Security Architectures for Generative-AI Systems

Muhammad Salar 21K-4619
Muhammad Hamza 21K-4579
Hamza Raza 21K-4699

***Abstract*—The rapid advancements in Generative AI (GenAI) have propelled its integration across various domains, yet its widespread adoption is marred by significant security and privacy vulnerabilities. This report consolidates insights from four seminal papers presented at SAGAI'24, each addressing distinct challenges and proposing innovative defences within the realm of GenAI security. The first paper introduces *spotlighting*, a novel prompt-engineering methodology designed to defend large language models (LLMs) against *indirect prompt injection (IPI) attacks*, achieving a substantial reduction in attack success rates with minimal impact on task performance. The second paper expands the scope of security in *multimodal large language models (MLLMs)* by proposing a two-stage defence mechanism comprising *input validation* and *prompt injection detection*, effectively safeguarding against adversarial image-based attacks. The third paper examines the *dual-use risks* of instruction-following LLMs, demonstrating how adversarial actors exploit their programmatic behaviours to generate malicious content efficiently, bypassing existing security filters. Finally, the fourth paper explores the potential of *self-supervised learning encoders* to significantly enhance the robustness and privacy-preserving capabilities of supervised learning models, particularly against adversarial examples, data poisoning, and inference attacks.**

Together, these works underscore the multifaceted nature of threats targeting GenAI systems and highlight the urgent need for robust, integrative defences. By addressing specific attack vectors and vulnerabilities, these contributions lay a critical foundation for future research aimed at fortifying the safety and integrity of GenAI applications in the information security domain.

***Index Terms*—Generative AI, Large Language Models, Multimodal Large Language Models, Indirect Prompt Injection Attacks, Image-Based Prompt Injection, Self-Supervised Learning, Privacy-Preserving Learning, Adversarial Attacks, Information Security.**

I. Introduction

A. The Evolution and Challenges of Generative AI

Generative AI (GenAI) has witnessed unprecedented growth in recent years, revolutionizing industries by automating tasks, generating creative content, and advancing human-computer interaction. By leveraging large-scale machine learning (ML) models trained on vast datasets, GenAI systems can perform diverse and complex tasks, ranging from natural language processing to image synthesis and multimodal understanding. Models such as GPT-4, Stable Diffusion, and others exemplify the potential of GenAI in transforming workflows and enriching user experiences.

However, the versatility and power of GenAI also come with significant risks. As these systems grow in capability and complexity, they become increasingly susceptible to adversarial attacks and vulnerabilities that can compromise their reliability, safety, and ethical application. These vulnerabilities arise not only from the inherent limitations of ML models but also from the intricate architectures that integrate multiple components, external APIs, and chained model outputs. As a result, securing GenAI systems requires addressing not just individual model weaknesses but also systemic threats.

B. Security Concerns in Generative AI Systems

The convergence of GenAI's immense potential and its critical vulnerabilities has made it a focal point for both innovation and exploitation. Adversarial threats targeting GenAI systems often exploit their:

- **Prompt Sensitivity:** The reliance on natural language prompts exposes GenAI models to *prompt injection attacks*, where malicious inputs manipulate model behaviour.
- **Architectural Complexity:** The integration of multiple models and external components introduces attack vectors such as *indirect prompt injections* and *multimodal prompt attacks*, often bypassing traditional defences.
- **Dual-Use Risks:** Instruction-following capabilities, essential for many applications, also enable the generation of malicious outputs, such as spam, misinformation, and phishing attempts.
- **Privacy Concerns:** Data used for training and inference often carries sensitive information, posing risks of inference attacks, data poisoning, and challenges in implementing privacy-preserving mechanisms.

C. Contributions of This Work

This report consolidates key insights from four foundational papers presented at SAGAI'24, each addressing a critical aspect of GenAI security:

1) **Defending Against Indirect Prompt Injection Attacks With Spotlighting:** This work introduces the concept of spotlighting—an innovative set of techniques to mitigate indirect prompt injection (IPI) attacks by enabling models to distinguish between trusted and untrusted inputs.
2) **Defending Language Models Against Image-Based Prompt Attacks via User-Provided Specifications:** Focused on multimodal large language models (MLLMs), this paper proposes a robust two-stage defence mechanism to combat malicious content embedded in visual data inputs.
3) **Exploiting Programmatic Behaviour of LLMs: Dual-Use Through Standard Security Attacks:** By exploring the dual-use nature of instruction-following LLMs, this study highlights how adversarial actors can economically and effectively exploit these systems for malicious purposes.

4) **Pre-Trained Encoders in Self-Supervised Learning for Secure and Privacy-Preserving Supervised Learning:** This work investigates how pre-trained encoders in self-supervised learning can enhance the robustness and privacy of supervised learning models, addressing long-standing challenges in adversarial and privacy-preserving applications.

D. Broader Implications for Information Security

The insights from these studies underscore the pressing need for a holistic approach to securing GenAI systems. Unlike traditional cybersecurity paradigms, GenAI security must account for:

- The evolving threat landscape, including novel attack vectors and adversarial strategies.
- The ethical and societal impacts of GenAI misuse, from misinformation to privacy violations.
- The integration of interdisciplinary solutions, spanning robust algorithm design, regulatory frameworks, and ethical AI practices.

This work not only synthesizes cutting-edge research but also aims to provide a roadmap for future directions in the information security domain. By addressing the vulnerabilities of GenAI systems holistically, we can unlock their potential while safeguarding against misuse, ensuring these transformative technologies serve as tools for progress rather than harm.

II. Related Work

The rapid proliferation of Generative AI (GenAI) systems has necessitated a deeper exploration of their inherent vulnerabilities and the development of robust defensive measures. Prior research has addressed critical issues in the domains of adversarial attacks, multimodal integration, dual-use risks, and privacy preservation, providing foundational insights for advancing GenAI security.

A. Defending Against Prompt Injection Attacks

Prompt injection attacks, a prominent threat to the integrity of language models, exploit the model's susceptibility to adversarially crafted inputs. Existing approaches to mitigate these attacks include fine-tuning and filtering mechanisms, but these often fall short against sophisticated attackers. Spotlighting, introduced as a novel defence, builds upon prior techniques by employing methods such as contextual delimiting, datamarking, and encoding to differentiate between trusted and malicious inputs. These advancements enable models to retain functionality while significantly reducing susceptibility to prompt-based adversarial manipulations.

B. Security for Multimodal Large Language Models

Multimodal large language models (MLLMs), which process text and visual inputs, expand the capabilities of AI systems but introduce unique vulnerabilities. Prior studies have revealed how attackers embed malicious payloads within visual data, bypassing conventional text-based defences. Current research emphasizes the importance of a two-stage security framework: input validation to preemptively filter malicious images and injection detection mechanisms to analyse intent and safeguard model responses. This approach builds on foundational work in adversarial robustness and extends it to the domain of multimodal inputs.

C. Addressing Dual-Use Risks in Instruction-Following Models

Instruction-following models have revolutionized human-computer interaction but also exacerbated the risks of malicious exploitation. Studies highlight how attackers exploit programmatic behaviours within these models to generate harmful outputs, circumventing existing safeguards. Techniques such as obfuscation, payload splitting, and content-specific bypassing are frequently employed to evade detection. Research in this area underscores the need for adaptive, context-aware defences that balance usability with robust security mechanisms, a challenge that remains at the forefront of GenAI safety efforts.

D. Enhancing Privacy-Preserving Learning

The integration of self-supervised learning into privacy-preserving AI systems addresses longstanding challenges in balancing security, efficiency, and performance. Traditional methods, such as differentially private training and randomized smoothing, often compromise accuracy or scalability. Recent advances leverage pre-trained encoders to provide high-quality feature representations, enabling models to achieve enhanced robustness against adversarial attacks and stronger privacy guarantees. These findings are pivotal in reshaping privacy-centric AI development, bridging the gap between theoretical guarantees and practical deployment.

Summary of Contributions

The existing body of work highlights the multifaceted nature of GenAI vulnerabilities and the diverse methodologies required to address them. From contextual defence mechanisms like spotlighting to systemic frameworks for multimodal security and privacy preservation, these contributions establish a strong foundation for advancing GenAI safety in increasingly complex and adversarial environments.

III. Methodology

The methodologies in the four studies employ innovative techniques tailored to address specific security and privacy challenges in Generative AI systems. Each approach presents a unique perspective while collectively advancing the robustness, privacy, and security of these systems.

A. Spotlighting for Indirect Prompt Injection Attacks

Spotlighting was proposed as a novel defence mechanism to mitigate indirect prompt injection (IPI) attacks. This methodology focuses on equipping models with tools to discern malicious inputs without compromising their ability to process legitimate queries.

Components of Spotlighting

: **Delimiting:**

- Inputs are structured with explicit textual boundaries to delineate instructions from an external context.
- For example, trusted inputs are enclosed within predefined markers (e.g., `<<START>> ... <<END>>`) to signal their authenticity.

Datamarking:

- Unique markers are embedded at various points in the input, creating traceable identifiers.
- Markers can be designed to persist through transformations, enabling forensic analysis if outputs deviate from expectations.

Encoding:

- Inputs undergo reversible transformations (e.g., Base64 encoding) to obfuscate their content from direct interpretation.
- Decoding is performed within a secure and controlled environment, minimizing exposure to adversarial manipulation.

Experimental Setup:

- A dataset of benign and adversarial prompts was curated, covering common scenarios like conversational queries and complex task instructions.
- LLMs were tested with and without spotlighting mechanisms to assess their resilience to IPIs.
- Metrics included attack success rates, response fidelity, and overall model performance in a simulated operational environment.

B. Multimodal Large Language Model Defence Framework

To address the unique vulnerabilities of multimodal systems like GPT-4 Vision and LLAVA, a two-stage defence framework was developed. This method integrates validation and detection mechanisms to ensure secure operation across text and image modalities.

Framework Design

: **Stage 1: Input Validation:**

- User-provided specifications define acceptable input formats and content.
- Images are analysed for pixel integrity and watermark consistency, while text inputs undergo lexical and semantic filtering.
- Validation leverages a whitelist and heuristic-based rules to block known attack patterns proactively.

Stage 2: Prompt Injection Detection:

- Intermediate representations of inputs, such as embeddings and activation maps, are analysed for malicious intent.
- Task-specific embeddings are compared against a database of adversarial patterns to flag suspicious activity.
- Detection employs machine learning classifiers trained on adversarial and non-adversarial multimodal datasets.

Evaluation and Testing:

- Real-world adversarial datasets, including manipulated images and poisoned text inputs, were used to challenge the framework.
- Key performance metrics included detection precision, recall, processing latency, and the model's overall accuracy when under attack.

C. Exploiting Programmatic Behaviour in Instruction-Following Models

This study explored how instruction-following large language models (LLMs) can be exploited for dual-use purposes, such as generating malicious content. The methodology involved designing and testing various attack vectors to evaluate model vulnerabilities.

Attack Techniques:

- **Obfuscation:** Attackers slightly altered prompts to bypass content filters, for example, introducing typos or encoding key phrases.
- **Payload Splitting:** Malicious instructions were divided into smaller, seemingly harmless fragments, which were then reconstructed by the model during processing.
- **Prompt Chaining:** Sequences of prompts were crafted to exploit the model's contextual understanding and bypass security mechanisms iteratively.

Evaluation Methodology:

- Models like GPT-4 and InstructGPT were subjected to crafted adversarial prompts to measure their ability to generate malicious content.
- Metrics included the success rate of bypassing content filters, the coherence of malicious outputs, and the economic feasibility of executing such attacks.

D. Privacy-Preserving Learning via Pre-Trained Encoders

This study leveraged pre-trained encoders from self-supervised learning to enhance the security and privacy of supervised learning systems. The methodology addressed limitations in adversarial robustness and differential privacy without sacrificing model performance.

Pre-Training Setup:

- Encoders were trained on large-scale public datasets using unsupervised objectives to capture general-purpose feature representations.
- For downstream supervised tasks, two approaches were evaluated:
  - **Linear Probing:** Only the final classification layer was trained while keeping encoder parameters frozen.
  - **Fine-Tuning:** Both the encoder and the classification layer were updated during training.

Privacy and Security Enhancements:

- **Adversarial Robustness:** Certified defences, such as randomized smoothing and bagging, were augmented with pre-trained encoders to improve accuracy against adversarial perturbations.

- **Differential Privacy:** The training process incorporated differential privacy mechanisms (e.g., DP-SGD) to ensure formal privacy guarantees for user data.
- **Exact Machine Unlearning:** Efficient unlearning protocols were implemented to remove specific user data from models without requiring full retraining.

Evaluation Metrics:

- Accuracy under benign and adversarial conditions was measured across datasets such as STL10, CIFAR10, and Tiny-ImageNet.
- Privacy guarantees were quantified using differential privacy budgets, while unlearning efficiency was assessed in terms of runtime and resource utilization.

Summary of Methodological Innovations

The methodologies adopted in these studies highlight diverse yet complementary approaches to addressing security and privacy challenges in Generative AI:

- Contextual defences like spotlighting offer practical solutions to input manipulation.
- Multimodal frameworks address the unique vulnerabilities of systems combining text and visual inputs.
- Exploration of programmatic vulnerabilities provides actionable insights into mitigating dual-use risks.
- Pre-trained encoders demonstrate transformative potential for privacy-preserving AI without compromising utility.

IV. Results

The results from the four studies provide significant insights into the effectiveness of the proposed methodologies in mitigating adversarial threats, safeguarding multimodal systems, addressing dual-use risks, and enhancing privacy-preserving supervised learning. Each study demonstrated measurable improvements in system robustness, security, and utility under adversarial conditions.

A. Spotlighting for Indirect Prompt Injection Attacks

Spotlighting effectively mitigated indirect prompt injection (IPI) attacks while preserving task performance in language models.

Attack Mitigation:

- Models equipped with spotlighting showed a 98% reduction in attack success rates compared to baseline systems without defences.
- Techniques like delimiting and datamarking improved the model's ability to distinguish malicious from legitimate inputs.

Performance Retention:

- The response accuracy for non-adversarial tasks was unaffected, with less than a 1% performance degradation observed in edge cases.

Scalability:

- Spotlighting proved scalable across a range of prompt types, including simple queries, chained instructions, and complex scenarios requiring contextual understanding.

B. Multimodal Large Language Model Defence Framework

The defence framework significantly enhanced the robustness of multimodal large language models (MLLMs) like GPT-4 Vision against adversarial inputs.

Input Validation:

- Successfully filtered 92% of adversarial image inputs and 95% of malicious text inputs during testing.
- User-provided specifications improved the flexibility and adaptability of the framework to novel attack patterns.

Injection Detection:

- Achieved a detection precision of 96% and a recall of 93% for adversarial prompts in a multimodal context.

Task Performance:

- Maintained system accuracy within 2% of baseline performance for non-adversarial inputs, ensuring operational reliability.

Resilience Across Modalities:

- Demonstrated robustness in handling complex adversarial scenarios involving simultaneous manipulation of text and images.

C. Exploiting Programmatic Behaviour in Instruction-Following Models

The study highlighted the vulnerabilities of instruction-following models to dual-use exploitation while providing valuable insights into potential mitigations.

Vulnerability Assessment:

- Attackers achieved a 100% success rate in generating malicious content using techniques like obfuscation and payload splitting on unprotected models.
- The malicious content generated was both coherent and contextually relevant, demonstrating the dual-use risks of instruction-following models.

Economic Feasibility:

- Malicious use of models was shown to be highly cost-effective:
  - Generation costs ranged between $0.0064 and $0.016 per query, significantly lower than traditional methods.
- Highlighted the economic incentives for adversarial actors to adopt these models for harmful purposes.

Potential Mitigations:

- Recommendations included integrating dynamic contextual defences and enhancing filtering systems to address programmatic vulnerabilities.

D. Privacy-Preserving Learning via Pre-Trained Encoders

Integrating pre-trained encoders into privacy-preserving supervised learning demonstrated significant benefits in robustness, privacy guarantees, and efficiency.

Adversarial Robustness:

- Models with pre-trained encoders achieved up to a 67% increase in certified robustness against adversarial examples compared to traditional supervised learning systems.

- Techniques like randomized smoothing and bagging benefited significantly from the high-quality feature representations of the encoders.

Privacy Guarantees:

- Differentially private classifiers with pre-trained encoders showed a 4x improvement in accuracy under a strict privacy budget.
- Exact machine unlearning tasks demonstrated orders of magnitude improvements in runtime and efficiency, reducing retraining costs while maintaining privacy compliance.

Utility and Scalability:

- Across datasets like STL10, CIFAR10, and Tiny-ImageNet, models achieved higher accuracy and robustness with reduced resource requirements.
- Fine-tuning approaches outperformed linear probing in adversarial settings, though linear probing offered greater efficiency in privacy-preserving tasks.

Summary of Results

The results across all four studies demonstrate the effectiveness of the proposed methods:

- Spotlighting provided a robust defence against IPI attacks with minimal impact on model performance.
- Multimodal frameworks safeguarded against complex adversarial inputs across text and visual modalities, ensuring operational reliability.
- Dual-use analysis exposed critical vulnerabilities in instruction-following models, emphasizing the need for adaptive countermeasures.
- Pre-trained encoders significantly improved adversarial robustness, privacy guarantees, and efficiency, making them a pivotal tool for secure AI deployment.

V. Discussion

The findings from the four studies underscore the growing complexity of securing Generative AI (GenAI) systems and highlight the effectiveness of innovative methodologies in mitigating emerging threats. This section explores the broader implications, strengths, limitations, and potential future directions of the methodologies employed in these studies.

A. Spotlighting as a Defence for Indirect Prompt Injection Attacks

Spotlighting demonstrates significant potential as a scalable and effective defence against indirect prompt injection (IPI) attacks.

Strengths:

- The techniques of delimiting, datamarking, and encoding provided robust mechanisms to distinguish trusted inputs from adversarial ones.
- The minimal performance degradation observed across tasks highlights the practicality of spotlighting in real-world applications.
- The approach is adaptable to diverse use cases, making it a versatile solution for securing language models in both conversational and task-oriented scenarios.

Limitations:

- The encoding step introduces an additional computational overhead, which may impact the real-time responsiveness of systems.
- Attackers could potentially adapt to the spotlighting mechanisms over time, requiring continuous updates to defence strategies.

B. Multimodal Large Language Model Defence Framework

The defence framework for multimodal large language models (MLLMs) addresses critical vulnerabilities unique to systems handling both textual and visual inputs.

Strengths:

- The two-stage approach of input validation and injection detection significantly enhanced the robustness of MLLMs against adversarial attacks.
- High detection precision and recall metrics demonstrated the reliability of the framework in identifying malicious inputs.
- The flexibility of user-provided specifications enabled tailored defences for specific use cases, enhancing the adaptability of the system.

Limitations:

- The framework relies on pre-defined user specifications, which may not fully capture emerging attack patterns.
- Processing multimodal inputs with advanced validation and detection mechanisms may introduce latency, limiting deployment in time-sensitive applications.

C. Dual-Use Risks in Instruction-Following Models

The exploration of dual-use risks highlights the urgent need to address the programmatic vulnerabilities of instruction-following models.

Strengths:

- The study effectively identified key techniques (e.g., obfuscation, payload splitting) used by attackers to exploit these models.
- Economic analysis provided actionable insights into the feasibility of adversarial use, emphasizing the need for proactive mitigation strategies.
- Highlighting the dual-use nature of these models paves the way for ethical AI frameworks and stricter governance policies.

Limitations:

- The study primarily focused on technical vulnerabilities, with limited exploration of ethical and regulatory measures to curb misuse.
- Addressing programmatic risks may require trade-offs in model usability and accessibility, impacting broader adoption.

D. Privacy-Preserving Learning via Pre-Trained Encoders

The use of pre-trained encoders for secure and privacy-preserving supervised learning offers a transformative approach to mitigating adversarial and privacy risks.

Strengths:

- Encoders significantly improved both robustness and accuracy in adversarial settings, validating their utility in high-risk applications.
- Differentially private training with pre-trained encoders achieved enhanced performance without compromising privacy guarantees, addressing a longstanding limitation in privacy-preserving AI.
- The efficiency gains in exact machine unlearning tasks underscored the practical scalability of this approach.

Limitations:

- Fine-tuning approaches, while effective in adversarial conditions, require higher computational resources compared to linear probing.
- The reliance on large-scale pre-trained encoders may pose challenges for resource-constrained environments or domains with limited public datasets.

Broader Implications for Information Security

The combined findings of these studies reflect a paradigm shift in the approach to GenAI security:

Holistic Defences:

- The methodologies demonstrate the importance of integrating multiple layers of defence, from input validation to advanced encoding and monitoring techniques.

Ethical Considerations:

- The exploration of dual-use risks underscores the ethical responsibilities associated with deploying advanced AI models.
- Striking a balance between innovation and security will require collaboration across technical, regulatory, and societal dimensions.

Adaptability and Scalability:

- The scalability of solutions like spotlighting and pre-trained encoders highlights their potential for widespread adoption, while future research must address resource efficiency and real-time applicability.

VI. Conclusion

The advancements outlined in these four studies underscore the critical importance of addressing security, privacy, and robustness in Generative AI (GenAI) systems. As these systems continue to permeate diverse applications, their vulnerabilities to adversarial manipulation, dual-use risks, and privacy breaches become increasingly evident. This report highlights the significant progress achieved through innovative methodologies, each contributing uniquely to strengthening the safety and reliability of GenAI systems.

Spotlighting for Indirect Prompt Injection Attacks introduced a scalable and effective defence mechanism that leverages techniques like delimiting, datamarking, and encoding. This approach demonstrated significant success in mitigating indirect prompt injection attacks while maintaining task performance, emphasizing its practicality for real-world deployment.

The Multimodal Large Language Model Defence Framework addressed the unique challenges of securing multimodal systems and achieving high precision and recall in detecting adversarial inputs. Its two-stage framework, comprising input validation and prompt injection detection, has set a benchmark for safeguarding complex, multimodal GenAI architectures.

The study on Dual-Use Risks in Instruction-Following Models illuminated the vulnerabilities of programmatic behaviours in AI systems. By demonstrating the economic feasibility of malicious exploits, this research underscored the urgency of proactive and adaptive defences, as well as the need for ethical and regulatory oversight.

Leveraging Pre-Trained Encoders for Privacy-Preserving Learning offered a transformative approach to enhancing adversarial robustness and privacy guarantees. The efficiency and scalability achieved through self-supervised learning techniques highlight the potential for secure and privacy-conscious AI systems without sacrificing utility.

Together, these methodologies provide a comprehensive roadmap for securing GenAI systems. They emphasize the need for multi-layered defences, dynamic adaptability, and ethical considerations to ensure that AI technologies remain safe and beneficial.

As the landscape of threats continues to evolve, future research must build upon these findings to address emerging vulnerabilities, balance security with usability, and foster collaboration between academia, industry, and policymakers. These efforts will be pivotal in shaping the next generation of secure, trustworthy, and ethically aligned AI systems.

VII. Future Directions

The findings from the four studies not only address current challenges but also reveal opportunities for enhancing the security, privacy, and robustness of Generative AI (GenAI) systems. Future research must build on these methodologies, focusing on specific areas of improvement and innovation as identified in each study.

A. Advancing Spotlighting Techniques for Indirect Prompt Injection Attacks

Spotlighting proved effective in mitigating indirect prompt injection (IPI) attacks, but further enhancements are necessary to address evolving adversarial strategies.

Dynamic Spotlighting:

- Research should explore adaptive methods to dynamically adjust delimiting, datamarking, and encoding strategies based on the detected context and threat level.

Performance Optimization:

- Developing lightweight encoding techniques to reduce computational overhead while maintaining security will be critical for real-time applications.

Expansion Across Domains:

- Future work could extend spotlighting to domain-specific applications, such as healthcare and finance, where sensitive data interactions demand higher levels of security.

B. Enhancing Multimodal Large Language Model Defenses

The multimodal defense framework showed promising results, but further refinements are required to ensure broader adaptability and operational efficiency.

Automated Specification Generation:

- Incorporating AI-driven tools to automatically generate user specifications based on real-time usage patterns and evolving threats can enhance the framework's adaptability.

Latency Reduction:

- Optimizing the two-stage defense framework to process multimodal inputs faster, especially in high-frequency applications, will improve usability in time-sensitive contexts.

Cross-Modality Generalization:

- Extending the defense framework to handle a wider variety of modalities, such as audio and video, will ensure comprehensive protection for emerging multimodal systems.

C. Mitigating Dual-Use Risks in Instruction-Following Models

Addressing the dual-use risks highlighted in instruction-following models is critical to minimizing their exploitation for malicious purposes.

Context-Aware Filtering:

- Developing advanced filtering systems that analyze the intent and context of prompts in real-time can prevent the generation of harmful outputs without hindering legitimate use cases.

Adaptive Mitigations:

- Research into machine learning models capable of learning and countering new adversarial tactics as they emerge will be vital to staying ahead of evolving threats.

Collaborative Governance Models:

- Establishing interdisciplinary collaborations between researchers, policymakers, and industry stakeholders to define ethical standards and legal frameworks for deploying instruction-following models.

D. Expanding the Utility of Pre-Trained Encoders for Privacy-Preserving Learning

Pre-trained encoders demonstrated significant improvements in both robustness and privacy, but further exploration is required to unlock their full potential.

Domain-Specific Encoders:

- Designing pre-trained encoders tailored to specific domains or industries (e.g., medical imaging, cybersecurity) can enhance their effectiveness in specialized applications.

Hybrid Training Techniques:

- Combining fine-tuning and linear probing approaches can strike a balance between efficiency and performance, particularly for resource-constrained environments.

Improved Adversarial Defenses:

- Exploring novel augmentation techniques during encoder pre-training could further enhance robustness against sophisticated adversarial attacks.

Scalable Privacy Guarantees:

- Research into new methods for integrating differential privacy mechanisms at scale, without significant accuracy trade-offs, will be key to enabling privacy-preserving AI on a broader scale.

E. Broader Implications for Future Research

The future of Generative AI security lies in adopting a holistic approach that integrates innovations from multiple domains:

Real-Time Threat Detection:

- Developing systems capable of dynamically detecting and mitigating threats in real-time, leveraging insights from spotlighting and multimodal frameworks.

Interdisciplinary Collaboration:

- Bridging gaps between technical, ethical, and regulatory domains to ensure that AI advancements align with societal values and legal standards.

Resource-Efficient Security:

- Prioritizing the development of lightweight, scalable solutions that maintain robustness without imposing significant computational or operational costs.

By addressing these specific directions, future research can build on the foundations established in these studies to create secure, trustworthy, and efficient GenAI systems that meet the growing demands of diverse applications and adversarial environments.

References

[1] K. Hines, G. Lopez, M. Hall, F. Zarfati, Y. Zunger, and E. Kıcıman, ”Defending Against Indirect Prompt Injection Attacks With Spotlighting,” *Microsoft Technical Report*, 2024.

[2] R. K. Sharma, V. Gupta, and D. Grossman, ”Defending Language Models Against Image-Based Prompt Attacks via User-Provided Specifications,” *IEEE Security and Privacy Workshops*, 2024.

[3] D. Kang, X. Li, I. Stoica, C. Guestrin, M. Zaharia, and T. Hashimoto, ”Exploiting Programmatic Behavior of LLMs: Dual-Use Through Standard Security Attacks,” *arXiv preprint arXiv:2302.05733*, 2023.

[4] H. Liu, W. Qu, J. Jia, and N. Z. Gong, ”Pre-trained Encoders in Self-Supervised Learning Improve Secure and Privacy-preserving Supervised Learning,” *arXiv preprint arXiv:2212.03334*, 2022.