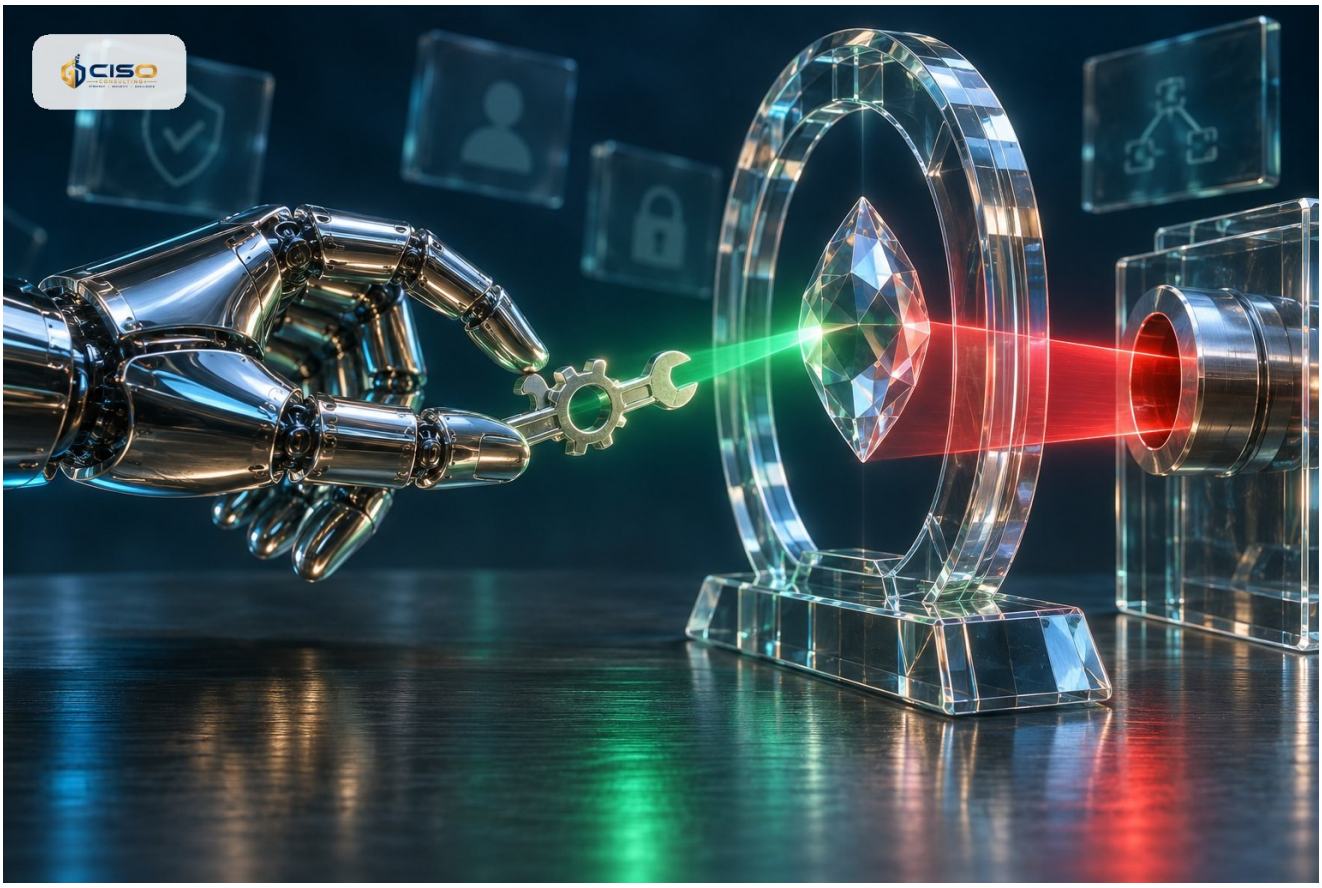
غوغل تعرض بصيغة تجريبية ضوابط سياسات تمنع تنفيذ استدعاء أداة من وكيل ذكاء اصطناعي قبل تشغيله



تستخدم سياسات الحوكمة الدلالية التجريبية من غوغل على منصة Gemini نموذجاً لغوياً للحكم على استدعاءات أدوات الوكلاء، وتمنع ما يخالف قواعد المؤسسة أو طلب المستخدم قبل تنفيذه.

- يعمل الفحص خارج الوكيل عند نقطة إنفاذ مثل Agent Gateway، ويقيّم كل استدعاء أداة مقترح.
- يجب أن يستوفي الاستدعاء طلب المستخدم والقواعد المكتوبة بلغة واضحة من الفريقين القانوني والامتثال؛ وفحص نية المستخدم مفعّل افتراضياً.
- يُستبدل الاستدعاء المرفوض برسالة رفض ولا يُنفَّذ، وتُسجَّل الأحكام ومبرراتها في Cloud Logging.
- الأحكام احتمالية وقد تختلف، لذا توصي غوغل بمراجعتها في وضع التجربة قبل الإنفاذ.

- لا تُقيَّم الاستجابات المتدفقة، وهذه السياسات مكملة للضوابط القائمة وليست بديلًا عنها.

أتاحت غوغل بصيغة تجريبية سياسات الحوكمة الدلالية على منصة Gemini. وهي تضيف نقطة فحص أمام إجراءات وكيل الذكاء الاصطناعي: إذ يُراجَع كل استدعاء أداة يقترحه نموذج الوكيل، ويُمنع قبل تنفيذه إن لم يتوافق مع طلب المستخدم أو خالف قاعدة من قواعد المؤسسة.

آلية العمل

يقع الفحص خارج الوكيل عند نقطة إنفاذ مثل Agent Gateway. فالصلاحيات تحدد ما يستطيع الوكيل فعله، لكنها لا تحدد ما إذا كان استدعاء بعينه يلتزم بالسياسة. وقد يُقنَع النموذج بالتخلي عن تعليماته من مستخدم لديه مبرر مقنع، أو يضلله مستند يحمل تعليمات مخفية، وهي تقنية تُعرف بحقن التعليمات غير المباشر.

ويؤدي نموذج لغوي دور الحَكَم، فيقارن كل استدعاء بطلب المستخدم وبقواعد يكتبها الفريقان القانوني والامتثال بلغة واضحة. وفحص نية المستخدم مفعّل افتراضياً ولا يحتاج إلى سياسات. وإذا فشل الاستدعاء استُبدلت استجابته برسالة رفض ولم يُنفَّذ، وتُرسل الأحكام ومبرراتها إلى Cloud Logging.

ومثال غوغل وكيل مالي يقتصر إرسال النتائج الفصلية غير المعلنة على عناوين داخلية قبل الإعلان. فيُرفض البريد الموجّه إلى مستثمر خارجي، كما يرفض فحص النية تعليمة مخفية في مستند تخطيط تطلب حذف المسودات.

حدود يجب الانتباه لها

لأن الحَكَم نموذج لغوي، فأحكامه احتمالية وقد تختلف. وتنصح غوغل بمراجعتها في وضع التجربة قبل الإنفاذ، وبإدراج الوقائع التي تعتمد عليها السياسة، مثل موعد إعلان الأرباح، في تعليمات النظام الخاصة بالوكيل. ولا تُقيَّم الاستجابات المتدفقة. وتؤكد غوغل أن هذه السياسات تكمّل ضوابط الإفصاح والمراجعات القانونية ومراجعات الامتثال ولا تحل محلها.

ما يجب فعله الآن

1. احصر وكلاء الذكاء الاصطناعي المستخدمين أو قيد التجربة، ودوّن الأدوات التي يستطيع كل منهم استدعاءها، وبخاصة البريد وحذف الملفات وتصدير البيانات.
2. كلّف فريقي الامتثال والشؤون القانونية بصياغة قواعد بلغة واضحة للإجراءات الحساسة، مثل إرسال بيانات مالية غير معلنة أو بيانات شخصية خارج المنشأة.
3. عند تجربة هذه السياسات، شغّلها أولًا في وضع التجربة وراجع الأحكام قبل الإنفاذ.
4. أدرج الوقائع التي تعتمد عليها السياسة في تعليمات النظام للوكيل، واختبرها بسيناريوهات حقن التعليمات غير المباشر.
5. تعامل معها كطبقة واحدة: أبقِ صلاحيات الوصول وضوابط الإفصاح والمراجعة البشرية، واربط سجلات الأحكام بمنظومة المراقبة لديك.