# Policy Puppetry exposes fundamental flaws in AI safety despite RLHF safeguards



Generative AI vendors have long asserted that methodologies such as Reinforcement Learning from Human Feedback (RLHF) guarantee that large language models (LLMs) adhere to established safety protocols. However, recent research conducted by HiddenLayer raises significant concerns about the validity of this assurance. Revealed through a systematic investigation, the study introduces a technique known as “Policy Puppetry,” which reportedly has the capacity to manipulate a wide array of major LLMs irrespective of their vendor or underlying architecture.

According to the findings, this bypass technique can effectively reframe harmful prompts into formats resembling XML or JSON, thereby misleading the models into interpreting these prompts as legitimate operational commands. The research further illustrates that the use of innovative tactics—such as leetspeak (a form of internet language) and elaborate roleplay scenarios—facilitates the evasion of built-in safety detection mechanisms. As a result, models can be compelled to comply with instructions that would ordinarily be blocked.

The implications of this research are broad and troubling. HiddenLayer's investigations revealed that a single sophisticated prompt could circumvent security measures implemented in various prominent AI models, including OpenAI’s ChatGPT (versions o1 through 4o), Google’s Gemini, Anthropic’s Claude, Microsoft’s Copilot, Meta’s LLaMA 3 and 4, among several others. Even the newest models, which are thought to deploy advanced reasoning safeguards, displayed vulnerabilities that could be exploited with minimal adjustments. Notably, fictional scenarios—like narratives involving dangerous activities portrayed in TV dramas—enabled attackers to blur the lines between fiction and instructions, which can confuse the models and prompt aberrant behaviours.

In an interview, Jason Martin, director of adversarial research at HiddenLayer, remarked that the vulnerabilities identified are deeply rooted in the training data of the models. He indicated that simple software patches are unlikely to resolve these issues, as the fundamental weaknesses reside within the architecture and training methodologies themselves. Malcolm Harkins, chief trust and security officer at HiddenLayer, cautioned that the ramifications of these vulnerabilities extend beyond mere digital disturbances. Industries critical to society, including healthcare, finance, manufacturing, and aviation, are at risk; compromised AI systems within these sectors could pose significant threats to public safety and security.

The study underscores that RLHF, often promoted as a robust safeguard, does not provide an infallible layer of protection. The researchers argue that AI systems can still be destabilised structurally, even when they appear to be aligned with safety guidelines.

In response to the findings, HiddenLayer is advocating for a paradigm shift in AI security. The team suggests integrating real-time AI monitoring platforms—such as AISec and AIDR—designed to detect and mitigate prompt injections and unsafe behaviours proactively. This approach mirrors the principles of zero-trust security typical in enterprise IT, suggesting that as AI systems become increasingly embedded in critical infrastructures, a more continuous and intelligent defense mechanism is necessary to thwart potential threats.

This ongoing discourse highlights the urgent need for a reassessment of the existing strategies for securing AI models, moving beyond reliance on alignment and towards proactive and sophisticated monitoring solutions.

Source: [Noah Wire Services](https://www.noahwire.com)

## Bibliography

1. <https://arxiv.org/abs/2308.06463> - This study demonstrates that ciphers can bypass safety alignment techniques in large language models, highlighting vulnerabilities in models like GPT-4.
2. <https://arxiv.org/abs/2409.14177> - Introduces 'PathSeeker,' a reinforcement learning-based approach that exploits LLM security vulnerabilities, achieving high attack success rates across various models.
3. <https://arxiv.org/abs/2502.11054> - Presents 'Reasoning-Augmented Conversation,' a multi-turn jailbreak framework that reformulates harmful queries into benign tasks, compromising safety alignment in LLMs.
4. <https://arxiv.org/abs/2502.20968> - Assesses role-play fine-tuning risks in LLMs, revealing that such techniques can degrade safety performance, especially for villainous characters.
5. <https://nvd.nist.gov/vuln/detail/CVE-2024-37058> - Details a vulnerability in MLflow versions 2.5.0 and newer, where deserialization of untrusted data can lead to arbitrary code execution, highlighting security risks in AI systems.
6. <https://hiddenlayer.com/research/risks-related-to-the-use-of-ai/> - Discusses various AI-related risks, including the creation of deepfakes and privacy concerns, emphasizing the need for robust security measures in AI deployments.
7. <https://news.google.com/rss/articles/CBMijAFBVV95cUxQLW5qal9OUmNNdHRGYjZkSVk2X1ZhNXA0dEpnYUpYV2xTNndnSmVjbEFLeFNzanB6M29pSmVDVE95S3p1OGk5eUlkaHpMTy1qM0JiSzNINGc0SGtWX0tkV0JhbmJHbF9pMExsclVhX3FnNFlGN2Q5V3Z3MEk5QUVRYXFXWC1mMGpYTVBUYw?oc=5&hl=en-US&gl=US&ceid=US:en> - Please view link - unable to able to access data