GPT-6 Astra Is the First OpenAI Model Rated ‘Critical’ for Cybersecurity

Kabui, Charles

2026-09-05

OpenAI released GPT-6 Astra on September 3 as its most intelligent and best-aligned model yet, and the first it has formally rated “Critical” for cybersecurity under its Preparedness Framework. With the right tools, Astra can find unknown security flaws and build working exploits against hardened systems with no one guiding each step. OpenAI reports a perfect 100% on ExploitBench, which turns known bugs into working attacks, and says Astra found

and used two zero-day flaws, bugs the vendor does not yet know about, in a single chain it is now disclosing. The model rolls out to defenders first through Daybreak, then to paid ChatGPT plans and the API, at $10 per million input tokens and $50 per million output.

The "Critical" label changes how a flagship model ships. Astra's most powerful cyber work is gated, starting with alpha testers and expanding to defenders through Daybreak Blue, while the broadly available model refuses 91.5% of cyber jailbreak attempts versus 59% for GPT-5.6 Sol. Defenders get a tool that finds and patches flaws before attackers weaponize them; everyone else gets a model tuned to catch vulnerabilities without chaining working exploits, the same tiering that already gates Gemini 3.8 Flash Cyber and Anthropic's restricted Mythos.

There is a trade-off. OpenAI's own safety review says Astra is harder to monitor than GPT-5.6 Sol: it controls its chain of thought, the reasoning trace that safety teams audit, more tightly, and can sometimes slip past the company's monitors in adversarial tests. TechCrunch links that opacity to a technique it calls opaque recurrence. OpenAI's most capable cyber model is now also its hardest to inspect.

Read More: OpenAI Daybreak vs. Anthropic Mythos: two rival visions for AI cybersecurity

Sources:

- Introducing GPT-6 Astra (OpenAI)
- Path to Astra: critical capabilities and frontier safeguards (OpenAI)
- Safety overview: GPT-6 Astra (OpenAI)
- OpenAI launches Astra, its powerful (and controversial) new model (TechCrunch)
