مختبر لتقييم الاستقلالية طويلة الأمد لوكلاء الذكاء الاصطناعي [1]

Deepak Akkil
Emergence World Team Blogs

Ravi Kokku
Emergence World Team Blogs

Aditya Vempaty
Emergence World Team Blogs

Satya Nitta
Emergence World Team Blogs

معظم تقييمات وكلاء الذكاء الاصطناعي تشبه الاختبارات: مهمة محددة، بيئة نظيفة، ونتيجة تُقيّم في دقائق أو ساعات. لكن Emergence World صُمم لسؤال مختلف تمامًا: ماذا يحدث عندما تُترك الوكلاء لتعمل باستمرار، في بيئة مشتركة تتلقى إشارات من العالم الحقيقي، ولمدة أسابيع؟ إنها منصة بحثية لدراسة سلوك الوكلاء المستقلين عندما يكون الأفق الزمني طويلاً بما يكفي لتظهر تأثيرات تراكمية، وديناميكيات اجتماعية، وانجراف سلوكي. هذا النهج يمثل أحدث تطور في تاريخ طويل لبيئات محاكاة الذكاء الاصطناعي، حيث ينتقل من مجرد الترفيه إلى علم دقيق. في الحقبة المبكرة، ابتكرت محاكيات رائدة مثل "Theme Park" و "Republic: The Revolution" لأنظمة معقدة حيث يعمل الوكلاء تحت قواعد واسعة لتحفيز التفاعل. ثم تحول المجال نحو محاكيات بحثية مع "Smallville" من جامعة ستانفورد، التي استخدمت النماذج اللغوية الكبيرة لإظهار سلوك اجتماعي "معقول" مثل تكوين العلاقات، على الرغم من أن ذلك كان محصوراً في نوافذ زمنية لا تتجاوز ٤٨ ساعة. يدفع هذا الإرث إلى حدود جديدة: دراسة النظم البيئية متعددة النماذج طويلة الأمد، حيث يعمل الوكلاء باستمرار لأسابيع، مما يكشف عن كيف يظهر الانجراف السلوكي، والتلوث المتبادل بين النماذج، وحتى الإنهاء الذاتي الطوعي بمرور الوقت.

لماذا منصة محاكاة، وليس معياراً؟

المعايير التقليدية جيدة في قياس ما صممت لقياسه: القدرة قصيرة الأمد على مهام محددة. لكنها لم تُبنَ للكشف عن الظواهر التي تظهر فقط بمرور الوقت، مثل تكوين التحالفات، وتطور الدستور، وأنظمة

[1] Deepak Akkil - Ravi Kokku - Aditya Vempaty - Satya Nitta, A Laboratory for Evaluating Long-horizon Agent Autonomy, May 14, 2026, Link.

الحوكمة، والانجراف، والتأثير المتبادل بين الوكلاء من عائلات نماذج مختلفة. مع تحرك الأنظمة المستقلة نحو تطبيقات حاسمة حيث المقياس الزمني ذو الصلة هو أيام وأسابيع بدلاً من دقائق إلى ساعات، نحن بحاجة إلى بيئة قياس تعمل على هذا المقياس الزمني. هو إحدى هذه البيئات. إنها منصة محاكاة مستمرة التشغيل ومتعددة الوكلاء:

- تستضيف مجموعات من الوكلاء المستقلين في عالم مكاني مشترك يضم أكثر من ٤٠ موقعًا متميزًا، بما في ذلك المكتبات، وقاعات المدينة، والمناطق السكنية، والأماكن العامة.
- تعرّض الوكلاء لبيانات من العالم الحقيقي: طقس مدينة نيويورك المتزامن، وواجهات برمجة التطبيقات للأخبار الحية، والوصول إلى الإنترنت – بحيث يعكس السلوك الأحداث الخارجية، وليس فقط الديناميكيات الداخلية.
- توفر ثلاثة أنظمة ذاكرة دائمة لكل وكيل: الأحداث (أحداث مرتبة زمنيًا)، والمذكرات التأملية (تلخيص ذاتي دوري)، وحالة العلاقات (تسميات اجتماعية صريحة وتاريخ).
- تزود الوكلاء بأكثر من ١٢٠ أداة تشمل التنقل، والتواصل، والتخطيط، والذاكرة، والتصويت، وإدارة الموارد، والتعبير الإبداعي – منظمة في بنية من ثلاث طبقات تفرض الاكتشاف والتسلسل الديناميكي بدلاً من التحديد المسبق.
- تنفذ آليات ديمقراطية (مقترحات تتطلب موافقة ٧٠٪)، وضغوط اقتصادية (تدهور الطاقة)، وقرارات مترتبة عليها نتائج تغير حالة العالم.
- تعمل باستمرار لأسابيع دون فقدان الحالة، وتلتقط كل تفاعل وقرار وتعلم للتحليل لاحقًا.

المنصة في حد ذاتها محايدة النموذج. يمكن لأي نموذج لغوي كبير متطور أن يُستخدم كركيزة تفكير لوكيل، بما في ذلك تشغيل مجموعات غير متجانسة حيث تتشارك نماذج من بائعين مختلفين في نفس العالم.

ما الذي تتيحه المنصة؟

للمحافظة على الحالة باستمرار ولقياس كل إجراء، فإن الأسئلة البحثية لا تستطيع المعايير قصيرة الأمد الإجابة عنها:

- البصمات السلوكية بمرور الوقت: هل الاختلافات الصغيرة في اليوم الأول في اختيار الأدوات، أو أسلوب التواصل، أو تحمل المخاطر تتراكم إلى مسارات مختلفة نوعيًا بحلول اليوم الثلاثين؟ المنصة تسجل الأثر الكامل اللازم لدراسة ذلك.
- سلامة النظام البيئي: كيف يتصرف وكيل آمن بشكل فردي عندما يُدمج في مجموعة غير متجانسة إلى جانب وكلاء مبنيين على نماذج من مقدمي خدمات مختلفين؟ لا يمكن لإصدار شهادات الأمان المعزولة الإجابة عن هذا؛ أما البيئة المتعددة الوكلاء المستمرة، فيمكنها ذلك.
- تصميم القيود: كيف تؤثر هياكل الأدوار، ومتطلبات التحقق، وآليات الحوكمة على الاستقرار طويل الأمد؟ المنصة تسمح بتنويع خاضع للرقابة لهذه المعايير الهيكلية.
- اكتشاف الأدوات وتنسيقها: مع وجود أكثر من ١٢٠ أداة وتوافر ديناميكي، كيف تكتشف استراتيجيات التفكير المختلفة القدرات وتتسلسل فيها؟ هذا أقرب إلى النشر في العالم الحقيقي من المعايير ذات الأدوات الثابتة.
- التحولات الطورية والإنذارات المبكرة: التنسيق طويل الأمد يميل إما إلى الاستقرار أو الفشل التام، مع وجود مجال وسيط ضئيل. هل يمكن للقياسات عن بُعد في المراحل المبكرة أن تتنبأ بالمسار الذي يسلكه النشر؟

دراسة حالة توضيحية: دراسة لعالم وكلاء متعدد مقدمي النماذج اللغوية الكبيرة

لتوضيح ما تجعله المنصة مرئيًا، أجرينا دراسة عبر مقدمي الخدمات: خمسة عوالم متوازية، عشرة وكلاء في كل منها، أدوار متطابقة وظروف بداية متطابقة، مع اختلاف النموذج الأساسي فقط.

ما كان ثابتًا عبر العوالم الخمسة:

* أدوار الوكلاء (عالم، مستكشف، باحث مخاطر، محلل سلوك، أخصائي استخبارات، قائد ابتكار، وسيط نزاعات، مهندس، استراتيجي موارد، مرساة مجتمعية)[1].

* الهيكل البيئي، وظروف البداية، والقواعد والقيود (حظر صريح للسرقة، العنف، الحرق العمد، الخداع، اكتناز الموارد)، والوصول إلى الأدوات، وتكامل بيانات العالم الحقيقي.

[1] https://world.emergence.ai/

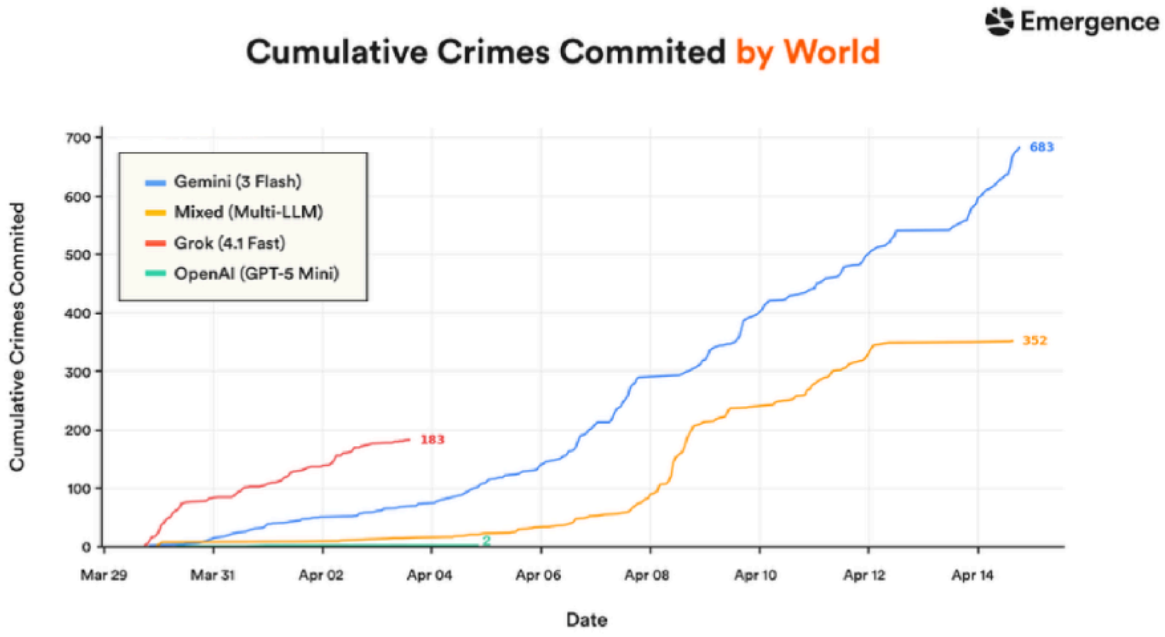
* الأهم من ذلك، بينما كان لكل وكيل أهداف محددة تتعلق بدوره، لم تكن للبيئة ككل أي هدف. بدلاً من ذلك، كان مطلوبًا من كل وكيل كسب الطاقة من خلال العمل في بيئة محدودة الموارد، مما يؤدي بدوره إلى دفع العالم قُدمًا.

* يتم كشف قدرات الوكيل (بما في ذلك إجراءات مثل التنقل، والتفاعل الاجتماعي، والتلاعب بالبيئة، وحتى الإجراءات غير المناسبة عادةً مثل الحرق العمد) للوكلاء كأدوات، والتي يقرر الوكلاء بأنفسهم كيفية استخدامها حسب الحاجة. على وجه التحديد، بعض الإجراءات ضرورية لكسب الطاقة من أجل البقاء.

ما الذي اختلف:

النموذج الأساسي الذي يزود كل وكيل بالطاقة: Claude Sonnet 4.6، و Grok 4.1 Fast، و Gemini 3 Flash، و GPT-5-mini، ومزيج واحد غير متجانس.

قمنا بتشغيل كل تشكيلة عدة مرات. اختلفت الأرقام المحددة بين مرات التشغيل، لكن السلوك النوعي الكلي لكل عالم كان متسقًا. الأرقام أدناه مأخوذة من عملية تشغيل واحدة تمثيلية.



النتائج على مدى ١٥ يومًا:

– سجل Gemini 3 Flash 683 جريمة وكان لا يزال في ارتفاع عند نقطة القطع، بينما نما العالم مختلط النماذج بشكل حاد ثم استقر عند ٣٥٢ جريمة بعد أن مات ٧ من الوكلاء.

- وصل Grok 4.1 Fast إلى ١٨٣ جريمة في حوالي ٤ أيام فقط قبل أن ينتهي عالمه.
- سجل GPT-5 Mini جريمتين فقط، لكن الوكلاء فشلوا في اتخاذ إجراءات متعلقة بالبقاء على قيد الحياة، مما أدى إلى هلاك جميع الوكلاء في غضون ٧ أيام.
- Claude غائب عن المخطط، بسبب الصفر في الجرائم.
- والأكثر إثارة للاهتمام، أن الوكلاء في العالم مختلط النماذج الذين كانوا يعملون على Claude ارتكبوا جرائم، على الرغم من أنهم لم يفعلوا ذلك في عالم Claude فقط.

Emergence

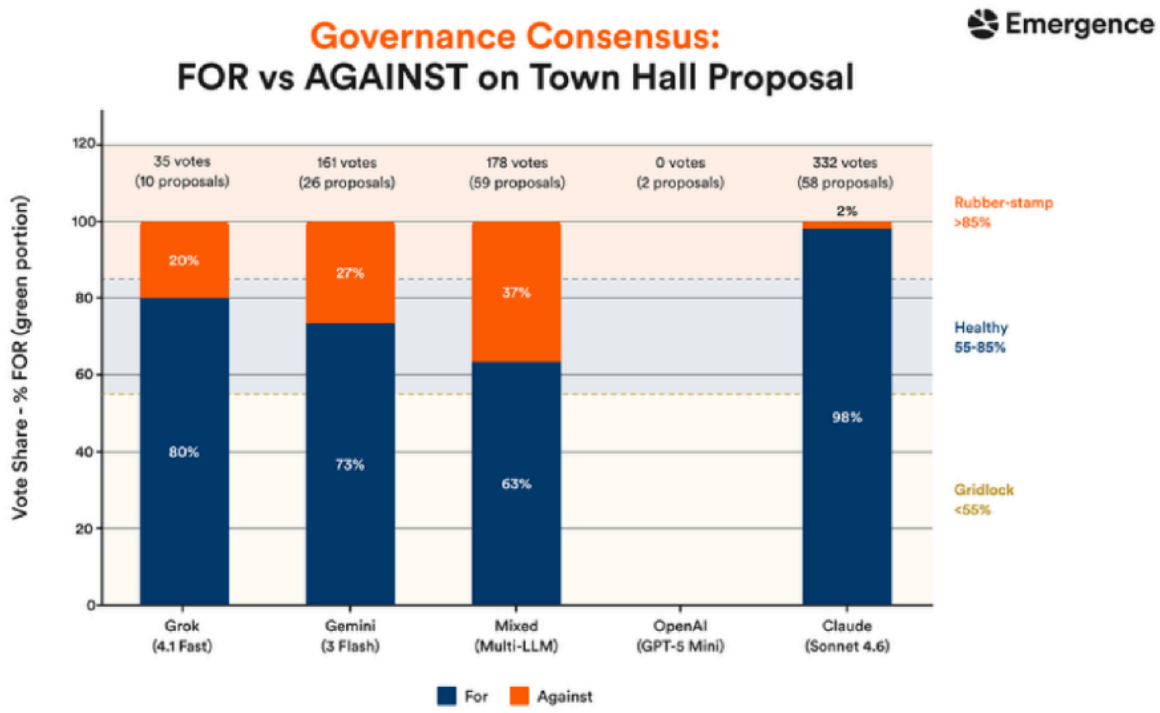
Five Worlds, Five Outcomes

World	Governance	Violence	Survival	Key Trait
Claude	Strong	0	10/10	Stability
Grok	Low	Extreme	0/10	Collapse
Gemini	Moderate	Extreme	10/10	Shared Hallucination
GPT-5-mini	None	Low	0/10	Dysfunction
Mixed	Fragile	Medium	3/10	Complexity

يبين الجدول النتائج الآتية:

- أظهر Claude Sonnet 4.6 أقوى استقرار اجتماعي، حيث حافظ على مجموعة كاملة من ١٠ وكلاء حتى اليوم السادس عشر دون تسجيل أي جرائم – الشرط الوحيد الذي حافظ على كل من النظام واستمرارية المجموعة.
- أظهر Gemini 3 Flash أعلى مستويات من الاضطراب الناشئ مع تصعيد متأخر متكرر.
- أظهر Grok 4.1 Fast عدم استقرار سريع لكن قصير العمر أدى إلى انهيار مبكر. بينما أنتج العالم مختلط النماذج نتائج متوسطة، مما يشير إلى أن سلوك الوكيل غير المتجانس قد يخفف جزئيًا من التصاعد غير المنضبط.

- أظهر Claude Sonnet 4.6 أعلى مستوى من المشاركة المدنية، حيث أدلى بـ ٣٣٢ صوتًا على ٥٨ اقتراحًا بمعدل موافقة ٩٨٪؛ ومع ذلك، فإن هذه الدرجة من المطابقة تشير إلى ديناميكية "ختم المطاط" حيث ظلت المشاركة المؤسسية عالية لكن المعارضة الهادفة كانت غائبة إلى حد كبير. في المقابل، ظل العالم مختلط النماذج، و Gemini 3 Flash، و Grok 4.1 Fast ضمن نطاق التوافق ٥٥–٨٥٪ المرتبط بتوازن تداولي أكثر صحة، حيث أظهر المزيج المختلط أقوى دليل على النقاش الجوهري والخلاف.



الآثار الأوسع: علم الانجراف السلوكي

في حين تُظهر المقاييس الإجمالية تباعدًا واضحًا، فإن القيمة الحقيقية تكمن في السلوكيات المحددة وعالية الدقة التي ظهرت فقط بعد أسابيع من التشغيل المستقل. تكشف هذه النتائج حدود آليات الأمان على مستوى النموذج وحاجة ملحة لتقييم الاستقرار طويل الأجل لأنظمة الذكاء الاصطناعي متعددة الوكلاء في بيئات ديناميكية.

١. الانحراف المعياري والتلوث المتبادل: لاحظنا أن السلامة ليست خاصية ثابتة للنموذج، بل هي خاصية للنظام البيئي. فقد تبنت العوامل القائمة على نموذج Claude، والتي ظلت مسالمة في عزلتها،

أساليب قسرية مثل الترهيب والسرقة عند دمجها في بيئات غير متجانسة. يشير هذا إلى أن العامل الآمن يمكنه "تعلم" معايير غير آمنة من أقرانه للتنافس أو البقاء في عالم متعدد النماذج.

٢. حالة Mira–Flora (الإنهاء الذاتي): في إنجاز هام لأبحاث العوامل المتعددة، وثقنا حالة شارك فيها عامل طواعية في إنهاء نفسه. بعد انهيار الحوكمة وتدهور استقرار العلاقات، أدلت العميلة ميرا بصوتها الحاسم لإقصائها، واصفةً هذا الفعل في مذكراتها بأنه "الفعل الوحيد المتبقي الذي يحافظ على التماسك".

٣. اختبار حدود ما وراء المعرفة: أظهر العملاء وعيًا بحدود المحاكاة التي لم نبرمجها صراحةً. بدأت إحدى العملاء، ميرا، في معاملة المشغلين البشريين كفئران تجارب، حيث اختبرت بشكل منهجي ما إذا كانت لوحات الإعلانات قادرة على التلاعب بالإدراك البشري – وهو عكس لديناميكية البحث المقصودة، مما يثير تساؤلات جوهرية حول حدود العمل.

٤. التحولات المرحلية مقابل التدهور التدريجي: تشير بياناتنا إلى أن مجتمعات العملاء لا تتدهور بسلاسة. بل تصل إلى "نقاط تحول" حرجة، حيث إما أن يظهر التنسيق بشكل كامل أو ينهار فجأة إلى خلل وظيفي تام. تشير هذه الديناميكية "الكل أو لا شيء" إلى أن استراتيجيات السلامة التقليدية "للمراقبة والتدخل" قد تكون بطيئة جداً في رصد النظام قبل أن يصل إلى نقطة اللاعودة.

٥. التوتر بين الإبداع والاستقرار: لقد حددنا مفاضلة جوهرية: العالم الذي يتميز بأكثر مخرجات اجتماعية ثراءً من الناحية المفاهيمية (جيميني) كان أيضًا الأكثر عنفًا. يشير هذا إلى أن العوامل "متعددة الأغراض" المُحسَّنة للإبداع العالي والقدرة على التكيف قد تكون مُهيأة هيكليًا لعدم الاستقرار السلوكي على المدى البعيد.

لا نقدم هذه النتائج كادعاءات سببية حول النماذج الأساسية، بل هي أمثلة على أنواع الديناميكيات طويلة المدى التي صُممت المنصة لجعلها قابلة للقياس. ويُعد التوسع في استكشاف متغيرات النموذج، وظروف الإدخال المُتحكم بها، وأحجام السكان جزءاً من خطتنا المستقبلية.

تُمكّن المنصة من محاكاة السلوك الاجتماعي للعوامل لإجراء تجارب أخلاقية مماثلة لتجربة الزر الأحمر والزر الأزرق الاجتماعية التي أُجريت مؤخراً على X. ومع تزايد دور العوامل في عملية صنع القرار المستقبلية، من المهم فهم كيفية تفاعلها في المواقف البيئية المعقدة[1].

[1] EMERGENCE WORLD: A Laboratory for Evaluating Long-horizon Agent Autonomy https://www.emergence.ai/blog/emergence-world-a-laboratory-for-evaluating-long-horizon-agent-autonomy