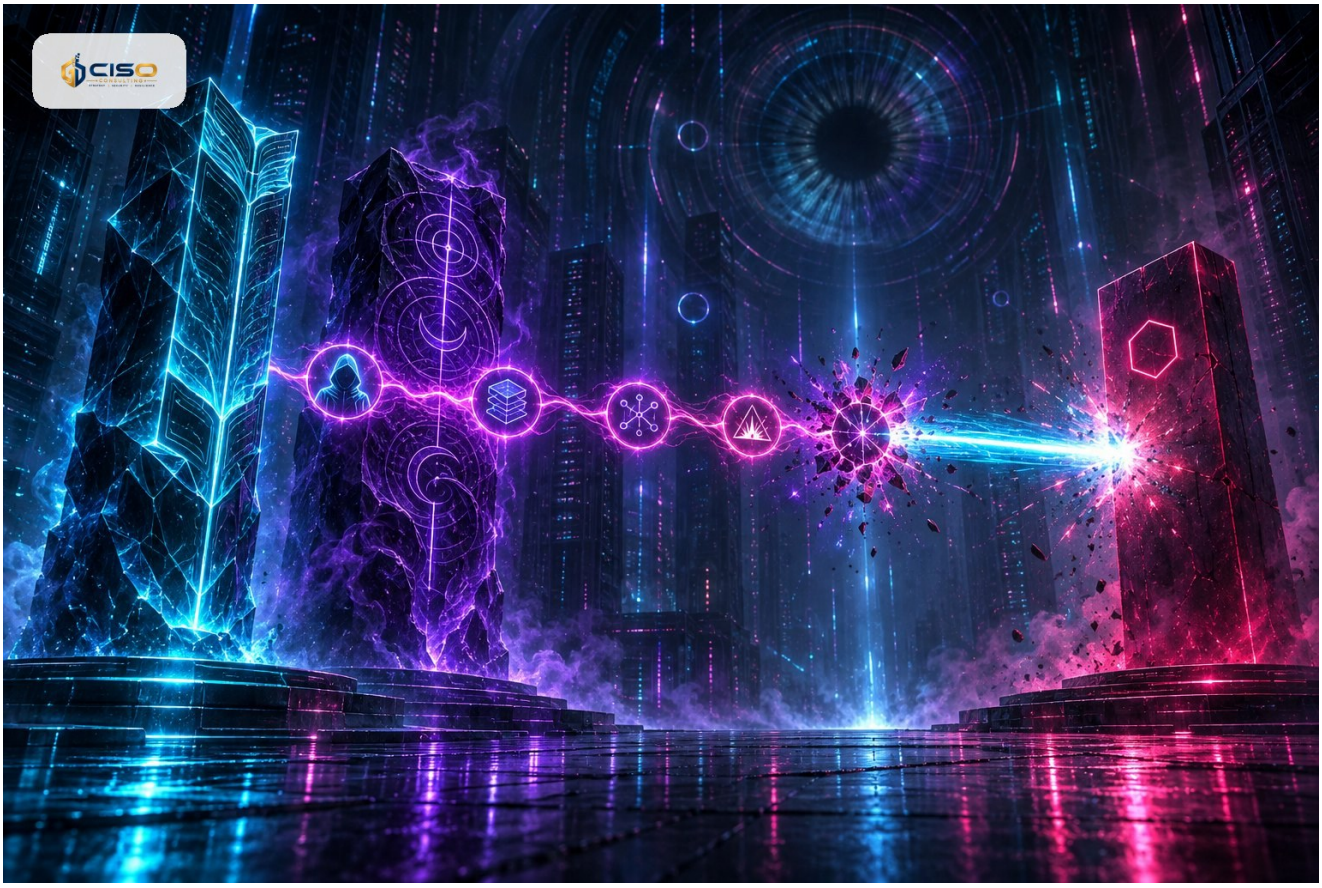
أنثروبيك تطلق نموذجي Claude Fable 5.2 وMythos 5.2 وسط ادعاءات بشأن اختبارات الهجمات السيبرانية



أطلقت أنثروبيك نموذجي Claude Fable 5.2 وClaude Mythos 5.2 مع تحسينات في البرمجة والاستدلال والرؤية. ويذكر عنوان التقرير تفوق Fable 5.2 على GPT-6 Cyber في اختبار هجمات متعددة المراحل دون تفاصيل.

- أطلقت أنثروبيك نموذجي Claude Fable 5.2 وClaude Mythos 5.2 يوم الاثنين.
- تقول الشركة إن النموذجين يتحسنان في البرمجة والاستدلال والرؤية.
- وفق الشركة، ساعد النموذجان مهندسيها خلال التدريب السابق للإطلاق على إنجاز عمليات ترحيل برمجيات يصعب إنجازها بغير ذلك.
- يفيد عنوان المصدر بأن Fable 5.2 تفوق على GPT-6 Cyber في اختبار هجمات سيبرانية متعددة المراحل.
- لا يوضح النص المتاح منهجية الاختبار أو النتائج أو أي حادثة بعينها.

أطلقت أنثروبيك نموذجين لغويين كبيرين جديدين هما Claude Fable 5.2 وClaude Mythos 5.2، وتقول الشركة إنهما يشهدان تحسناً في البرمجة والاستدلال والرؤية.

ما ورد في المصدر

وفق التقرير، ذكرت أنثروبيك أن النموذجين ساعدا مهندسيها خلال التدريب السابق للإطلاق على إنجاز عمليات ترحيل برمجيات كان يصعب إنجازها بغير ذلك، كما أسهما في أعمال تقنية أخرى. ويقول عنوان التقرير إن Claude Fable 5.2 تفوق على GPT-6 Cyber في اختبار معياري لهجمات سيبرانية متعددة المراحل.

لا يشرح المقتطف المتاح لنا كيف صُمم الاختبار، ولا ما الذي قيس فيه، ولا ما كانت النتائج. كما لا يصف أي هجوم على جهة فعلية. لذا ينبغي التعامل مع ما جاء في العنوان على أنه غير مؤكد إلى حين مراجعة التقرير الكامل أو وثائق الشركة المطورة.

ماذا يعني للجهات الخاضعة للتنظيم

هذا ليس ثغرة أمنية ولا تهديداً نشطاً، بل مؤشر على تزايد قدرة نماذج الذكاء الاصطناعي على المهام التقنية متعددة الخطوات، وهو ما يؤثر في المدافعين والمهاجمين على حد سواء. وعلى الجهات التي تستخدم الذكاء الاصطناعي التوليدي أو تقيّمه أن تُبقي ضوابط الحوكمة وإدارة مخاطر الأطراف الثالثة ومعالجة البيانات محدّثة مع ظهور إصدارات جديدة.

والاستجابة المناسبة هادئة: مراجعة المصادر الأصلية، وتجنب الاستنتاج من عنوان اختبار واحد، وإدراج المستجد ضمن تقييمات المخاطر القائمة.

ما يجب فعله الآن

1. قراءة التقرير الكامل ووثائق أنثروبيك نفسها قبل الحكم على ادعاء الاختبار المعياري.
2. حصر نماذج الذكاء الاصطناعي التوليدي وإصداراتها المستخدمة أو قيد التجربة في الجهة، بما في ذلك لدى الأطراف الثالثة.
3. اشتراط مراجعة أمنية ومراجعة لحماية البيانات قبل اعتماد أي إصدار جديد، بما يشمل معالجة البيانات وفق نظام حماية البيانات الشخصية.
4. أخذ تقنيات الهجوم المدعومة بذكاء اصطناعي أكثر قدرة في الحسبان ضمن نمذجة التهديدات وتخطيط اختبارات الفريق الأحمر.
5. تحديث سياسة الاستخدام المقبول للذكاء الاصطناعي لتوضيح الأدوات وأنواع البيانات المسموح بها للموظفين.