OpenAI comenzó el viernes a desplegar Codex Security, un agente de seguridad impulsado por inteligencia artificial (IA) diseñado para detectar vulnerabilidades, validarlas y sugerir posibles soluciones.

La función está disponible como vista previa de investigación para clientes de ChatGPT Pro, Enterprise, Business y Edu a través de la web de Codex, con uso gratuito durante el próximo mes.

"Construye un contexto profundo sobre tu proyecto para identificar vulnerabilidades complejas que otras herramientas basadas en agentes no logran detectar, ofreciendo hallazgos con mayor nivel de confianza junto con correcciones que realmente fortalecen la seguridad del sistema, evitando además el ruido generado por errores poco relevantes," señaló la compañía.

Codex Security representa una evolución de Aardvark, una herramienta que OpenAI presentó en beta privada en octubre de 2025 con el objetivo de permitir a desarrolladores y equipos de seguridad identificar y corregir fallos de seguridad a gran escala.

Durante los últimos 30 días, Codex Security ha analizado más de 1.2 millones de commits en repositorios externos durante la fase beta, detectando 792 hallazgos críticos y 10,561 de alta severidad. Entre ellos se encuentran vulnerabilidades en distintos proyectos de código abierto como OpenSSH, GnuTLS, GOGS, Thorium, libssh, PHP y Chromium, entre otros. Algunos ejemplos se listan a continuación:

- GnuPG – CVE-2026-24881, CVE-2026-24882
- GnuTLS – CVE-2025-32988, CVE-2025-32989
- GOGS – CVE-2025-64175, CVE-2026-25242
- Thorium – CVE-2025-35430, CVE-2025-35431, CVE-2025-35432, CVE-2025-35433, CVE-2025-35434, CVE-2025-35435, CVE-2025-35436

Según la compañía de IA, esta nueva versión del agente de seguridad para aplicaciones aprovecha las capacidades de razonamiento de sus modelos más avanzados y las combina

con validación automatizada para reducir el riesgo de falsos positivos y proporcionar correcciones que puedan aplicarse de forma práctica.

Los análisis realizados por OpenAI sobre los mismos repositorios a lo largo del tiempo muestran un aumento progresivo en la precisión y una disminución en los falsos positivos, los cuales se han reducido en más del 50% en todos los repositorios evaluados.

En una declaración, OpenAI indicó que Codex Security fue diseñado para mejorar la relación señal-ruido al basar la detección de vulnerabilidades en el contexto del sistema y validar los hallazgos antes de presentarlos a los usuarios.

Concretamente, el agente opera en tres etapas: primero analiza el repositorio para comprender la estructura del sistema relevante para la seguridad del proyecto y genera un modelo de amenazas editable que describe su funcionamiento y los puntos donde podría estar más expuesto.

Una vez que se establece el contexto del sistema, Codex Security lo utiliza como base para detectar vulnerabilidades y clasificar los hallazgos según su impacto real. Posteriormente, los problemas identificados se someten a pruebas en un entorno aislado para confirmar su validez.

"Cuando Codex Security se configura con un entorno adaptado a tu proyecto, puede validar posibles problemas directamente en el contexto del sistema en ejecución," explicó OpenAI. *"Esta validación más profunda permite reducir aún más los falsos positivos y facilita la creación de pruebas de concepto funcionales, proporcionando a los equipos de seguridad evidencias más sólidas y una ruta más clara para su corrección."*

La etapa final consiste en que el agente proponga soluciones que se ajusten al comportamiento del sistema, con el objetivo de minimizar regresiones y facilitar su revisión e implementación.

La noticia sobre Codex Security llega pocas semanas después de que Anthropic lanzara

Claude Code Security, una herramienta destinada a ayudar a los usuarios a analizar bases de código en busca de vulnerabilidades y sugerir parches de corrección.