April 14, 2026 Security Safety

Trusted access for the next era of cyber defense

We continue to evolve trusted access, safeguards, and ecosystem support to help cyber defenders protect us all.

Listen to article 9:59 Share

We are scaling up our Trusted Access for Cyber (TAC) program to thousands of verified individual defenders and hundreds of teams responsible for defending critical software. For years, we've been building a cyber defense program on the principles of democratized access, iterative deployment, and ecosystem resilience. In preparation for increasingly more capable models from OpenAI over the next few months, we are fine-tuning our models specifically to enable defensive cybersecurity use cases, starting today with a variant of GPT-5.4 trained to be cyber-permissive: GPT-5.4-Cyber. In this post, we share how we expect our approach of scaling cyber defense in lockstep with increasing model capabilities to guide the testing and deployment of future releases.

The progressive use of AI accelerates defenders – those responsible for keeping systems, data, and users safe – enabling them to find and fix problems faster in the digital infrastructure everyone relies on. Similarly, AI is being used by attackers looking to cause harm. We've been preparing for this. Since 2023, we've supported defenders through our Cybersecurity Grant Program and strengthened safeguards through our Preparedness Framework. The same year, we started evaluating our models' cyber capabilities, and in 2025, we began including cyber-specific safeguards in our model deployments. Earlier this year, we furthered our support for defenders with the launch of Codex Security to identify and fix vulnerabilities at scale. Our approach to this continuous advancement of capabilities is guided by three principles:

- **Democratized access:** Our goal is to make these tools as widely available as possible while preventing misuse. We design mechanisms which avoid arbitrarily deciding who gets access for legitimate use and who doesn't. That means using clear, objective criteria and methods – such as strong KYC and identity verification – to guide who can access more advanced

available to legitimate actors large and small, including those responsible for protecting critical infrastructure, public services, and the digital systems people depend on every day.

- **Iterative deployment:** We learn the most by putting these systems into the world carefully and improving them over time. As we better understand both their capabilities and risks, we update our models and safety systems accordingly. This includes understanding the differentiated benefits and risks of specific models, improving resilience to jailbreaks and other adversarial attacks, and improving defensive capabilities — while mitigating harms.
- **Investing in ecosystem resilience:** We support and accelerate the community of defenders through trusted access pathways, targeted grants, contributions to open-source security initiatives, and technologies like Codex Security that help defenders more rapidly find and patch vulnerabilities.

Our strategy for cybersecurity resilience and defensive acceleration

For years, our cybersecurity strategy has been to invest in research, prevent misuse, and accelerate defenders. As model capabilities have advanced, we have expanded our programs toward these goals, which are grounded in the following convictions:

- **Cyber risk is already here and accelerating, but we can act.** Digital infrastructure has already been vulnerable for years, before advanced AI even came along. Now, existing models can help find vulnerabilities, reason across codebases, and support meaningful parts of the cyber workflow, and threat actors are experimenting with novel AI-driven approaches. We've seen sophisticated harnesses elicit stronger and stronger capabilities by using more test-time compute with existing models. That means safeguards cannot wait for a single future threshold.
- **Expand access based on who is using these systems and how they're being used.** Cyber capabilities are inherently dual-use, so risk isn't defined by the model alone. It also depends on the user, the trust signals around them, and the level of access they're given.
    - Broad access to general models with safeguards can coexist with more granular controls for higher-risk capabilities, supported by stronger verification, clearer signals of intent, and better visibility into use.
    - To enable responsible use at scale, we need systems that can validate trustworthy users and use cases in more automated and more objective ways. This allows us to expand access based on evidence and real signals of trust, rather than relying on manual decisions. We don't think it's practical or appropriate to centrally decide who gets to defend themselves. Instead, we aim to enable as many

- **Defenses should be continually scaled with capability.** As model capabilities increase, defenses need to scale alongside them. We've seen steady improvements in agentic coding, which have direct implications for cybersecurity and we've adapted our approach in step.
    - We began cyber-specific safety training with GPT-5.2, then expanded it with additional safeguards through GPT-5.3-Codex and GPT-5.4, where we also classified the model as "high" cyber capability under our Preparedness Framework. In parallel, we increased support for defenders: launching a $10M Cybersecurity Grant Program, reached over 1,000 open source projects with Codex for Open Source which provides free security scanning, and continued to improve Codex Security.
    - Codex Security, which launched in private beta six months ago, and as a research preview earlier this year, automatically monitors codebases, validates issues, and proposes fixes. As models have improved, so has the system's precision and usefulness. Since the recent launch, Codex Security has contributed to over 3,000 critical and high fixed vulnerabilities, along with many more lower-severity fixed findings across the ecosystem.
    - Across these releases, we've also refined how models handle sensitive requests, calibrating refusal boundaries while expanding trusted access through programs like TAC.
- **Software development itself must be made more secure.** The strongest ecosystem is one that continuously identifies, validates, and fixes security issues as software is written. By integrating advanced coding models and agentic capabilities into developer workflows, we can give developers immediate, actionable feedback while they are building, shifting security from episodic audits and static bug inventories to ongoing, tangible risk reduction.

Scaling Trusted Access for Cyber and GPT-5.4-Cyber

We want to empower defenders by giving broad access to frontier capabilities, including models which have been tailor-made for cybersecurity. In February, we introduced Trusted Access for Cyber (TAC) with both automated identity verification for individuals in order to reduce the friction of safeguards on cybersecurity-related tasks and partner with a limited set of organizations for more cyber-permissive models.

Today we're expanding this program by introducing additional tiers of access for users willing to work with OpenAI to authenticate

additional cyber capabilities and with fewer capability restrictions. This is a version of GPT-5.4 which lowers the refusal boundary for legitimate cybersecurity work and enables new capabilities for advanced defensive workflows, including binary reverse engineering capabilities that enable security professionals to analyze compiled software for malware potential, vulnerabilities and security robustness without needing access to its source code.

Because this model is more permissive, we are starting with a limited, iterative deployment to vetted security vendors, organizations, and researchers. Access to permissive and cyber-capable models may come with limitations, especially around no-visibility uses like Zero-Data Retention (ZDR). This is particularly true for developers and organizations accessing our models through third-party platforms where OpenAI may have less direct visibility into the user, the environment, or the purpose of the request.

Gaining access to TAC is straightforward:

- Individual users can verify their identity at chatgpt.com/cyber.
- Enterprises can request trusted access for their team through their OpenAI representative.

All customers approved through this process will gain access to versions of existing models with reduced friction around safeguards which might trigger on dual-use cyber activity, allowing them to continue to support security education, defensive programming, and responsible vulnerability research. Customers already in TAC willing to further authenticate themselves as legitimate cyber defenders can express interest in additional tiers of access, including requesting access to GPT-5.4-Cyber.

Looking ahead to our upcoming model release and beyond

Our cybersecurity defenses are the result of many months of iterative improvement. We believe the class of safeguards in use today sufficiently reduce cyber risk enough to support broad deployment of current models. We expect versions of these safeguards to be sufficient for upcoming more powerful models, while models explicitly trained and made more permissive for cybersecurity work require more restrictive deployments and appropriate controls.

Over the long term, to ensure the ongoing sufficiency of AI safety in cybersecurity, we also expect the need for more expansive defenses for future models, whose capabilities will rapidly exceed even the best purpose-built models of today.

Author

OpenAI

Keep reading

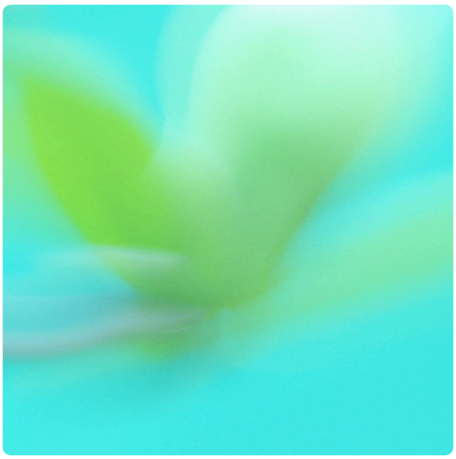
View all

Accelerating the cyber defense ecosystem that protects us all

Security Apr 16, 2026

Our response to the Axios developer tool compromise

Security Apr 10, 2026



Introducing the Child Safety Blueprint

Safety Apr 8, 2026

OpenAI

Pricing

API log in ↗

Documentation ↗

Developer Forum ↗

Stories

Academy

Livestreams

Podcast

RSS

English United States