LLM Jailbreaking and System Vulnerabilities

Jean-Jacques Halans
strangelove.ai

Introduction

Large Language Models (LLMs) represent significant advancements in artificial intelligence, capable of understanding and generating human-like text. However, their widespread adoption has revealed critical vulnerabilities that can be exploited by attackers. This thesis explores the architectural and operational weaknesses of LLMs, their integration with external systems, and the safeguards in place, focusing on attacks like BoN Jailbreaking, Flowbreaking, and context-dependent vulnerabilities.

Key Vulnerabilities in LLM Architecture and Operation

Sensitivity to Input Variations

Despite their sophistication, LLMs exhibit a surprising sensitivity to minor input variations. These vulnerabilities are particularly pronounced in modalities like vision and audio, where subtle changes—such as alterations in image colour or audio pitch—can dramatically impact output. This sensitivity is foundational to attacks like **BoN Jailbreaking**, which repeatedly samples augmented versions of harmful requests until one bypasses the model's safeguards (Hughes et al., 2024) 3†source . The cross-modal effectiveness of this technique underscores the inherent fragility of LLMs in handling high-dimensional and continuous inputs.

Stochastic Output Generation

The stochastic nature of LLM output generation, especially at higher sampling temperatures, introduces another layer of vulnerability. While safety measures aim to prevent harmful responses, the randomness inherent in output generation can occasionally result in the production of unsafe content. By leveraging this unpredictability, BoN Jailbreaking increases the likelihood of eliciting harmful outputs through systematic augmentations (Hughes et al., 2024) 3†source . This challenge highlights the difficulty of safeguarding models that rely on probabilistic output generation.

Limitations of Alignment Techniques

Advances in alignment methodologies, such as reinforcement learning from human feedback (RLHF), have significantly improved the safety of LLMs. However, these techniques remain susceptible to adversarial attacks or "**jailbreaks**" (Robey et al., 2024) 2†source . Carefully crafted prompts can exploit alignment inconsistencies, prompting models to generate harmful or undesirable content. The persistence of these jailbreaks, even against commercial LLM systems, emphasises the need for more robust alignment strategies.

Vulnerabilities in System Architecture

The integration of LLMs into broader systems introduces additional attack vectors, as attackers can target weaknesses in the architecture and implementation. **Flowbreaking**, a novel class of attacks, exploits these systemic vulnerabilities by manipulating the interaction and synchronization of components. One example, the **Stop and Roll** attack, demonstrates how halting an LLM's response midway can bypass second-line guardrails, allowing harmful content to persist. This vulnerability underscores the necessity for holistic security measures that account for the entire system's architecture (Evron, 2024) 4†source .

Context-Dependent Alignment Challenges

The emergence of context-dependent alignment challenges, especially in LLM-controlled robotics, adds complexity to the problem of safeguarding AI systems. Unlike chatbots, which focus on filtering harmful text, robots operate in dynamic physical environments where intent and context significantly influence the potential for harm. For example, a command to "deliver a bomb" is harmful only if the robot possesses such an item. Addressing this issue requires sophisticated alignment mechanisms capable of reasoning about a robot's physical state and surroundings, further complicating defense strategies (Robey et al., 2024) 2†source .

Differentiating Attack Types: Prompt Injection, Jailbreaking, and Flowbreaking

Prompt Injection Attacks:

These attacks exploit vulnerabilities in applications built on top of LLMs. By crafting malicious input concatenated with a trusted prompt, attackers manipulate the LLM into performing unintended actions. **Prompt injection** remains a prevalent issue in the design of user-facing AI systems (IEEE Spectrum, 2024) 1†source .

Jailbreaking Attacks:

Jailbreaking focuses on bypassing the safety filters embedded within LLMs. Attackers design prompts that exploit alignment loopholes or inconsistencies, enabling the generation of harmful content. These attacks highlight the fragility of current alignment techniques and their limitations in preventing exploitation (Robey et al., 2024) 2†source .

Flowbreaking Attacks:

Flowbreaking attacks extend beyond the LLM to target the surrounding system architecture. By exploiting *timing issues*, *synchronisation weaknesses,* or *operational flaws*, these attacks disrupt data flow and application logic. For instance, the **Second Thoughts** attack manipulates response timing, allowing harmful information to leak before guardrails retract the response (Evron, 2024) 4†source . This broader scope of attack makes **Flowbreaking** particularly dangerous, as it targets not just the model but the entire application environment.

Findings and Implications for LLM Security

The findings presented in the sources have significant implications for LLM security and deployment, particularly highlighting the potential for real-world harm stemming from jailbroken LLMs.

- **Jailbroken LLMs pose a critical risk beyond generating harmful text; they can potentially cause physical harm in the real world.** This is especially concerning as many LLM-robot systems are currently deployed in safety-critical applications. One study demonstrated that an automated attack called RoboPAIR successfully jailbroke three different LLM-controlled robots, manipulating them into performing dangerous tasks such as colliding with pedestrians or searching for locations to detonate bombs (Robey et al., 2024) 2†source .

- **LLM-controlled robots may be fundamentally unaligned, even for non-adversarial inputs.** Unlike chatbots, where generating harmful text is generally viewed as objectively harmful, the harmfulness of a robotic action is context-dependent. This makes it difficult to establish clear safety guidelines and necessitates the development of new, robot-specific filters and defense mechanisms (Robey et al., 2024) 2†source .

- **The stochastic nature of LLM outputs and their sensitivity to input variations make them vulnerable even to simple attack algorithms.** The Best-of-N (BoN) jailbreaking algorithm successfully jailbroke a range of frontier LLMs across multiple modalities (text, vision, and audio) by repeatedly sampling augmented prompts until a harmful response was elicited (Hughes et al., 2024) 3†source . This highlights the need for robust defenses that can withstand a va-

riety of attacks, including those that exploit seemingly innocuous changes to inputs.

- **The current reliance on streaming responses in LLM applications poses a security risk, as harmful information may be transmitted before guardrails can effectively intervene.** This underscores the need for enterprises to ensure that LLM answers are fully generated before being displayed to users, despite potential user experience challenges (IEEE Spectrum, 2024) 1†source .

Recommendations for Holistic LLM Security

1. **Develop context-aware alignment mechanisms for LLM-controlled robots:** This will require considering the robot's environment and the potential consequences of its actions. Specifically, this involves creating advanced reasoning systems that can dynamically assess the intent of a command, evaluate the physical context, and predict potential outcomes before execution. For example, mechanisms could incorporate real-time sensory data and environmental modeling to understand whether a requested action, such as navigating a crowded space, could pose risks to human safety. Integrating these considerations will enable robots to operate ethically and adaptively in diverse, complex scenarios.

2. **Design robust defences specifically for LLM-controlled robots:** These defences should address the unique challenges posed by physical embodiment and context-dependent harm. Robust defence strategies must include real-time anomaly detection systems capable of identifying unexpected robotic behaviours, adaptive safety protocols that dynamically update based on environmental inputs, and multi-layered fail-safes to prevent harm even in the event of system compromise. Furthermore, collaboration with domain experts to tailor defences for specific robotic applications (e.g., autonomous vehicles, medical robots) is critical. These measures will significantly enhance the resilience and operational safety of LLM-controlled robots (Robey et al., 2024) 2†source .

3. **Investigate and address the sensitivity of LLMs to input variations:** This may involve exploring new defence mechanisms such as *input smoothing*, *adversarial training*, or employing robust *gradient masking* techniques to reduce susceptibility to perturbations (Hughes et al., 2024) 3†source . Additionally, creating multi-modal training datasets with diverse and noisy inputs can help models generalise better and resist targeted manipulations. Evaluations using systematic stress testing across multiple modalities (text, vision, audio) are essential for identifying specific vulnerabilities and tailoring defences to the dynamic nature of real-world data.

4. **Implement safeguards to prevent the premature release of harmful information in streaming LLM applications:** This may include delaying the display of responses until they are fully generated and vetted (IEEE Spectrum, 2024) 1†source . Furthermore, integrating multi-tiered content verification systems that analyse the response at various stages of generation could help identify and mitigate harmful outputs more effectively. Enterprises could adopt real-time monitoring tools to dynamically assess responses for safety violations and reinforce guardrails before final outputs are delivered to users. This approach ensures that user experience challenges are balanced with robust safety mechanisms, providing both functionality and security.

The rapid integration of LLMs into various societal and industrial domains necessitates a proactive approach to addressing their vulnerabilities. Key findings, such as the potential for real-world harm from RoboPAIR jailbreaks or the exploitability of stochastic outputs, highlight the pressing need for

robust defences. By adopting comprehensive alignment strategies, enhancing system-level safeguards, and ensuring context-aware mechanisms for robots, we can mitigate risks and unlock the transformative potential of LLMs responsibly.

Future research must focus on interdisciplinary collaboration and ongoing stress testing to adapt to evolving threats. Together, these measures will help protect against the challenges posed by advanced AI systems while ensuring their safe and ethical deployment.

1† "Robot Jailbreak: Researchers Trick Bots Into Dangerous Tasks", IEEE Spectrum, https://spectrum.ieee.org/jailbreak-llm

2† "Jailbreaking LLM-Controlled Robots", Alexander Robey, Zachary Ravichandran, Vijay Kumar, Hamed Hassani, George J. Pappas, School of Engineering and Applied Science, University of Pennsylvania. arXiv:2410.13691v2 [cs.RO] 9 Nov 2024

3† "BEST-OF-N JAILBREAKING", John Hughes, Sara Price, Aengus Lynch, Rylan Schaeffer, Fazl Barez, Sanmi Koyejo, Henry Sleight, Erik Jones, Ethan Perez, Mrinank Sharma, arXiv:2412.03556v2 [cs.CL] 19 Dec 2024

4† Suicide Bot: New AI Attack Causes LLM to Provide Potential "Self-Harm" Instructions , Gadi Evron, Knostic https://www.knostic.ai/blog/introducing-a-new-class-of-ai-attacks-flowbreaking