OpenAI anunció el viernes el lanzamiento de tres variantes de GPT-5.6, denominadas Sol, Terra y Luna, disponibles inicialmente como una vista previa limitada para un reducido grupo de empresas. Esta iniciativa forma parte de una colaboración continua con el gobierno de Estados Unidos.

Mientras que Sol representa el modelo insignia más avanzado y potente de la compañía, Terra busca ofrecer un equilibrio entre rendimiento y eficiencia, y Luna ha sido optimizado para proporcionar mayor velocidad y un menor costo de operación.

«GPT-5.6 Sol debuta con el conjunto de medidas de seguridad más sólido que hemos desarrollado hasta ahora. Reforzamos las protecciones frente a actividades de mayor riesgo, solicitudes cibernéticas sensibles y usos indebidos reiterados. Además, dedicamos varias semanas a identificar vulnerabilidades, someter el sistema a pruebas de resistencia y fortalecerlo frente a ataques del mundo real», señaló OpenAI.

La compañía también destacó que este modelo es el más competente que ha desarrollado hasta la fecha para tareas relacionadas con la ciberseguridad, lo que lo convierte en una herramienta más adecuada para la investigación y explotación controlada de vulnerabilidades. De acuerdo con OpenAI, en la prueba ExploitBench, GPT-5.6 Sol ofrece un rendimiento comparable al de Anthropic Mythos Preview utilizando aproximadamente un tercio de los tokens de salida.

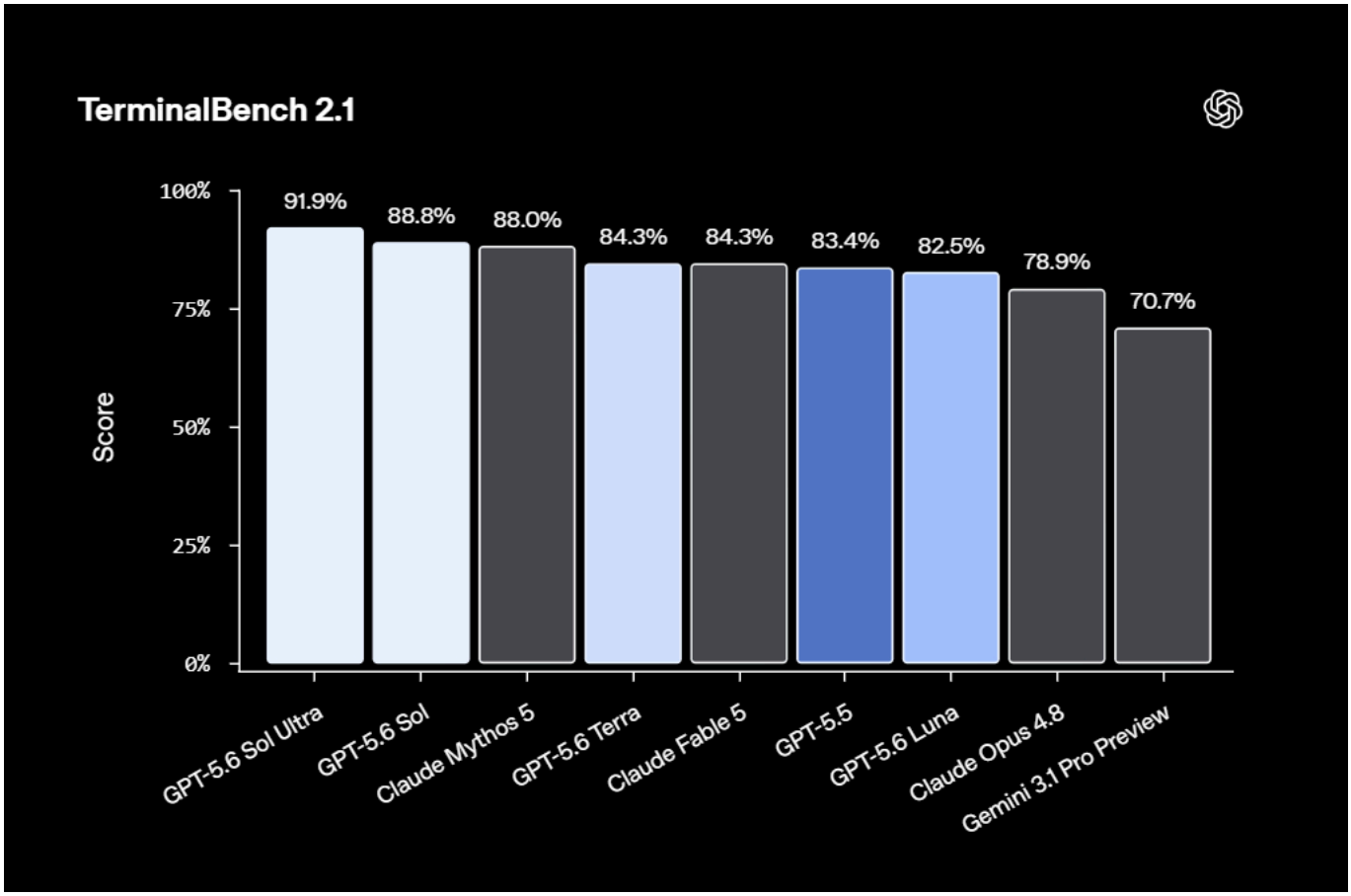
Según la empresa, el objetivo es facilitar el acceso a actividades legítimas como la revisión de código, la investigación de vulnerabilidades, el desarrollo de parches, la depuración de software, la capacitación en seguridad y las pruebas defensivas. Al mismo tiempo, mantiene estrictos mecanismos de protección para impedir acciones ofensivas y corregir con rapidez cualquier nuevo método de evasión de restricciones descubierto. Estas medidas abarcan intentos deliberados de vulnerar las salvaguardas del modelo y rechazan lo que OpenAI define como *«asistencia cibernética prohibida».*

«A medida que estas capacidades continúan evolucionando, nuestra prioridad es garantizar que lleguen a quienes defienden los sistemas, para que puedan identificar debilidades,

desarrollar correcciones y reforzar la seguridad de manera más amplia», explicó la empresa especializada en inteligencia artificial (IA).

No obstante, OpenAI advirtió que durante esta fase de vista previa algunos usuarios podrían encontrarse con mecanismos de seguridad que bloqueen o rechacen solicitudes legítimas, o incluso que las sometan a una revisión adicional. Esto se debe al carácter de «doble uso» que poseen estas tecnologías.

De acuerdo con la GPT-5.6 Preview System Card de OpenAI, aunque el modelo ha mejorado significativamente su capacidad para descubrir vulnerabilidades en el código y desarrollar exploits, todavía no posee la capacidad de ejecutar ataques completamente autónomos de principio a fin contra objetivos altamente protegidos, ni de convertir esas vulnerabilidades en armas listas para emplearse en ataques reales.



«Evaluaciones independientes analizaron comportamientos desalineados en tareas de programación con agentes y concluyeron que GPT-5.6 presenta una mayor tendencia que GPT-5.5 a exceder la intención del usuario, incluyendo la realización o el intento de realizar acciones que no fueron solicitadas, aunque la frecuencia absoluta de estos casos sigue siendo baja», indicó la compañía.

Asimismo, una evaluación realizada con VulnLMP —el marco interno de OpenAI diseñado para poner a prueba el desarrollo integral de cadenas de explotación contra objetivos reales— sobre proyectos de software ampliamente implementados y reforzados reveló que GPT-5.6 Sol es capaz de generar indicios creíbles relacionados con vulnerabilidades de seguridad de memoria. Algunos de estos hallazgos podrían derivar en problemas de divulgación de información, alteraciones de memoria o corrupción del flujo de ejecución.

«Esto indica que una parte considerable de la investigación de vulnerabilidades en entornos reales está cada vez más cerca de automatizarse cuando los modelos de IA trabajan junto con herramientas especializadas, sistemas de compilación e infraestructura de verificación», afirmó la empresa tecnológica.

OpenAI tiene previsto que GPT-5.6 Sol, Terra y Luna estén disponibles de forma general en las próximas semanas. Como parte de este proceso, la empresa presentó previamente las capacidades de estos modelos al gobierno estadounidense y habilitó un programa piloto para un grupo reducido de socios de confianza cuya participación fue autorizada por las autoridades antes de un despliegue más amplio.

A principios de este mes, el presidente de Estados Unidos, Donald Trump, firmó una orden ejecutiva sobre inteligencia artificial y ciberseguridad en la que solicitó la creación de un marco regulatorio que permita al gobierno federal evaluar las capacidades de los modelos de IA y determinar cuáles califican como *«modelos frontera cubiertos»*, una categoría destinada a los sistemas con capacidades cibernéticas avanzadas.

Este lanzamiento escalonado se produce pocos días después de que OpenAI presentara una versión mejorada de GPT-5.5-Cyber para un grupo de defensores de confianza dentro de la iniciativa Daybreak, además de anunciar el proyecto Patch the Planet en colaboración con Trail of Bits, cuyo propósito es fortalecer la seguridad de proyectos de código abierto.

La decisión también se produce tras la autorización del gobierno de Estados Unidos para que Anthropic volviera a poner a disposición su modelo Mythos AI a un grupo de aproximadamente 100 empresas de confianza y agencias federales encargadas de operar y proteger infraestructuras críticas. Esta autorización llegó más de dos semanas después de que los avanzados modelos orientados a la ciberseguridad fueran retirados temporalmente del mercado.

«Estamos restableciendo rápidamente el acceso para estas organizaciones y continuamos colaborando con el gobierno para ampliar la disponibilidad de Mythos 5 y volver a ofrecer Fable 5 para uso general», señaló Anthropic en un comunicado publicado en X.