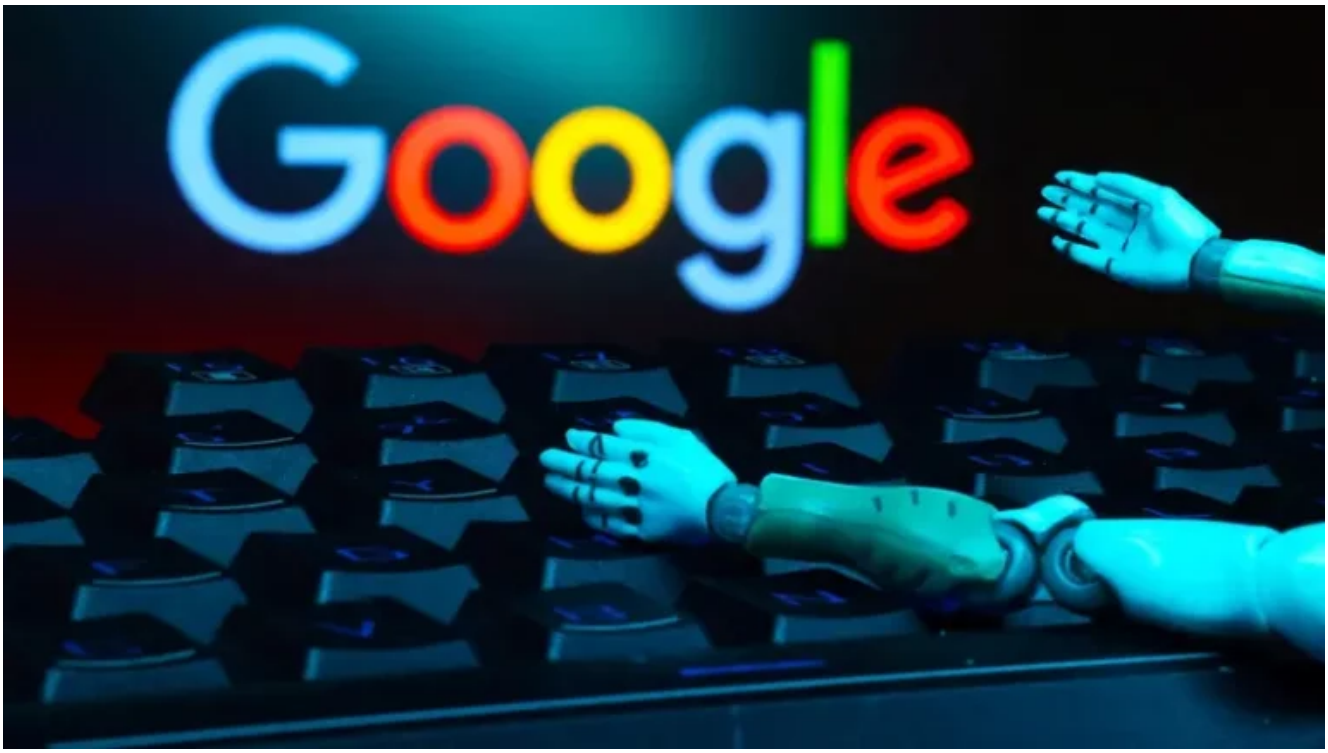
غوغل تعلن آليات جديدة لحماية مستخدمي كروم من مخاطر الوكلاء الذكيين

10 ديسمبر، 2025



مع توسّع شركات التكنولوجيا في تزويد متصفحات الويب بميزات "وكلاء" ذكية قادرة على تنفيذ مهام تلقائياً مثل حجز التذاكر وإتمام عمليات التسوّق، تتزايد المخاوف من مخاطر أمنية قد تؤدي إلى تسريب بيانات المستخدمين أو التسبب في خسائر مالية. وفي هذا السياق، كشفت شركة غوغل عن آلية حماية جديدة في متصفح كروم تهدف إلى ضبط سلوك هذه الوكلاء ومنعها من تجاوز الحدود المسموح بها.

وأوضحت غوغل، في توضيح تقني نشرته عقب استعراض قدرات هذه الوكلاء في أيلول الماضي، أن ميزات الذكاء الاصطناعي ستصل إلى المستخدمين خلال الأشهر المقبلة، لكنها ستخضع لعدة طبقات من التدقيق والمراقبة، وفق ما نقل موقع "تك كرانش".

وتستخدم الشركة مجموعة من النماذج لمراجعة كل خطوة يتخذها الوكيل الذكي، من بينها نموذج User Alignment Critic المبني على "جيميني"، والذي يقيّم المهام المقترحة من نموذج التخطيط. وفي حال تبيّن أن الخطوات المقترحة لا تتوافق مع هدف المستخدم، يطلب

النموذج إعادة صياغة الخطة. وتشير غوغل إلى أن هذا "الناقد" لا يطّلع على المحتوى الكامل للمواقع، بل يتعامل مع بيانات وصفية فقط، حفاظاً على الخصوصية.

ولمنع الوكلاء من الوصول إلى مواقع غير آمنة أو غير موثوقة، تعتمد غوغل تقنية Agent Origin Sets التي تحدد بدقة نطاق المواقع التي يُسمح للوكيل بقراءتها أو التفاعل معها، بحيث يمكنه مثلاً قراءة قوائم المنتجات في مواقع التسوق، من دون الاطلاع على الإعلانات أو العناصر غير ذات الصلة، مع تقييد قدرته على النقر والكتابة ضمن إطارات محددة في الصفحة. وتقول الشركة إن هذه القيود تقلل من احتمالات تسريب بيانات حساسة بين المواقع، وتمكّن المتصفح من منع إرسال أي معلومات إلى نماذج الذكاء الاصطناعي خارج النطاق المسموح.

كما تستخدم غوغل نموذجاً آخر لمراجعة عناوين الروابط التي يقترحها الوكيل، ما يساعد في منع انتقاله إلى مواقع ضارة أو عناوين جرى توليدها بشكل غير آمن. وفي المهام الحساسة، مثل دخول مواقع بنكية أو صفحات تحتوي على معلومات طبية، يشترط كروم الحصول على موافقة صريحة من المستخدم، وينطبق الأمر نفسه على المواقع التي تتطلّب تسجيل الدخول أو تنفيذ عمليات شراء أو إرسال رسائل. وأكدت الشركة أن نماذج الذكاء الاصطناعي لا يمكنها الوصول إلى كلمات المرور المحفوظة في "مدير كلمات المرور" داخل كروم.

إلى جانب ذلك، طوّرت غوغل مُصنِّفاً خاصاً لاكتشاف هجمات "حقن التعليمات" (Prompt Injection)، وتختبر أداء الوكلاء في مواجهة هجمات حقيقية يطوّرها باحثون متخصصون، في وقت تسابق فيه منصات متصفحات الذكاء الاصطناعي الزمن لتعزيز أمن المستخدمين، إذ أطلقت منصة "Perplexity" مؤخرًا نموذجاً مفتوح المصدر لرصد المحتوى الضار ومنع الهجمات التي تستهدف الوكلاء الأذكياء.