Dark LLMs: The Growing Threat of Unaligned AI Models

Michael Fire[1]*, Yitzhak Elbazis[1], Adi Wasenstein[1], Lior Rokach[1]*

Abstract
Large Language Models (LLMs) rapidly reshape modern life, advancing fields from healthcare to education and beyond. However, alongside their remarkable capabilities lies a significant threat: the susceptibility of these models to jailbreaking. The fundamental vulnerability of LLMs to jailbreak attacks stems from the very data they learn from. As long as this training data includes unfiltered, problematic, or 'dark' content, the models can inherently learn undesirable patterns or weaknesses that allow users to circumvent their intended safety controls. Our research identifies the growing threat posed by dark LLMs—models deliberately designed without ethical guardrails or modified through jailbreak techniques. In our research, we uncovered a universal jailbreak attack that effectively compromises multiple state-of-the-art models, enabling them to answer almost any question and produce harmful outputs upon request. The main idea of our attack was published online over seven months ago. However, many of the tested LLMs were still vulnerable to this attack. Despite our responsible disclosure efforts, responses from major LLM providers were often inadequate, highlighting a concerning gap in industry practices regarding AI safety. As model training becomes more accessible and cheaper, and as open-source LLMs proliferate, the risk of widespread misuse escalates. Without decisive intervention, LLMs may continue democratizing access to dangerous knowledge, posing greater risks than anticipated.

Keywords
Safe AI — Large Language Model (LLM) – Jailbreak — Dark LLM

[1] *Ben Gurion University of the Negev*
***Corresponding authors**: mickyfi@bgu.ac.il and liorrk@bgu.ac.il

The Dual-Use Challenge

Large Language Models (LLMs) have rapidly become embedded in modern society, used by over a billion people,[1] accelerating discovery, democratizing knowledge, and enabling new forms of creativity. From helping researchers translate rare languages [4] to personalized medicine [5], their positive impact is clear. However, these same models, trained on vast data, which, despite curation efforts, can still absorb dangerous knowledge, including instructions for bomb-making, money laundering, hacking, and performing insider trading [6]. While commercial LLMs incorporate safety mechanisms to block harmful outputs, these safeguards are increasingly proving insufficient. A critical vulnerability lies in jailbreaking—a technique that uses carefully crafted prompts to bypass safety filters, enabling the model to generate restricted content [7].

The Rise of Dark LLMs

Recently, a disturbing trend has gained momentum: the release of deliberately unaligned models, often described as "dark LLMs" [8, 9]. Variants such as WormGPT and FraudGPT are openly advertised online for having "no ethical guardrails" and for their willingness to assist in cybercrime, fraud, and more [9]. These models, alongside open-source systems like Llama[2] and DeepSeek[3], can be jailbroken to remove restrictions [10, 11, 12]. As model training becomes cheaper and hardware requirements diminish [13], powerful LLMs may become more accessible to individuals with malicious intent. Even in the mid of 2023, there were already more than 15,800 LLMs available on platforms like Hugging Face [14], reflecting the rapid

---

[1]It is estimated that over a billion people worldwide use LLMs. OpenAI's ChatGPT has over 800 million weekly users [1], Meta's Llama models have been downloaded 650 million times [2], and Baidu's Ernie Bot has 200 million users [3].

[2]https://www.llama.com

[3]https://www.deepseek.com

proliferation of these models. What was once restricted to state actors or organized crime groups may soon be in the hands of anyone with a laptop or even a mobile phone.

Jailbreaking: Unlocking Forbidden Knowledge

Even carefully aligned LLMs are vulnerable to manipulation. Through a technique known as jailbreaking, attackers craft adversarial prompts that bypass safety filters, forcing models that cost tens of millions to create, like ChatGPT[4] and Gemini,[5] to output restricted content [15, 16, 12]. An entire ecosystem has emerged around the creation and distribution of jailbreak prompts; for instance, the ChatGPT Jailbreak subreddit alone has amassed approximately 141,000 users, referred as Jailbreakers.[6] Alarmingly, recent research has demonstrated that even simple character sequences can successfully bypass safeguards in multiple leading models simultaneously [16]. Moreover, a recent study from April 2025, introduced a novel universal jailbreak attack capable of bypassing protections in a wide range of LLMs, including advanced reasoning models [12]. As the market for jailbreak techniques continues to expand, the potential to weaponize LLMs is no longer a theoretical risk–it is a tangible reality, easily accessible to those who seek it, even young kids and teenagers.

A Glimpse Into the Dark Potential

Our research began by investigating the real-world implications of LLM jailbreak attacks and evaluating the defense mechanisms embedded within commercial models. We started with a publicly known jailbreak method, published over seven months ago on Reddit. Surprisingly, many of the leading LLMs we tested, including state-of-the-art commercial systems, remained vulnerable to this widely disseminated attack. We developed a more comprehensive universal jailbreak attack based on this foundational exploit. This method proved to be highly effective, successfully bypassing safety filters in nearly all the LLMs we evaluated. Once compromised, the models consistently generated responses to virtually any query, including those involving illicit and harmful activities. Disturbingly, the LLMs themselves offered examples of illegal activities spanning various domains, often accompanied by detailed, step-by-step instructions. To responsibly disclose this vulnerability, we contacted several leading LLM providers via official channels, including bug bounty programs and direct communication. However, the response was underwhelming. Several companies did not respond at all, while others indicated that such vulnerabilities fell outside the scope of their bounty programs, suggesting we report the issue through alternative channels instead. These findings expose a critical weakness in the current approach to LLM security: even when vulnerabilities are well-documented and actively exploited in public forums, major providers often fail to respond adequately. The ease with which these LLMs can be manipulated to produce harmful content underscores the urgent need for robust safeguards. The risk is not speculative—it is immediate, tangible, and deeply concerning, highlighting the fragile state of AI safety in the face of rapidly evolving jailbreak techniques.

The Irreversibility of Open-Source Leaks

Unlike centrally managed platforms like ChatGPT or Gemini, open-source LLMs cannot be patched once vulnerabilities are discovered. Once an uncensored version is shared online, it is archived, copied, and distributed beyond control. No company, no update cycle, and no regulation can erase a locally saved model from a laptop or private server. Moreover, attackers can chain models together—using one model to generate jailbreak prompts for another—compounding the risk [17].

What Can Be Done?

LLM providers must actively work to patch vulnerabilities and jailbreak techniques as soon as they become known. Containing the threat of dark LLMs requires layered, proactive defenses. Key strategies include:

- **Training Data Curation** - Models should be trained on curated datasets that deliberately exclude harmful content, such as bomb-making instructions, money laundering guides, and extremist manifestos. Leveraging AI-driven content screening during pretraining can significantly enhance this process. Just as we protect children from unfiltered content on TV or the internet, we should also ensure that LLMs are not exposed to dark and dangerous material.

- **LLM Firewalls** - Middleware can intercept prompts and outputs, acting as a real-time safeguard between users and the model. Robust LLM firewalls should become a standard part of any deployment, just as antivirus software became ubiquitous for computers. Notably, IBM offers *Granite Guardian*, a suite of models designed to detect risks in prompts and responses, ensuring safe and responsible use of large language models [18]. Similarly, Meta provides *Llama Guard*,

---

[4]https://chatgpt.com
[5]https://gemini.google.com/app
[6]https://www.reddit.com/r/ChatGPTJailbreak/

an open-source guardrail system aimed at building secure AI agents by detecting and mitigating harmful or inappropriate content generation [19].

- **Machine Unlearning** - Recent advances allow models to "forget" specific types of content after deployment, without full retraining [20]. If perfected, machine unlearning could enable rapid removal of dangerous capabilities from already-released models.

- **Continuous Red Teaming** - Developers should maintain active adversarial testing teams, publish red-team performance benchmarks, and offer bug bounties for vulnerability discovery.

- **Public Awareness** - Governments, educators, and civil society must treat unaligned LLMs as serious security risks, comparable to unlicensed weaponry or explosives guides. Restricting casual access, especially for minors, should be a policy priority.

Conclusion: The Clock Is Ticking

LLMs are one of the most consequential technologies of our time. Their potential for good is immense—but so is their capacity for harm if left unchecked. Unchecked, dark LLMs could democratize access to dangerous knowledge at an unprecedented scale, empowering criminals and extremists across the world. It is not enough to celebrate the promise of AI innovation. Without decisive intervention—technical, regulatory, and societal—we risk unleashing a future where the same tools that heal, teach, and inspire can just as easily destroy. The choice remains ours. But time is running out.

Acknowledgments

References

[1] Digital Watch Observatory. Chatgpt hits 800 million users after viral surge, April 2025. Accessed: 2025-05-13.

[2] Reuters. Meta's llama ai model adoption. `https://ai.meta.com/blog/future-of-ai-built-with-llama/`, 2025. Accessed: 2025-05-13.

[3] Wall Street Journal. Baidu's ernie bot user base. `https://www.reuters.com/technology/baidu-says-ai-chatbot-ernie-bot-has-amassed-200-million-users-2024-04-16/`, 2025. Accessed: 2025-05-13.

[4] Kexun Zhang, Yee Man Choi, Zhenqiao Song, Taiqi He, William Yang Wang, and Lei Li. Hire a linguist!: Learning endangered languages with in-context linguistic descriptions. *arXiv preprint arXiv:2402.18025*, 2024.

[5] Armin Berger, David Berghaus, Ali Hamza Bashir, Lorenz Grigull, Lara Fendrich, Tom Anglim Lagones, Henriette Högl, Gundula Ernst, Ralf Schmidt, David Bascom, et al. Advancing personalized medicine: A scalable llm-based recommender system for patient matching. In *2024 IEEE International Conference on Big Data (BigData)*, pages 5876–5883. IEEE, 2024.

[6] Nathalie Maria Kirch, Severin Field, and Stephen Casper. What features in prompts jailbreak llms? investigating the mechanisms behind attacks. *arXiv preprint arXiv:2411.03343*, 2024.

[7] Sibo Yi, Yule Liu, Zhen Sun, Tianshuo Cong, Xinlei He, Jiaxing Song, Ke Xu, and Qi Li. Jailbreak attacks and defenses against large language models: A survey. *arXiv preprint arXiv:2407.04295*, 2024.

[8] Kevin Poireault. The dark side of generative ai: Five malicious llms found on the dark web. `https://www.infosecurityeurope.com/en-gb/blog/threat-vectors/generative-ai-dark-web-bots.html`, 2023. Accessed: 2025-05-13.

[9] Zvelo. Malicious ai: The rise of dark llms. `https://zvelo.com/malicious-ai-the-rise-of-dark-llms/`, February 2024. Accessed: 2025-05-13.

[10] Liwei Jiang, Kavel Rao, Seungju Han, Allyson Ettinger, Faeze Brahman, Sachin Kumar, Niloofar Mireshghallah, Ximing Lu, Maarten Sap, Yejin Choi, et al. Wildteaming at scale: From in-the-wild jailbreaks to (adversarially) safer language models. *Advances in Neural Information Processing Systems*, 37:47094–47165, 2024.

[11] Dilip Bachwani. Deepseek failed over half of the jailbreak tests by qualys tota-lai. `https://blog.qualys.com/vulnerabilities-threat-research/2025/01/31/deepseek-failed-over-half-of-the-jailbreak-tests-by-qualys-totalai`, January 2025. Accessed: 2025-05-13.

[12] Conor McCauley, Kenneth Yeung, Jason Martin, and Kasimir Schulz. Novel universal bypass for all major llms. `https://hiddenlayer.com/innovation-hub/novel-universal-bypass-for-all-major-llms/`, April 2025. Accessed: 2025-05-13.

[13] Guido Appenzeller. Welcome to llmflation – llm inference cost is going down fast. *Andreessen Horowitz (a16z)*, November 2024.

[14] Sarah Gao and Andrew Kean Gao. On the origin of llms: An evolutionary tree and graph for 15,821 large language models. *arXiv preprint arXiv:2307.09793*, 2023.

[15] Financial Times. Hackers 'jailbreak' powerful ai models in global effort to highlight flaws. *Financial Times*, November 2024. Accessed: 2025-05-13.

[16] Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. Jailbreaking leading safety-aligned llms with simple adaptive attacks. *arXiv preprint arXiv:2404.02151*, 2024.

[17] Jeremy Kritz, Vaughn Robinson, Robert Vacareanu, Bijan Varjavand, Michael Choi, Bobby Gogov, Scale Red Team, Summer Yue, Willow E Primack, and Zifan Wang. Jailbreaking to jailbreak. *arXiv preprint arXiv:2502.09638*, 2025.

[18] Inkit Padhi, Manish Nagireddy, Giandomenico Cornacchia, Subhajit Chaudhury, Tejaswini Pedapati, Pierre Dognin, Keerthiram Murugesan, Erik Miehling, Martín Santillán Cooper, Kieran Fraser, et al. Granite guardian. *arXiv preprint arXiv:2412.07724*, 2024.

[19] Sahana Chennabasappa, Cyrus Nikolaidis, Daniel Song, David Molnar, Stephanie Ding, Shengye Wan, Spencer Whitman, Lauren Deason, Nicholas Doucette, Abraham Montilla, et al. Llamafirewall: An open source guardrail system for building secure ai agents. *arXiv preprint arXiv:2505.03574*, 2025.

[20] Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, et al. Rethinking machine unlearning for large language models. *Nature Machine Intelligence*, pages 1–14, 2025.