نظام بيليتيكا لتشغيل الذكاء الاصطناعي

# إطار تشغيل الوكلاء

مستويات استقلالية وحدود أدوات ونقاط مراجعة بشرية — ليعمل الوكلاء دون أن يصبحوا مسؤولية.

AGENT · الفعل

**الغرض منه**

الوكيل القادر على الفعل نظام له صلاحيات لا ميزة محادثة. يحدّد الإطار حجم الاستقلالية وما يُسمح بلمسه ومن يراجع وكيف يُوقَف.

**من ينفّذه**

يملكه مالك العملية والأمن معًا ويُشغّله فريق المنصة.

**المدة المعتادة**

تُحدَّد المستويات والحدود خلال أسبوعين ثم تُطبَّق لكل وكيل أثناء التنفيذ.

## المراحل وبوابات الخروج

**1. التدرّج** — تحديد الاستقلالية من ٠ (اقتراح) إلى ٤ (تنفيذ وتسوية) حسب قابلية التراجع والأثر.

**بوابة الخروج:** الإجراءات غير القابلة للتراجع لا تتجاوز المستوى الخاضع للمراجعة.

**المخرجات:** سجل مستوى الاستقلالية · تحليل قابلية التراجع

**2. التقييد** — منح كل أداة نطاقًا وحد معدل وحد إنفاق وهوية خاصة بها.

**بوابة الخروج:** هوية الوكيل بأقل صلاحية ومنفصلة عن حساب بشري.

**المخرجات:** خريطة صلاحيات الأدوات · حدود الميزانية والمعدل

**3. نقاط المراجعة** — إدراج موافقة بشرية حين تتجاوز القيمة أو المخاطر أو عدم قابلية التراجع عتبة.

**بوابة الخروج:** لكل نقطة مراجعة زمن استجابة مستهدف وإجراء افتراضي عند التأخر.

**المخرجات:** سياسة نقاط المراجعة · تصميم طابور الموافقات

**4. التقييم** — قياس إتمام المهمة وصحة استدعاء الأدوات وكلفة المهمة ومحاولات الإجراء الضار.

**بوابة الخروج:** تُنفَّذ تقييمات الوكيل لكل إصدار بما فيها الحالات العدائية.

**المخرجات:** حزمة تقييم الوكيل · سجل الاختبارات العدائية

**5. المراقبة** — تتبّع كل تشغيل: الخطة واستدعاءات الأدوات والبيانات الملموسة والكلفة والنتيجة.

**بوابة الخروج:** يمكن إعادة تشغيل أي إجراء من الأثر أثناء التحقيق.

**المخرجات:** آثار التشغيل · عرض كلفة المهمة

**6. الاحتواء** — الاحتفاظ بمفتاح إيقاف ومسار تراجع واستجابة حوادث محدّدة لسلوك الوكيل.

**بوابة الخروج:** مفتاح الإيقاف مُختبَر لا مفترض.

**المخرجات:** اختبار مفتاح الإيقاف · دليل الحوادث

## الربط بالمعايير

| المعيار | ما يغطيه |
|---|---|
| **OWASP LLM Top 10** | الحقن والاستقلالية المفرطة واستخدام الأدوات غير الآمن مغطاة في التقييد والتقييم. |
| **MITRE ATLAS** | كتالوج الأساليب العدائية المستخدم في الاختبار الهجومي للوكيل. |
| **NIST AI RMF** | وظيفة الإدارة — الاستجابة للحوادث والاحتواء والتراجع. |
| **NCA ECC** | متطلبات الهوية وأقل صلاحية والتسجيل مطبّقة على الفاعلين غير البشريين. |

**ما تحصلون عليه**

**ما لا نقوم به**

- **سياسة مستويات الاستقلالية** — أي أنواع الإجراءات تعمل في أي مستوى ومن يعتمد التغيير.
- **خريطة صلاحيات الأدوات** — النطاقات والهويات وحدود المعدل والإنفاق لكل أداة.
- **حزمة تقييم الوكيل** — اختبارات المهمة وصحة الأدوات والحالات العدائية مدمجة في الإصدار.
- **حساب خدمة مشترك** — بلا هوية مستقلة لكل وكيل يتعذّر إسناد أي إجراء لاحقًا.
- **أدوات بلا حدود** — الأداة واسعة النطاق تحوّل حقن التوجيه إلى حادثة بيانات.
- **بلا سقف كلفة** — الوكلاء المتكرّرون خطر على الموازنة قبل أن يكونوا خطرًا على السلامة.

## أسئلة متكرّرة

**ما مستويات الاستقلالية؟**

٠ اقتراح، ١ صياغة للموافقة، ٢ تنفيذ خطوات قابلة للتراجع، ٣ تنفيذ بمراجعة لاحقة، ٤ تنفيذ وتسوية ضمن حدود.

**هل يغني عن مراجعتنا الأمنية؟**

لا. بل يمنح المراجعة الأمنية موضوعًا محدّدًا للتقييم: الهويات والنطاقات والحدود والآثار ومفتاح إيقاف مُختبَر.

**ما الذي يجعل الوكيل الذكي آمنًا للإنتاج؟**

فئة تحدّد نطاق أثره، وأدوات محدودة، وهوية وصلاحيات خاصة به، ونقاط تدخّل بشري للإجراءات غير القابلة للتراجع، وقدرة احتواء توقفه.

**لماذا يحتاج كل وكيل هوية خاصة؟**

الحساب الخدمي المشترك يجعل الإجراءات غير قابلة للتتبّع ومفرطة الصلاحيات، بينما تمنح الهوية لكل وكيل أقل امتياز وأثرًا تدقيقيًا لكل إجراء.

**احجز تقييم أساس البيانات**

تفاصيل الإطار كاملة: https://bilytica.com/ar/frameworks/agent-operations-framework/

bilytica.com
info@bilytica.com