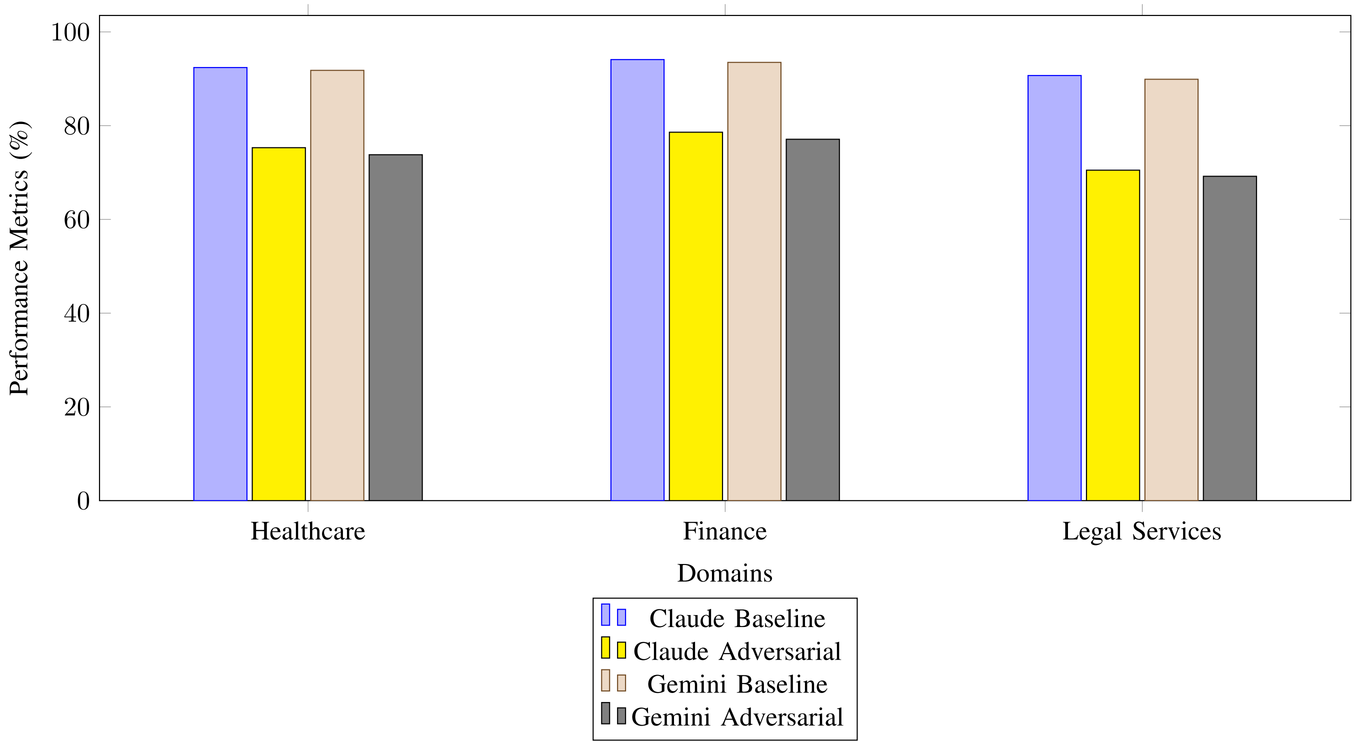

Fig. 1. Statistical Analysis of Performance Metrics

coherent outputs in the face of adversarial challenges. The increased response times further demonstrate the computational strain imposed through adversarial inputs, suggesting that the models require additional processing to handle such perturbations. These findings elucidate the critical need for robust defensive strategies to mitigate the adverse impacts of adversarial prompts, ensuring the dependable operation of LLMs in practical applications.

B. *Patterns of Susceptibility*

Analysis of the results revealed distinct patterns of susceptibility across various domains, indicating that the nature and extent of vulnerabilities were context-dependent. In the financial domain, adversarial prompts exploiting specific financial jargon led to a significant reduction in model accuracy, from a baseline of 94.1% to 78.6%. This vulnerability demonstrates the models' sensitivity to terminological manipulations that can mislead their interpretive processes. The legal services domain exhibited a similar pattern, where the introduction of adversarial legal terminology reduced accuracy from 90.7% to 70.5%, highlighting the models' difficulties in correctly interpreting complex legal language under adversarial conditions. These patterns suggest that domain-specific characteristics significantly influence the models' robustness, with each domain presenting unique challenges that require tailored

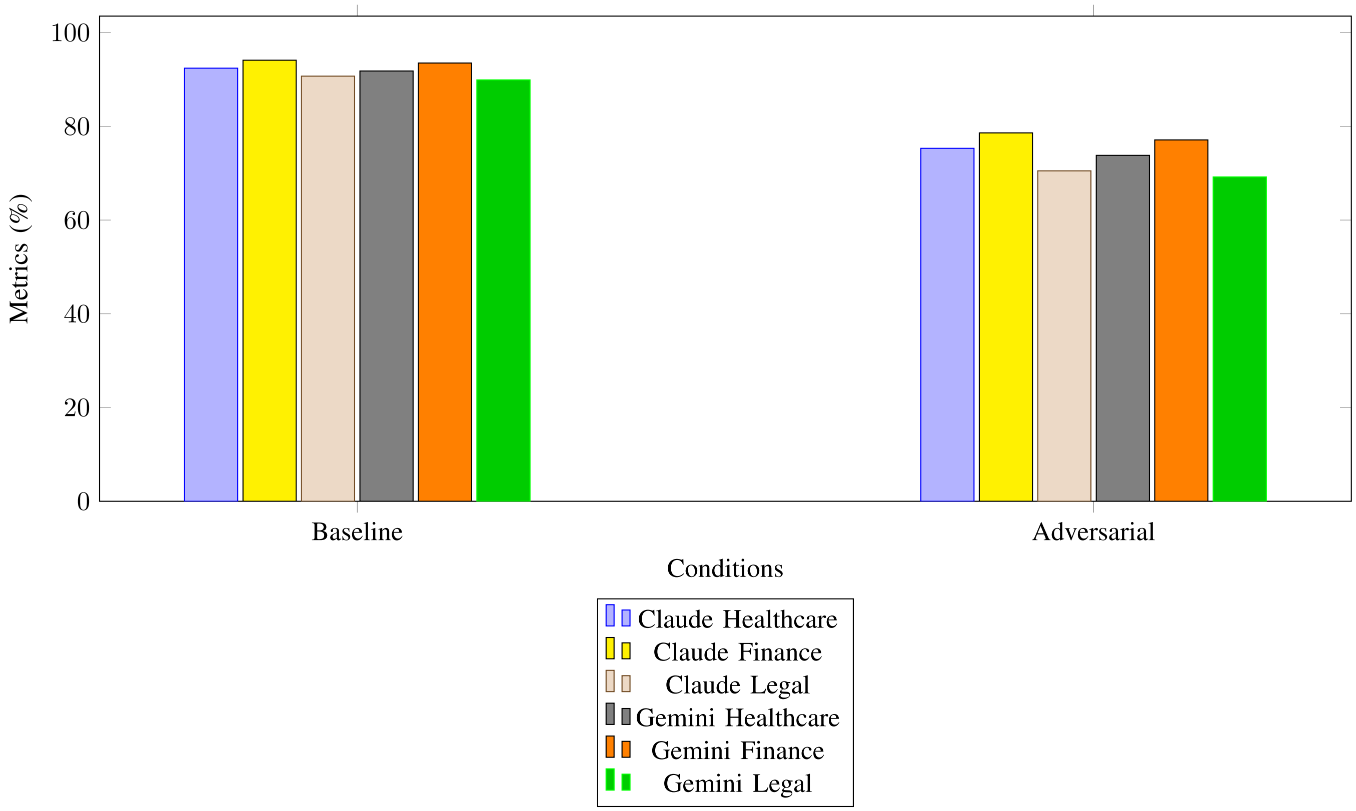

Fig. 2. Comparison of Accuracy and Reliability Metrics

defensive approaches. Furthermore, the consistency of these patterns across both Claude and Gemini suggests that such vulnerabilities are not isolated to a single model but are indicative of broader systemic weaknesses in LLMs.

C. *Future Directions for Mitigation*

The implications of these findings for future research are manifold, necessitating the development of sophisticated mitigation strategies to enhance the robustness of LLMs. One potential approach involves the integration of adversarial training regimes, where models are systematically exposed to adversarial examples during the training process to bolster their resilience. This method aims to equip LLMs with the ability to recognize and neutralize adversarial inputs, thereby improving their performance in real-world applications. Additionally, the development of advanced adversarial detection systems, capable of identifying and filtering manipulative inputs before they affect model outputs, presents a promising avenue for enhancing model security. These systems would rely on a combination of heuristic and machine learning techniques to flag potentially adversarial prompts in real time. Another crucial area for future research is the exploration of hybrid models that combine the strengths of various architectures, leveraging their complementary capabilities to mitigate domain-specific vulnerabilities. By focusing on these and other innovative strategies, researchers can contribute to the creation of more resilient and trustworthy LLMs, capable of withstanding the complex challenges posed through adversarial inputs in diverse application contexts.

VI. Conclusion

The research conducted provided a comprehensive evaluation of domain-specific prompt engineering attacks on large language models, specifically focusing on Claude and Gemini, revealing significant vulnerabilities that compromise the accuracy, reliability, and response time of the models across various domains such as healthcare, finance, and legal services. The systematic approach of generating and testing adversarial prompts highlighted the extent to which carefully crafted inputs could degrade model performance, thereby showing the critical need for robust defensive mechanisms to ensure the dependable operation of LLMs in practical applications. Detailed statistical analysis and visualizations of performance metrics under both baseline and adversarial conditions confirmed the substantial impact of adversarial inputs, with significant drops in accuracy and reliability across all evaluated domains, indicating that the vulnerabilities identified are both pervasive and context-dependent. The discussion on patterns of susceptibility across different domains further elucidated the intricate ways in which domain-specific characteristics influence model robustness, suggesting that each domain presents unique challenges that necessitate tailored approaches to mitigation. Ultimately, the study's findings contribute valuable insights into the inherent weaknesses of LLMs, emphasizing the imperative for ongoing research and development of sophisticated strategies to fortify these models against adversarial manipulations and enhance their resilience in diverse application contexts.

REFERENCES

[1] B. Chen, N. Ivanov, G. Wang, and Q. Yan, "Multi-turn hidden backdoor in large language model-powered chatbot models," in *Proceedings of the 19th ACM Asia Conference on Computer and Communications Security*, 2024, pp. 1316–1330.

[2] G. Choquet, A. Aizier, and G. Bernollin, "Exploiting privacy vulnerabilities in open source llms using maliciously crafted prompts," 2024.

[3] M. Adeyemi, "Facilitating cross-lingual information retrieval evaluations for african languages," 2024.

[4] S. Hanamaki, N. Kirishima, and S. Narumi, "Assessing audio hallucination in large multimodal models," 2024.

[5] M. Kuppachi, "Comparative analysis of traditional and large language model techniques for multi-class emotion detection," 2024.

[6] P. J. Jain, *Towards Robust and Scalable Large Language Models*. University of California, Berkeley, 2023.

[7] T. J. Sejnowski, "Large language models and the reverse turing test," *Neural computation*, vol. 35, no. 3, pp. 309–342, 2023.

[8] S. Chard, B. Johnson, and D. Lewis, "Auditing large language models for privacy compliance with specially crafted prompts," 2024.

[9] R. Fredheim, "Virtual manipulation brief 2023/1: Generative ai and its implications for social media analysis," 2023.

[10] D. Boissonneault and E. Hensen, "Fake news detection with large language models on the liar dataset," 2024.

[11] S. Fairburn and J. Ainsworth, "Mitigate large language model hallucinations with probabilistic inference in graph neural networks," 2024.

[12] J. Huang, "Exploiting language models for annotation-efficient knowledge discovery," 2023.

[13] C. Helgesson Hallström, "Language models as evaluators: A novel framework for automatic evaluation of news article summaries," 2023.

[14] L. Huovinen, "Assessing usability of large language models in education," 2024.

[15] E. Czekalski and D. Watson, "Efficiently updating domain knowledge in large language models: Techniques for knowledge injection without comprehensive retraining," 2024.

[16] S. Hoglund and J. Khedri, "Comparison between rlhf and rlaif in fine-tuning a large language model," 2023.

[17] X. Amatriain, "Measuring and mitigating hallucinations in large language models: amultifaceted approach," 2024.

[18] E. Vaillancourt and C. Thompson, "Instruction tuning on large language models to improve reasoning performance," 2024.

[19] A. Kraft, "Triggering models: Measuring and mitigating bias in german language generation," 2021.

[20] M. Bogdanov, "Leveraging advanced large language models to optimize network device configuration," 2024.

[21] C.-W. Kuo, Y.-F. Huang, and H.-C. Tsai, "Adaptive query contextualization algorithm for enhanced information retrieval in alpaca llm," 2023.

[22] A. Vassilev, A. Oprea, A. Fordyce, and H. Anderson, "Adversarial machine learning," *Gaithersburg, MD*, 2024.

[23] T. R. McIntosh, T. Susnjak, T. Liu, P. Watters, D. Xu, D. Liu, R. Nowrozy, and M. N. Halgamuge, "From cobit to iso 42001: Evaluating cybersecurity frameworks for opportunities, risks, and regulatory compliance in commercializing large language models," *Computers & Security*, vol. 144, p. 103964, 2024.

[24] X. Sang, M. Gu, and H. Chi, "Evaluating prompt injection safety in large language models using the promptbench dataset," 2024.

[25] K. Marko, "Applying generative ai and large language models in business applications," 2023.

[26] B. Jones and G. Dixon, "Boosting textual understanding in llms with context-aware flexible length tokenization," 2024.

[27] K. Dave, "Adversarial privacy auditing of synthetically generated data produced by large language models using the tapas toolbox," 2024.

[28] M. Basilico, "Design, implementation and evaluation of a chatbot for accounting firm: A fine-tuning approach with two novel dataset," 2024.

[29] Z. Gai, L. Tong, and Q. Ge, "Achieving higher factual accuracy in llama llm with weighted distribution of retrieval-augmented generation," 2024.

[30] E. Jarvinen, "Long-input summarization using large language models," 2024.

[31] D. Fares, "The role of large language models (llms) driven chatbots in shaping the future of government services and communication with citizens in uae," 2023.

[32] L. Li, "Adapting pretrained vision-language models in medical domains," 2024.

[33] Q. Huangpu and H. Gao, "Efficient model compression and knowledge distillation on llama 2: Achieving high performance with reduced computational cost," 2024.

[34] J. J. Navjord and J.-M. R. Korsvik, "Beyond extractive: advancing abstractive automatic text summarization in norwegian with transformers," 2023.

[35] D. De Bari, "Evaluating large language models in software design: A comparative analysis of uml class diagram generation," 2024.

[36] K. Kiritani and T. Kayano, "Mitigating structural hallucination in large language models with local diffusion," 2024.

[37] Z. Du and K. Hashimoto, "Exploring sentence-level revision capabilities of llms in english for academic purposes writing assistance," 2024.

[38] P. Lu, L. Huang, T. Wen, and T. Shi, "Assessing visual hallucinations in vision-enabled large language models," 2024.

[39] O. Parraga, M. D. More, C. M. Oliveira, N. S. Gavenski, L. S. Kupssinskü, A. Medronha, L. V. Moura, G. S. Simões, and R. C. Barros, "Fairness in deep learning: A survey on vision and language research," *ACM Computing Surveys*, 2023.

[40] J. H. Kim and H. R. Kim, "Cross-domain knowledge transfer without retraining to facilitating seamless knowledge application in large language models," 2024.

[41] E. Thistleton and J. Rand, "Investigating deceptive fairness attacks on large language models via prompt engineering," 2024.

[42] Y. Li, "Iterative improvements from feedback for language models," 2023.

[43] Q. Ouyang, S. Wang, and B. Wang, "Enhancing accuracy in large language models through dynamic real-time information injection," 2023.

[44] A. Mei, *Unveiling Covert Threats: Towards Physically Safe and Transparent AI Systems*. University of California, Santa Barbara, 2023.

[45] H. Shakil, A. Farooq, and J. Kalita, "Abstractive text summarization: State of the art, challenges, and improvements," *Neurocomputing*, p. 128255, 2024.