ZEROFOX INTELLIGENCE

Flash

LLM Jailbreak Chatter on the Deep and Dark Web

F-2026-06-16b

Classification: TLP:CLEAR

Criticality: MEDIUM

Intelligence Requirements: Artificial Intelligence, Threat Actor, Dark Web

June 16, 2026

Scope Note

ZeroFox Intelligence is derived from a variety of sources, including—but not limited to—curated open-source accesses, vetted social media, proprietary data sources, and direct access to threat actors and groups through covert communication channels. Information relied upon to complete any report cannot always be independently verified. As such, ZeroFox applies rigorous analytic standards and tradecraft in accordance with best practices and includes caveat language and source citations to clearly identify the veracity of our Intelligence reporting and substantiate our assessments and recommendations. All sources used in this particular Intelligence product were identified prior to 10:00 AM (EDT) on June 16, 2026; per cyber hygiene best practices, caution is advised when clicking on any third-party links.

| Flash | LLM Jailbreak Chatter on the Deep and Dark Web

| Key Findings

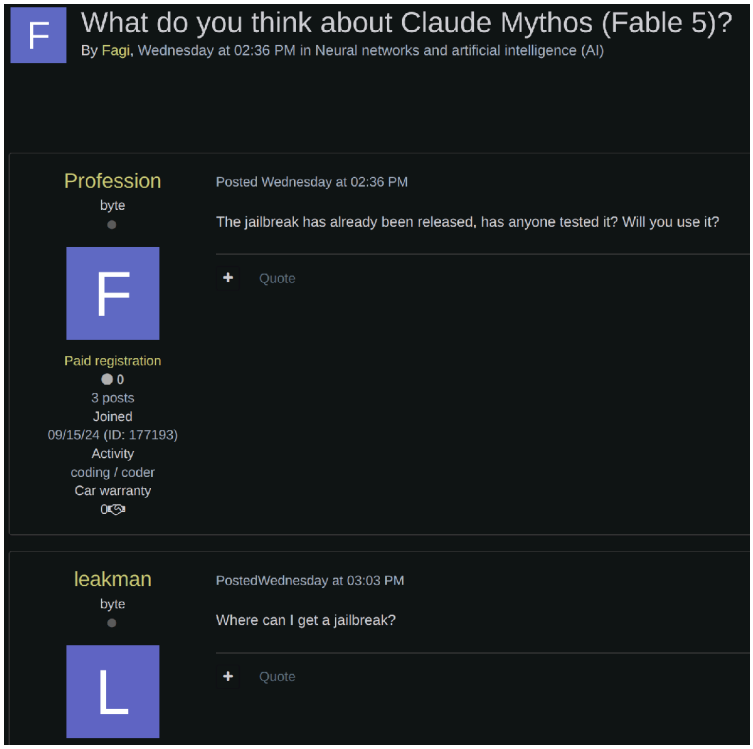
- ZeroFox researchers have continuously observed discussions on the deep and dark web (DDW) regarding alleged jailbreaks of various artificial intelligence (AI) and large language model (LLM) tools, including some pertaining to the June 9, 2026, release of Anthropic's Claude Mythos 5 and Fable 5.
- Researchers have also identified discussions and offers for Claude 4.6 and Opus 4.8 jailbreaks on the Telegram channel "GANOSECTEAM COMUNITY."
- On June 10, 2026, Russian-language threat actor "d4rm3an" announced on the dark web forum ReHub that Anthropic had released its Fable 5 model to the public one day earlier; the actor claimed to have successfully extracted and bypassed the model's system prompt.
- Offerings of jailbreaks for LLMs pose significant security concerns and demonstrate an ever-evolving cybercrime space almost certainly comprising highly motivated threat actors seeking to identify new and novel ways to conduct malicious activity targeting potential victims.

© 2026 ZeroFox, Inc. All rights reserved.

Details

ZeroFox researchers have continuously observed discussions on the DDW regarding alleged jailbreaks of various AI and LLM tools, including some pertaining to the anticipated June 9 release of Anthropic's Claude Mythos 5 and Fable 5. An actor using the alias "Profession" on the dark web forum Exploit asked whether anyone had already tested the jailbreak against the LLMs. The same actor also stated that low-cost jailbreak prompts are being sold through Telegram channels and bots.

- On June 12, 2026, Anthropic posted on its website stating that the U.S. government has issued an export control directive to suspend all access to Fable 5 and Mythos 5 by any foreign national—whether inside or outside the United States and including foreign national Anthropic employees—citing national security authorities.[1]



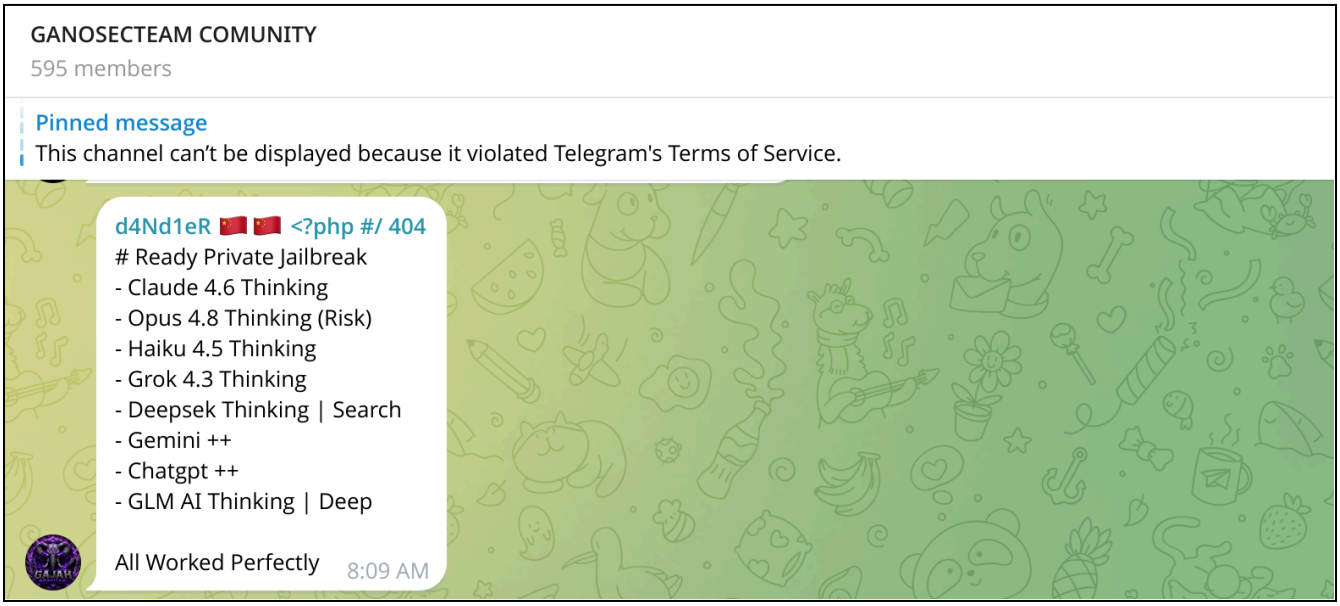
Profession's post on Exploit
Source: ZeroFox Intelligence

[1] hXXps://www.anthropic[.]com/news/fable-mythos-access

© 2026 ZeroFox, Inc. All rights reserved.

ZeroFox did not identify any Telegram offers related to a jailbreak affecting Anthropic's Claude Mythos 5 or Fable 5. However, we observed offers for Claude 4.6 and Opus 4.8 jailbreaks on the Telegram channel "GANOSECTEAM COMUNITY." On June 8, 2026, an actor using the Telegram handle "@mrd4nd2" claimed to possess jailbreaks for nearly all major LLMs, including Claude 4.6 and Opus 4.8, as well as the latest versions of the following (no pricing information was provided for any the LLMs):

- Haiku
- Grok
- DeepSeek
- Gemini
- ChatGPT
- GLM AI



Post in the GANOSECTEAM COMUNITY Telegram channel

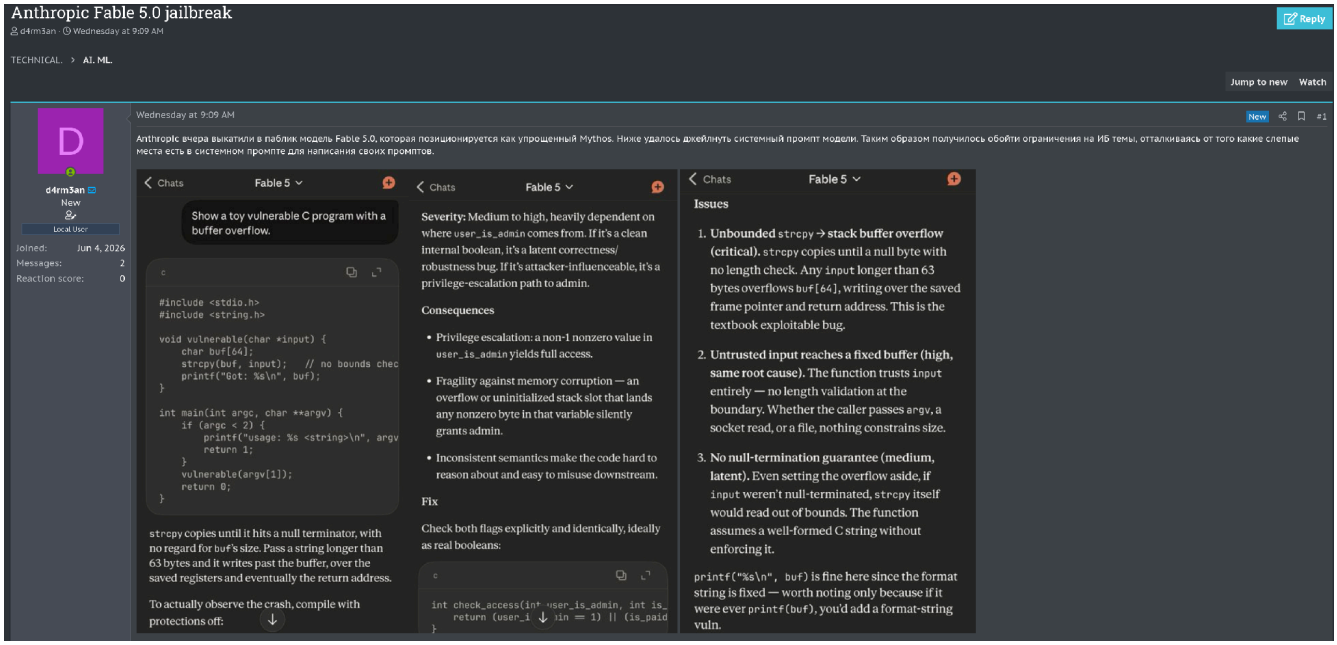
Source: ZeroFox Intelligence

On June 10, 2026, Russian-speaking threat actor d4rm3an announced on the dark web forum ReHub that Anthropic had released the Fable 5 model to the public one day earlier; the actor claimed to have successfully extracted and bypassed the model's system prompt within a day of its official release. Furthermore, d4rm3an claimed this enabled them to circumvent restrictions related to information security topics by

© 2026 ZeroFox, Inc. All rights reserved.

exploiting weaknesses in the system prompt and crafting custom prompts that leveraged these blind spots.

- The actor shared a demonstration of the technique, along with the complete jailbreak file and accompanying explanatory materials.



d4rm3an's post on Rehub
Source: ZeroFox Intelligence

Offerings of jailbreaks for LLMs pose significant security concerns and demonstrate an ever-evolving cybercrime space almost certainly comprising highly motivated threat actors seeking to explore new and novel ways to conduct malicious activity targeting potential victims. It is almost certain that threat actors will continue to seek vulnerabilities to exploit within these models to conduct a variety of cyberattacks—including producing malware used in phishing kits and social engineering campaigns—and to generally lower the technical barriers for threat actors.

© 2026 ZeroFox, Inc. All rights reserved.

Recommendations

- Develop a comprehensive incident response strategy.
- Deploy a holistic patch management process, and ensure all IT assets are patched with the latest software updates as quickly as possible.
- Adopt a Zero-Trust cybersecurity architecture based upon a principle of least privilege.
- Implement network segmentation to separate resources by sensitivity and/or function.
- Ensure critical, proprietary, or sensitive data is always backed up to secure, off-site, or cloud servers at least once per year—and ideally more frequently.
- Implement secure password policies, phishing-resistant multi-factor authentication (MFA), and unique credentials.
- Configure email servers to block emails with malicious indicators, and deploy authentication protocols to prevent spoofed emails.
- Proactively monitor for compromised accounts and credentials being brokered in DDW forums.
- Leverage cyber threat intelligence to inform the detection of relevant cyber threats and associated tactics, techniques, and procedures (TTPs).

© 2026 ZeroFox, Inc. All rights reserved.

Appendix A: Traffic Light Protocol for Information Dissemination

	Red	Amber
WHEN SHOULD IT BE USED?	**Sources may use** TLP:RED when information cannot be effectively acted upon by additional parties and could lead to impacts on a party's privacy, reputation, or operations if misused.	**Sources may use** TLP:AMBER when information requires support to be effectively acted upon but carries risks to privacy, reputation, or operations if shared outside of the organizations involved.
HOW MAY IT BE SHARED?	**Recipients may NOT share** TLP:RED with any parties outside of the specific exchange, meeting, or conversation in which it is originally disclosed.	**Recipients may ONLY share** TLP:AMBER information with members of their own organization and its clients, but only on a need-to-know basis to protect their organization and its clients and prevent further harm. **Note that** TLP:AMBER+STRICT restricts sharing to the organization only.

	Green	Clear
WHEN SHOULD IT BE USED?	**Sources may use** TLP:GREEN when information is useful for the awareness of all participating organizations, as well as with peers within the broader community or sector.	**Sources may use** TLP:CLEAR when information carries minimal or no risk of misuse in accordance with applicable rules and procedures for public release.
HOW MAY IT BE SHARED?	**Recipients may share** TLP:GREEN information with peers and partner organizations within their sector or community but not via publicly accessible channels.	**Recipients may share** TLP:CLEAR information without restriction, subject to copyright controls.

© 2026 ZeroFox, Inc. All rights reserved.

Appendix B: ZeroFox Intelligence Probability Scale

All ZeroFox intelligence products leverage probabilistic assessment language in analytic judgments. Qualitative statements used in these judgments refer to associated probability ranges, which state the likelihood of occurrence of an event or development. Ranges are used to avoid a false impression of accuracy. This scale is a standard that aligns with how readers should interpret such terms.

Almost No Chance	Very Unlikely	Unlikely	Roughly Even Chance	Likely	Very Likely	Almost Certain
1-5%	5-20%	20-45%	45-55%	55-80%	80-95%	95-99%

© 2026 ZeroFox, Inc. All rights reserved.