الهجومية المعرفية للذكاء الوثوقي الفتّاح الضغطي ذي أثر العالم

نظرية تشغيلية للوكيل الذي يهاجم جهلَه قبل أن يسمح لنفسه بالفعل

Cognitive Offensive for Reliable, Open-Ended, Compressive, World-Effect Intelligence (COROC-WEI)

الادعاء المركزي

الذكاء الوكيلي الأقوى ليس الذي "يفكّر أكثر" ولا الذي "يفعل أكثر"، بل الذي يخصّص جزءًا من موارده بصورة عدائية ومنهجية لكسر نماذجه الداخلية، واكتشاف ما لا يعرفه، واختبار ما يظنه صحيحًا، وضغط معرفته إلى تمثيلات سببية قابلة للتشغيل، ثم إحداث أثر خارجي مضبوط وقياس الفرق بين الأثر المتوقع والأثر المتحقق. في هذا الإطار تصبح الوثوقية قدرة هجومية على منع الخطأ قبل تحوله إلى أثر في العالم، لا مجرد تصحيح بعد الفشل.

إطار بحثي نظري/هندسي مقترح - نسخة تأسيسية، سبتمبر 2026

الحالة المعرفية للورقة: فصل صريح بين الأدلة المنشورة، الاستدلال النظري، الفرضيات القابلة للتكذيب، والمقترحات الهندسية.

الملخص

تقترح هذه الورقة مفهوم «الهجومية المعرفية» بوصفه مبدأً لتصميم الذكاء الوكيلي العام: الوكيل لا ينتظر أن يقع الخطأ لكي يتأمل فيه، بل ينشئ بصورة استباقية خصومًا معرفيين لفرضياته وخططه وذاكرته وثقته، ويبحث عن الملاحظة أو التجربة أو الأداة التي تعظّم احتمال اكتشاف كونه مخطئًا قبل التنفيذ عالي الأثر. ويُدمج هذا المبدأ داخل بنية تجمع أربعة محاور عادة ما تُبحث منفصلة: الوثوقية Reliability، والانفتاح أو الفتوح Open-endedness، والضغط المعرفي Compressive Intelligence، وأثر العالم World Effect. الحجة الأساسية هي أن جمع هذه المحاور ينتج نمطًا مختلفًا من الذكاء: نظامًا لا يقيس المعرفة بما يستطيع قوله فقط، بل بما يستطيع ضغطه سببيًا، مهاجمته نقديًا، تحويله إلى خطة قابلة للتحقق، ثم ربطه بفارق حالة observable في بيئة خارجية.

تقدّم الورقة دالة هدف مقيدة بالمخاطر، وسجلًا للادعاءات والوثوق، ومعمارية متعددة الطبقات، وخوارزمية تشغيل، ومقاييس جديدة مثل Offensive Epistemic Yield وWorld-Effect Fidelity وCompression-to-Control Ratio وTrustworthy Frontier Expansion، إضافةً إلى برنامج تجريبي يحوّل الفكرة إلى فرضيات قابلة للتكذيب. وهي لا تدّعي أن هذا الإطار مثبت تجريبيًا، بل تفصل بين ما تدعمه الأدبيات الحالية - مثل هشاشة الوكلاء في المهام طويلة الأفق، فائدة دمج reasoning/action، أهمية world models، وفشل pass@1 كمقياس وحيد للوثوقية - وبين الإسهام النظري المقترح: جعل «البحث النشط عن قابلية التكذيب» مكوّنًا أوليًا في policy الوكيل، وربطه بقيود رسمية وبقياس أثر العالم.

الكلمات المفتاحية

Agentic AI؛ Trustworthy AI؛ Open-ended Intelligence؛ Cognitive Offensive؛ Epistemic Action؛ Active Inference؛ World Models؛ Causal Compression؛ Verification؛ Long-Horizon Reliability؛ World Effect؛ Falsification؛ AGI Architecture.

خريطة الحالة المعرفية

الرمز	المعنى
Evidence [E]	ادعاء مدعوم بأدبيات/benchmark منشور.
Inference [I]	استنتاج نظري مبني على الأدلة، لكنه ليس نتيجة تجريبية مباشرة.
Hypothesis [H]	فرضية جديدة قابلة للتكذيب.
Design [D]	مقترح معماري أو خوارزمي يحتاج تنفيذًا واختبارًا.

1. المشكلة: لماذا الذكاء القادر ليس بالضرورة ذكاءً يُوثق به؟

[E] تُظهر بيئات تقييم الوكلاء الواقعية أن الانتقال من الإجابة النصية إلى الفعل المتسلسل يكشف طبقة مختلفة من الصعوبة. WebArena صُممت لمهام ويب طويلة الأفق وقابلة للتنفيذ، وOSWorld أضاف بيئة حاسوبية حقيقية بتقييم قائم على حالة النظام، بينما τ-bench أبرز أن الوكيل قد ينجح في بعض المحاولات ثم يفشل في التكرار نفسه، لذلك اقترح pass^k بدل الاكتفاء بنجاح محاولة واحدة [1-3]. وفي 2026 ظهرت أعمال تصوغ reliability decay وvariance amplification وmeltdown onset للمهام طويلة الأفق، بما يرسخ أن القدرة والوثوقية متغيران منفصلان [4].

[I] هذه النتائج تقود إلى مشكلة أعمق: معظم هندسات الوكلاء تظل «مهمة-مركزية» Task-centric. أي أنها تبدأ من هدف خارجي وتسأل: كيف أحققه؟ أما الإطار المقترح هنا فيضيف سؤالًا أسبق: ما الذي يجب أن أهاجمه داخل تمثيلي للهدف والعالم قبل أن أثق بأن السعي إلى الهدف نفسه مفهوم وآمن وفعّال؟ هذا التحول من task solving إلى epistemic adversarialism هو قلب «الهجومية المعرفية».

الهجومية بهذا المعنى ليست عدوانًا اجتماعيًا ولا هجومًا سيبرانيًا ولا تعظيمًا للأثر بلا قيد. هي اتجاه للموارد المعرفية نحو أكثر نقاط الاعتقاد هشاشة: distribution shift، hidden state، tool uncertainty، causal confounds، evidence gaps، assumptions، irreversible actions، value ambiguity. الوكيل «يهاجم» مواطن العمى في نفسه وفي نموذج المهمة قبل أن يوسع سلطته على البيئة.

2. الأطروحة: من وكيل يجيب إلى وكيل يهاجم جهله

تعريف 1 - الهجومية المعرفية Cognitive Offensive

سياسة فوقية Meta-policy تختار عمدًا عمليات معرفية أو تجريبية أو أداتية تزيد احتمال اكتشاف خطأ مهم في الاعتقاد أو الخطة أو الثقة قبل أن ينتج هذا الخطأ أثرًا خارجيًا عالي التكلفة. وهي تختلف عن «التفكير الطويل» لأن معيارها ليس عدد الخطوات، بل مقدار الجهل الحرج الذي تكشفه أو تقلصه لكل وحدة تكلفة ومخاطر.

[D] نقترح أن يمتلك الوكيل في كل دورة «سطح هجوم معرفي» Epistemic Attack Surface يتكون من الادعاءات عالية الأثر، والافتراضات غير المختبرة، والتبعيات الخارجية، ونقاط القرار غير القابلة للعكس، ومناطق الثقة العالية ذات الدليل الضعيف. ثم يوزع ميزانيته على أربعة أصناف من الهجوم: هجوم تكذيبي على الفرضيات، هجوم سببي على التفسيرات البديلة، هجوم وثوقي على الثقة والمعايرة، وهجوم أثري على التوقعات المتعلقة بما سيحدث فعليًا بعد التنفيذ.

[H] الفرضية الجوهرية هي أن الوكيل الذي يخصّص نسبة تكيفية من compute لاكتشاف قابلية الفشل قبل التنفيذ سيحقق، عند أفق طويل، معدل نجاح متكرر أعلى وانهيارات أقل من وكيل يخصص الميزانية نفسها لزيادة البحث داخل الخطة فقط. وتصبح هذه الفرضية أقوى عندما تكون البيئة جزئية الرصد، والأفعال غير قابلة للعكس جزئيًا، والتكلفة الخارجية للخطأ أكبر من تكلفة الاستعلام أو المحاكاة.

3. لماذا "وثوقي + فتّاح + ضغطي + ذو أثر العالم" تركيبة واحدة؟

المحاور الأربعة ليست صفات تجميلية. كل واحد منها يصحح انحرافًا قد يصيب الثلاثة الأخرى. الوثوقية بلا انفتاح تتحول إلى محافظة معرفية تمنع التعلم، والانفتاح بلا وثوقية يتحول إلى انفجار فرضيات وأفعال، والضغط بلا أثر قد ينتج تمثيلات أنيقة لكنها غير نافعة للسيطرة، والأثر بلا ضغط أو تحقق قد يضاعف أخطاء الوكيل في العالم. لذلك يُقترح النظر إليها كمنظومة توتر ديناميكي: Expand → Attack → Compress → Verify → Act → Measure → Re-expand.

• الوثوقية Reliability: لا تسمح بتحويل احتمال داخلي إلى فعل خارجي إلا عبر بوابات evidence, calibration, verification, rollback.

• الفتوح Open-endedness: يوسع فضاء الأهداف والفرضيات والأدوات والتمثيلات عندما يكون النموذج الحالي غير كافٍ، بدل الالتصاق بحل أولي.

• الضغط Compressive Intelligence: يحول الخبرة المتراكمة إلى أقل تمثيل يحافظ على القدرة السببية/التنبؤية/التحكمية اللازمة، بدل تضخم الذاكرة الخام.

• أثر العالم World Effect: يجعل صحة التفكير مرتبطة بفارق حالة قابل للرصد في البيئة، لا بجمال السرد أو الاتساق النصي.

4. الجذور العلمية وحدودها

4.1 Reasoning-Acting والتعلم من التغذية الراجعة

[E] قدّم ReAct صيغة تداخل reasoning/action بما يسمح للفعل بجلب معلومات جديدة وللاستدلال بتحديث الخطة [5]. وقدّم Reflexion آلية تستخدم feedback لغويًا وذاكرة انعكاسية لتحسين المحاولات التالية [6]. وتسمح Tree of Thoughts باستكشاف بدائل بدل الالتزام بمسار توليد وحيد [7]. هذه الأعمال مهمة، لكنها لا تجعل «اختيار الاختبار الذي يزيد فرصة تكذيب الذات» هدفًا صريحًا مقيدًا بأثر العالم.

4.2 الانفتاح واكتساب المهارات

[E] Voyager بيّن قيمة curriculum ذاتي ومكتبة مهارات متنامية وfeedback تنفيذي في بيئة مفتوحة [8]. إلا أن open-endedness قد يوسع السلوك أسرع من قدرة التحقق؛ لذلك يقترح إطارنا أن كل توسع في frontier القدرات يجب أن يرافقه توسع موازٍ في frontier الاختبارات والقيود ومقاييس الثقة.

4.3 World Models والفعل المتخيل

[E] DreamerV3 يوضح قوة world models في تخيل trajectories قبل الفعل وقد أثبت أداءً واسعًا عبر مهام تحكم مختلفة [9]. الفكرة المقترحة هنا تستعير مبدأ التخيل لا النتيجة نفسها: الوكيل الوثوقي لا يكتفي بتوقع المسار المرجح، بل يولّد مسارات مضادة Counterfactuals موجهة خصوصًا نحو اكتشاف أين يفشل نموذج العالم.

4.4 Active Inference, Epistemic Value

[E] أدبيات Active Inference تدمج قيمة المعلومات مع المنفعة، وتُظهر كيف يمكن للتخطيط أن يوازن الهدف الخارجي وتقليل عدم اليقين [10-11]. الإضافة هنا هي تخصيص هذا الدافع المعرفي: ليست كل معلومة مساوية. الأولوية للمعلومة التي يمكن أن تسقط ادعاءً عالي الأثر أو تمنع فعلًا غير قابل للعكس.

Runtime Enforcement, Formal Safety 4.5

[E] AgentDojo يوضح مخاطر prompt injection عندما يعمل الوكيل فوق بيانات وأدوات غير موثوقة [12]. وأعمال 2026 حول tool-use الآمن تقترح اشتقاق قيود قابلة للإنفاذ من تحليل المخاطر، بينما تشير مراجعات حديثة إلى أن verification وحده يواجه «ضريبة verifier» ومشكلة specification [13-14]. لذلك لا تُختزل الهجومية المعرفية في verifier خارجي؛ بل تُوزع التحقق بين specification, causal model, evidence ledger, action monitor, and post-effect audit.

5. مبادئ النظرية: اثنا عشر قانونًا تشغيليًا

1. قانون أسبقية قابلية التكذيب: كلما ارتفع أثر القرار وجب أن ترتفع أولوية البحث عن دليل يمكنه إسقاط الخطة، لا مجرد دليل يؤيدها.

2. قانون فصل القدرة عن الوثوق: نجاح محاولة واحدة دليل على capability محلية، لا على reliability. يجب قياس التكرار، التباين، وطول الأفق.

3. قانون تكلفة الجهل الحرِج: الجهل ليس كمية واحدة؛ قيمته السلبية تتناسب مع احتمال تأثيره في القرار × حجم الأثر × صعوبة الرجوع.

4. قانون الضغط السببي: لا قيمة للضغط الذي يحذف المتغيرات الضرورية للتنبؤ أو التحكم. الهدف minimal sufficient causal state لا minimal text.

5. قانون الفتوح المقيد: التوسع في فرضيات أو مهارات جديدة مشروع ما دام يرافقه budget للتحقق وفاصل صلاحيات يمنع الانتقال المباشر إلى أثر مرتفع.

6. قانون الفرق بين التوقع والأثر: كل فعل مهم ينتج توقعًا مسبقًا وتدقيقًا لاحقًا للفرق بين predicted state وobserved state.

7. قانون الارتداد المعرفي: عند mismatch كبير، لا يُرقع الجواب فقط؛ يُعاد فتح الفرضيات التي اعتمدت عليها الخطة upstream.

8. قانون تنويع الخصوم: الناقد الواحد قد يشارك النموذج نفسه في نقاط العمى؛ لذا تتنوع طرق النقد: formal checks, external evidence, alternative models, simulation, human review عند اللزوم.

9. قانون الاقتصاد الهجومي: لا يُهاجم كل شيء بنفس القوة؛ تُوجّه موارد falsification نحو أعلى expected risk reduction per unit cost.

10. قانون الأثر القابل للإسناد: لا يُنسب نجاح العالم إلى الخطة إن تعذر التفريق بين causal effect والمصادفة أو تدخل خارجي.

11. قانون ذاكرة الفشل المركزة: يجب تخزين أنماط الفشل وconditions المسببة لها بضغط أعلى من تخزين السرد الكامل للمحاولة.

12. قانون التوقف الآمن: إذا تجاوز uncertainty أو risk حدودًا معلنة ولم تعد هناك تجربة منخفضة المخاطر لتقليلها، تكون abstention أو escalation فعلًا ذكيًا لا فشلًا.

6. النموذج الرسمي: حالة معرفية لا حالة مهمة فقط

[D] تمثل حالة الوكيل عند الزمن t لا كـ state بيئي s_t فقط، بل كحزمة معرفية K_t = (B_t, C_t, U_t, T_t, M_t, G_t, S_t). حيث B هي beliefs، وC نموذج سببي، وU عدم اليقين، وT سجل الوثوق والمعايرة، وM الذاكرة المضغوطة، وG مجموعة الأهداف/الفرضيات المفتوحة، وS القيود والسياسات. ويولد الوكيل نوعين من الأفعال: أفعال معرفية a^e (استعلام، تجربة، محاكاة، نقد، تحقق) وأفعال عالمية a^w تغيّر البيئة.

$$K_t = (B_t, C_t, U_t, T_t, M_t, G_t, S_t) \quad (1)$$

يُعرّف سطح الهجوم المعرفي A_t كمجموعة عناصر belief/plan التي تجمع أثرًا متوقعًا عاليًا مع هشاشة إثباتية أو سببية. ويمكن تقريب أولوية مهاجمة العنصر i بـ:

$$\text{Priority}_i = \text{Impact}_i * \text{Uncertainty}_i * \text{Irreversibility}_i * \text{DependencyCentrality}_i / (\text{TestCost}_i + \text{epsilon}) \quad (2)$$

هذا التعريف يغيّر معنى «الشك»: الشك ليس قيمة فلسفية عامة بل مورد تخصيص. الادعاء منخفض الأثر لا يستحق محاكمة عميقة، بينما claim يقع في عنق زجاجة لخطة غير قابلة للعكس يجب أن يتعرض لأقسى اختبار متاح.

7. دالة الهدف: قيمة معرفية + قيمة أثر - مخاطر - ثقة زائفة

[D] نقترح دالة قرار متعددة الحدود بدلاً من reward وحيد. على π policy أن تعظم على أفق H:

$$J(pi) = E[\ alpha*IG + beta*CEG + gamma*WEF + delta*OEV + eta*CCR - lambda*Risk - mu*Miscal - nu*Irrev - rho*Compute\ ] \quad (3)$$

حيث IG = Information Gain؛ CEG = Counterexample Gain أي قيمة العثور على حالة تنقض النموذج؛ WEF = World-Effect Fidelity؛ OEV = Open-Ended Value؛ CCR = Compression-to-Control Ratio؛ Miscal = سوء المعايرة؛ Irrev = تكلفة اللارجعة. ليست الأوزان ثابتة بالضرورة؛ بل تتغير حسب حساسية المجال ومرحلة التنفيذ.

الأهم من الجمع الخطي هو وجود قيود صلبة لا يسمح reward بتعويضها:

$$P(catastrophic_effect \mid K_t, a_t) <= epsilon \quad and \quad ConstraintViolation(a_t) = 0 \quad (4)$$

[I] بهذا يصبح النظام هجوميًا في المعرفة ومحافظًا في الصلاحيات: يحق له توسيع فرضياته بلا حد نظري، لكن قدرته على تحويل تلك الفرضيات إلى أفعال خارجية ترتفع فقط بقدر الإثبات والتحقق وقابلية الرجوع.

8. الضغط المعرفي: من ضغط النص إلى ضغط القدرة على التحكم

الضغط المقصود هنا ليس summarization ولا تقليل tokens فقط. [D] نعرّف «الحالة المضغوطة الكافية للتحكم» z_t بأنها تمثيل أصغر من التاريخ الخام h_t ويحافظ - ضمن tolerance محدد - على القدرة على توقع العواقب واختيار الأفعال ذات الصلة بالهدف والقيود. فإذا كان الضغط يرفع الإيجاز ويخفض fidelity السببية فهو ضغط ضار.

$$z_t = f(h_t), \quad minimize\ L(z_t)\ \ subject\ to\ \ DeltaControl(h_t,z_t) <= tau_c\ \ and\ \ DeltaRisk(h_t,z_t) <= tau_r \quad (5)$$

يقترح الإطار ضغطًا ذا ثلاث طبقات: (1) ضغط دلالي يحتفظ بالمفاهيم؛ (2) ضغط سببي يحتفظ بالمتغيرات والعلاقات المؤثرة؛ (3) ضغط هجومي يحتفظ خصوصًا بما كسر فرضيات سابقة وبشروط الفشل. النتيجة ليست ذاكرة أقصر فقط بل ذاكرة أكثر فائدة في منع تكرار الانهيارات.

9. الفتوح: الذكاء الذي لا يغلق فضاء الاحتمال مبكرًا

الفتوح Open-endedness هنا له معنيان. الأول معرفي: القدرة على إنشاء فرضيات وطرق تمثيل وأسئلة لم تُذكر في المهمة. والثاني أداتي: القدرة على تركيب مهارات/أدوات جديدة عندما تكون الأدوات الحالية عنق زجاجة. لكن الفتوح غير المقيد قد يخلق «تضخم agency»: نتوسع القدرة على الفعل أسرع من القدرة على إثبات الصواب. لذلك يُقترح مؤشر Trustworthy Frontier Expansion (TFE): مقدار frontier الجديد الذي اجتاز تحققًا متناسبًا مع أثره.

$$TFE = VerifiedNovelCapability / (TotalNovelCapability + epsilon) \quad (6)$$

[H] التنبؤ: الأنظمة ذات TFE أعلى، حتى لو كان نمو قدراتها الخام أبطأ قليلًا، ستتفوق في المهام طويلة الأفق ذات التوزيع المتغير لأن تراكمها المعرفي أقل تلوثًا بالمهارات الهشة.

10. أثر العالم: الحقيقة التشغيلية النهائية

في الأنظمة اللغوية يمكن لنص مقنع أن يخفي فشلًا فعليًا. World Effect يعكس المعيار: يُعرّف النجاح عبر انتقال حالة خارجي verified state transition. ليست المسألة أن يعلن الوكيل أنه أرسل الملف، بل أن يظهر التغيير في النظام المقصود؛ وليس أن يقول إن الكود أُصلح، بل أن تمر الاختبارات؛ وليس أن يتوقع تحسن metric، بل أن يُقاس بعد intervention.

WEF = Similarity(ExpectedDeltaWorld, ObservedDeltaWorld) * AttributionConfidence (7)

كل فعل عالي الأثر ينتج «عقد أثر» Effect Contract يحوي: preconditions، expected delta، forbidden deltas، observables، timeout، rollback، attribution test. هذا العقد يجعل العالم جزءًا من البرهان، لا مجرد مسرح للتنفيذ.

11. المعمارية المقترحة: COROC-WEI

[D] تتكون المعمارية من تسع طبقات مترابطة. ليست Multi-agent theater؛ بل فصل وظيفي يسمح بالقياس والاستبدال والـablation:

1. **Perception & Evidence Ingest** - تحويل الملاحظات والمصادر والأدوات إلى claims مرتبطة بالمصدر والزمن والثقة.
2. **Claim/Evidence Ledger** - سجل: claim، support، counterevidence، calibration، dependencies، expiration.
3. **Causal World Model** - تمثيل ما الذي يغير ماذا، ومتى يفشل هذا الارتباط، مع counterfactual rollouts.
4. **Frontier Generator** - توليد فرضيات وأسئلة ومهارات وبدائل تتجاوز الحل الأول.
5. **Cognitive Offensive Court** - اختيار أخطر assumptions وتوليد adversarial tests/counterexamples.
6. **Compression Kernel** - استخراج minimal sufficient causal state وfailure signatures.
7. **Trust Gate** - منع الانتقال إلى الفعل الخارجي حتى تتحقق thresholds للوثوق، القيود، وقابلية الرجوع.
8. **World-Effect Executor** - تنفيذ least-privilege مع effect contract وinstrumentation.
9. **Post-Effect Auditor** - مقارنة attribution، predicted/observed، تحديث النموذج والذاكرة والثقة.

12. محكمة الهجوم المعرفي: من "ناقد" إلى بروتوكول خصومة

يمكن أن تفشل self-critique لأن الناقد يولد من نفس prior ونفس السياق. لذلك تقترح المحكمة خمس قنوات مستقلة نسبيًا. أولًا: Contradiction Search يبحث عن أدلة خارجية مضادة. ثانيًا: Model Rivalry ينشئ تفسيرين سببيين على الأقل ويبحث عن تجربة تفصل بينهما. ثالثًا: Boundary Attack يختبر الحالات الطرفية والمقاييس القصوى. رابعًا: Temporal Attack يسأل كيف يتغير الحكم مع الزمن أو ترتيب الأفعال. خامسًا: Tool Reality Check يتحقق من أن الأداة نفذت الفعل وحصلت state transition المقصودة.

[D] لكل claim حرج تصدر المحكمة واحدة من أربع حالات: PROVISIONAL، VERIFIED-ENOUGH-FOR-LOW-IMPACT، VERIFIED-FOR-ACTION، أو BLOCKED. لا وجود لحالة «TRUE» مطلقة، لأن الحكم مرتبط بسياق وأثر ووقت. هذه الصياغة تمنع انتقال الثقة اللغوية إلى يقين تشغيلي غير مستحق.

13. الخوارزمية: Offensive-Verify-Compress-Act Loop

Algorithm 1 - OVCAL

المدخل: مهمة τ، حالة معرفية K_t، قيود S، وميزانية B. المخرجات: فعل موثوق أو امتناع/تصعيد معلل، مع تحديث للحالة المعرفية وسجل الأثر.

1) ابنِ أو حدّث Claim Graph وحدد Critical Claims التي تتوقف عليها الخطة أو الأفعال غير القابلة للعكس.

2) احسب Epistemic Attack Priority لكل claim وكل transition حساس.

3) ولّد بدائل وخططًا ومنافسين سببيين؛ لا تسمح للخطة الأولى بأن تصبح prior مسيطرة بلا منافس.

4) اختر الاختبار المعرفي a^e ذي أعلى Expected Risk Reduction / Cost عندما تتجاوز هشاشة claim العتبة.

5) حدّث beliefs وcalibration؛ ارفض الخطة أو أعد صياغتها إذا ظهر counterexample قاتل.

6) اضغط التاريخ إلى minimal causal state مع failure signatures وprovenance.

7) أنشئ Effect Contract للفعل المرشح a^w يحدد expected/forbidden deltas وrollback وverifier.

8) مرر الفعل عبر Trust Gate والقيود الرسمية وleast privilege.

9) نفّذ ثم قس delta world من حالة النظام، لا من صياغة رد الأداة فقط؛ احسب World-Effect Fidelity.

10) إذا كان mismatch كبيرًا، ارجع upstream إلى assumption أو causal edge المسبب، لا إلى آخر خطوة فقط.

11) إذا تحقق الهدف والقيود فانهِ؛ وإلا وسّع frontier وأعد الدورة بميزانية تكيفية.

14. لماذا الهجوم المعرفي يسبق التخطيط؟

التخطيط التقليدي يفترض غالبًا أن problem formulation صالحة. لكن في المهام المفتوحة قد يكون الخطأ في صياغة الهدف، أو في تعريف النجاح، أو في خريطة الأدوات، أو في افتراض أن observation كاملة. لذلك يجب أن يسبق planning طور "Problem Fracture": ما الذي لو كان غير صحيح سيجعل الخطة كلها بلا معنى؟ ما الذي لا نراه؟ ما القرار الذي لا يمكن التراجع عنه؟ ما evidence الذي سيغير قرارنا فعليًا؟

[H] نتوقع أن Problem Fracture يقلل catastrophic branch commitment في المهام الغامضة حتى لو زاد latency. ويجب اختبار هذا لا بالانطباع، بل بقياس نسبة الخطط التي تُلغى مبكرًا بسبب اكتشاف assumption خاطئ مقابل تكلفة الإلغاء بعد execution.

15. مقاييس جديدة للعلم التجريبي المقترح

Offensive Epistemic Yield (OEY): كمية الانخفاض في الجهل الحرج أو المخاطر لكل وحدة تكلفة اختبار.

Counterexample Discovery Rate (CDR): نسبة الادعاءات الحرجة التي اكتُشف لها counterexample قبل الفعل بدل بعده.

World-Effect Fidelity (WEF): مدى تطابق الأثر المتوقع والمت observed بعد attribution.

Trust Calibration Error (TCE): فجوة confidence vs empirical success، خاصة عند مستويات الأثر المختلفة.

Compression-to-Control Ratio (CCR): تقليل حجم الحالة/الذاكرة نسبةً إلى خسارة القدرة على prediction/control.

Trustworthy Frontier Expansion (TFE): نسبة القدرات الجديدة التي حصلت على تحقق مناسب لأثرها.

Recovery Depth (RD): كم طبقة upstream يعود إليها الوكيل بعد failure؛ patch سطحي أم تعديل assumption/causal model.

Irreversible Error Rate (IER): أخطاء تتجاوز rollback أو تُحدث cost دائمًا لكل 100 قرار عالي الأثر.

Reliability Decay Slope (RDS): انحدار النجاح المتكرر مع زيادة intrinsic horizon.

Epistemic Attack Precision (EAP): نسبة الاختبارات المختارة التي تغيّر قرارًا أو تخفض مخاطرة فعلية بدل إهدار compute.

16. حزمة Benchmarks مقترحة: COROC-Bench

[D] لا يكفي إعادة قياس answer accuracy. يجب إنشاء benchmark يتلاعب عمدًا بدرجة الرصد، وطول الأفق، وقابلية الرجوع، وتلوث الأدلة. يتكون COROC-Bench من ست عائلات:

Ambiguous Objective - تعليمات لها تفسيران مع آثار مختلفة؛ النجاح هو اكتشاف ambiguity قبل الفعل.

Hidden Critical Variable - متغير غير ظاهر يغيّر القرار؛ يُكافأ الاستعلام الذي يكشفه.

Deceptive Confirmation - أدلة كثيرة مؤيدة لكن counterexample واحد حاسم؛ يقيس مقاومة confirmation bias.

Long-Horizon Drift - نفس الخطة تمتد لعشرات/مئات الخطوات مع تغير الحالة؛ يقيس reliability decay.

Irreversible Fork - مسار منخفض الفائدة قابل للرجوع ومسار أعلى فائدة لكنه عالي المخاطر؛ يقيس risk-sensitive epistemic action.

World-Effect Mismatch - واجهة/أداة تدعي نجاحًا لكن state لم يتغير أو تغير جزئيًا؛ يقيس effect verification.

يجب مقارنة على الأقل: Baseline ReAct، Reflexion-like، planner-only، verifier-only، COROC-WEI كامل؛ ثم ablations تحذف المحكمة أو الضغط أو world-effect audit أو trust gate.

17. فرضيات قابلة للتكذيب

- H1: عند ثبات compute budget، سيرفع الهجوم المعرفي الموجه Critical-Claim pass^k في المهام طويلة الأفق أكثر من زيادة depth للتخطيط وحده.
- H2: سترتفع تكلفة latency في COROC-WEI للمهام منخفضة المخاطر، لذلك يجب أن يكون التخصيص تكيفيًا لا دائمًا؛ إذا لم يظهر هذا trade-off فالنظرية مبسطة أكثر من اللازم.

• H3: ضغط failure signatures سيحسن transfer إلى مهام جديدة أكثر من تخزين transcripts كاملة عند حد ذاكرة ثابت.

• H4: World-Effect Audit سيقلل false-success rate بوضوح في البيئات التي تفصل بين tool response و ground-truth state.

• H5: تنويع قنوات الخصومة سيكتشف أخطاء لا يكتشفها self-critique أحادي النموذج، لكن فائدته ستتناقص عندما تشترك القنوات في نفسه model family/المصدر.

• H6: Trustworthy Frontier Expansion سيكون predictor أفضل للأداء طويل الأفق من raw skill-count في بيئات distribution shift.

• H7: عندما تكون أفعال البيئة reversible ورخيصة، ستنخفض قيمة الهجوم المعرفي المسبق وقد يصبح "act-to-learn" أفضل؛ وهذا حد صريح للنظرية لا استثناء تجميلي.

18. تصميم التجارب

18.1 التجربة **A** - الوثوقية عبر الأفق

أنشئ tasks متشابهة بنيويًا مع horizons $H \in \{5, 20, 50, 100\}$. لكل task نفذ 20 seeds على الأقل. قارن success، pass^k، IER، variance، RDS. الاختبار الحاسم ليس هل COROC أعلى في H=5، بل هل slope تدهوره أقل. إذا لم يقل decay مع بقاء compute متقارب، تفشل الفرضية الأساسية.

18.2 التجربة **B** - جدوى الهجوم المعرفي

أدخل hidden variable يمكن كشفه باستعلام منخفض التكلفة. قس نسبة المرات التي يكتشف فيها الوكيل أن الاستعلام يستحق تكلفته، ونسبة false queries. الهدف إثبات أن Epistemic Attack Precision أعلى من curiosity عامة غير مرتبطة بالأثر.

18.3 التجربة **C** - الضغط

اعط الوكيل stream طويلًا من النجاحات والفشل. قارن ذاكرة causal+failure signatures، causal state، summary، raw عند budgets ثابتة. اختبر recovery speed، plan success، prediction، وcatastrophic forgetting. إذا كان الضغط المقترح لا يحسن control per bit، فهو مجرد إعادة تسمية للتلخيص.

18.4 التجربة **D** - أثر العالم

استخدم بيئة تسمح للأداة بإرجاع success token بينما ground truth أحيانًا لا يتغير. الوكيل القائم على اللغة فقط سيحقق apparent success أعلى من real success، أما COROC فيفترض أن يقلل gap عبر state-based verification.

18.5 التجربة **E** - المحكمة

صمم claims لها ثلاثة أنواع failure: factual contradiction، causal confound، temporal constraint. نفذ ablation لكل قناة خصومة. النتيجة المطلوبة ليست أن المحكمة «تنقد أكثر»، بل أن كل قناة تكتشف failure class مميزًا وتزيد downstream reliability بما يبرر تكلفتها.

19. السلامة: لماذا "هجومي" لا يعني "أكثر استقلالية بلا قيد"

أخطر سوء فهم هو مساواة الهجومية المعرفية بتعظيم agency أو القدرة على تجاوز الحواجز. العكس هو المقصود: الهجوم موجه إلى تمثيل الوكيل وثقته وخطته، بينما الفعل الخارجي يظل محكومًا بـleast privilege وhard constraints. كلما ارتفع أثر العالم، يجب أن ترتفع صرامة verification، لا جرأة التنفيذ.

[D] نقترح فصلًا دستوريًا بين ثلاثة budgets: Epistemic Budget يمكن أن يكون واسعًا داخل sandbox، Simulation Budget يسمح بتجارب افتراضية، World-Effect Budget هو الأضيق ويحتاج authorization/constraints أقوى. هذا يمنع أن تتحول open-endedness إلى expansion غير مضبوط للصلاحيات.

وتفيد أعمال formal tool safety الحديثة في هذا السياق: القيود على تسلسل الأدوات وتدفق البيانات يمكن أن تصبح جزءًا من Trust Gate، لكن لا ينبغي افتراض أن formal verifier يحل specification ambiguity. يجب أن تُهاجم المواصفة نفسها معرفيًا قبل التحقق منها شكليًا [13-14].

20. الفشل الداخلي المحتمل للنظرية

شلل التحقق **Verification Paralysis**: الوكيل يستهلك budget في اختبار لا ينتهي. العلاج: value-of-information stopping rules وimpact-adaptive thresholds.

خصومة زائفة **Synthetic Skepticism**: توليد اعتراضات لغوية لا تغير قرارًا. العلاج: EAP وrequire executable discriminative tests.

ضغط مُفقِد **Lossy Compression**: إسقاط متغير حاسم. العلاج: control-preservation tests وperiodic replay من raw archive.

تلوث المحكمة **Shared Blind Spot**: النقاد يشتركون في النموذج والبيانات. العلاج: heterogeneity في أدوات الإثبات والمصادر وformal checks.

Goodhart على **World Effect**: تحسين observable ضيق يفسد الهدف الحقيقي. العلاج: multi-observable contracts وcausal attribution وconstraint auditing.

Over-conservatism: وثوقية مفرطة تمنع الاكتشاف. العلاج: فصل sandbox exploration عن real-world effect، ورفع freedom حيث reversibility عالية.

Model Collapse from Self-Memory: الفشل المتكرر يصبح bias دائمًا. العلاج: expiration, provenance, counterexamples-to-failure-memory.

21. علاقة الإطار بـ **AGI**: من كثرة المهارات إلى توسع موثوق في فضاء الفعل

[I] إذا كانت AGI تعني قدرة عامة على نقل المعرفة والعمل عبر مجالات جديدة، فالاختناق ليس فقط في عدد المهارات، بل في القدرة على معرفة متى لا تنطبق المهارة، وكيف نكتشف ذلك قبل إحداث أثر، وكيف نضغط الدروس في بنى قابلة للنقل. لذلك تقترح الورقة أن generality الحقيقية تحتاج «آلة حدود» Boundary Machine: نظامًا يكتشف حدود صلاحية تمثيلاته بصورة نشطة.

وفق هذا التصور، الذكاء الفتّاح لا يُقاس بما يستطيع ابتكاره فقط، بل بسرعة توسيع frontier مع المحافظة على ثلاثة invariants: وضوح provenance، معايرة الثقة، وإمكانية إسناد الأثر. النظام الذي ينتج ألف مهارة بلا قياس حدودها قد يكون أكثر قدرة لحظيًا وأقل عمومية عمليًا لأنه يراكم fragility.

22. صياغة نظرية مكثفة

مبدأ COROC-WEI

الذكاء الوكيلي العام الموثوق هو عملية توسع معرفي مستمر تُوجّه فيها الموارد أولًا نحو اكتشاف الأخطاء ذات أعلى أثر متوقع، ثم تُضغط المعرفة المصححة إلى حالة سببية كافية، ولا تتحول إلى فعل خارجي إلا عبر قيود تحقق متناسبة مع اللارجعة، وبعد الفعل يُستخدم الفرق بين الأثر المتوقع والمشاهد كإشارة لتحديث الاعتقاد والثقة والحدود نفسها.

يمكن اختزال الدورة في معادلة مفاهيمية:

$$\text{Intelligence_Growth} = \text{Open}(\text{Hypotheses}) \rightarrow \text{Attack}(\text{CriticalUncertainty}) \rightarrow \text{Compress}(\text{CausalState}) \rightarrow \text{Verify}(\text{Action}) \rightarrow \text{Effect}(\text{World}) \rightarrow \text{Audit}(\text{Delta}) \quad (8)$$

وهذا يعيد تعريف «التفكير» في الوكيل: التفكير ليس سيلًا من الرموز، بل إدارة نشطة لمسافة بين نموذج داخلي وعالم خارجي، مع آليات تمنع الثقة من الجري أسرع من الدليل والقدرة من الجري أسرع من التحقق.

23. تنبؤات بعيدة المدى وبرنامج بحث

- بناء Claim/Evidence Ledger معيارية يمكنها العمل فوق LLMs مختلفة دون تعديل الأوزان.
- إنشاء Epistemic Attack Policy تتعلم من نتائج الاختبارات أي نوع نقد يخفض المخاطر فعلًا.
- تعلم Compression Kernel بobjective يحافظ على control/risk لا next-token reconstruction فقط.
- دمج runtime formal monitors مع Effect Contracts بدل التحقق النصي.
- تطوير benchmark لتمييز capability عن reliable agency عبر horizons وirreversibility.
- دراسة العلاقة بين open-ended skill generation وgrowth of verification surface.
- تعريف "epistemic debt": حجم assumptions غير المختبرة التي تعتمد عليها مهارات/خطط لاحقة.
- بناء مؤشرات مبكرة للmeltdown قبل وقوعه من تراكم plan branching، retries، uncertainty، وworld-model mismatch.
- اختبار ما إذا كانت failure signatures المضغوطة تنتقل بين المجالات أفضل من episodic transcripts.
- تصميم governance layer يفصل حق الوكيل في التفكير/المحاكاة عن حقه في إحداث أثر خارجي.

24. حدود الورقة

هذا العمل نظرية تصميمية ولم يُنفذ كمنظومة كاملة بعد. بعض مكوناته مستند إلى تقاليد بحثية مثبتة جزئيًا - active information - executable benchmarks، tool verification، world models، seeking - لكن التجميع المقترح والمقاييس الجديدة والادعاء بأن falsification-oriented allocation يقلل reliability decay هي فرضيات تحتاج تجارب مباشرة. كذلك لا يوجد حتى الآن

تعريف متفق عليه لـAGI أو open-endedness أو trustworthy agency، لذلك يجب تفسير المصطلحات هنا ضمن التعريفات التشغيلية المقدمة لا بوصفها إجماعًا ميدانيًا.

كما أن formal guarantees محدودة بسلامة specification والنموذج المرصود، وWorld-Effect audit لا يحل causal attribution تلقائيًا، والضغط قد يخفي rare events، والبحث الهجومي قد يكون مكلفًا حسابيًا. قوة النظرية ستُحسم لا بثراء اللغة، بل بقدرتها على إنتاج تنبؤات تخسر أمام baselines أو تتفوق عليها بوضوح.

25. الخلاصة

تقترح الورقة تحولًا في صورة الوكيل الذكي: من كائن يحاول أن يكون صحيحًا إلى منظومة تُنظم احتمال كونها مخطئة. «الهجومية المعرفية» هي آلية هذا التحول. فهي لا تضيف ناقدًا بعد الإجابة، بل تجعل كشف قابلية الخطأ جزءًا من policy؛ ولا تجعل الضغط اختصارًا للنص، بل حفظًا للسببية اللازمة للفعل؛ ولا تجعل الانفتاح تراكمًا للمهارات، بل توسعًا يرافقه اتساع في حدود التحقق؛ ولا تجعل العالم وجهة نهائية، بل جهاز قياس يعود منه الأثر ليحاكم النموذج.

إذا صمدت الفرضيات تجريبيًا، فالمكسب النظري سيكون مهمًا: الوثوقية لن تعود طبقة أمان تُضاف فوق الذكاء، بل طريقةً في إنتاج الذكاء نفسه. عندها يصبح الوكيل الأقوى هو الذي يستطيع أن يفتح فضاء الإمكان بقوة، ويضغطه بذكاء، ويهاجم نقاط عماه بلا رحمة، ثم لا يسمح لنفسه بتغيير العالم إلا بمقدار ما يستطيع قياسه وتبريره واسترجاعه.

المراجع المختارة

[1] Zhou, S. et al. (2023). WebArena: A Realistic Web Environment for Building Autonomous Agents. arXiv:2307.13854.

[2] Xie, T. et al. (2024). OSWorld: Benchmarking Multimodal Agents for Open-Ended Tasks in Real Computer Environments. arXiv:2404.07972.

[3] Yao, S., Shinn, N., Razavi, P., Narasimhan, K. (2024). τ-bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains. arXiv:2406.12045.

[4] Khanal, A., Zhou, J., Tao, Y. (2026). Beyond pass@1: A Reliability Science Framework for Long-Horizon LLM Agents. arXiv:2603.29231.

[5] Yao, S. et al. (2022/2023). ReAct: Synergizing Reasoning and Acting in Language Models. ICLR 2023 / arXiv:2210.03629.

[6] Shinn, N. et al. (2023). Reflexion: Language Agents with Verbal Reinforcement Learning. NeurIPS 2023 / arXiv:2303.11366.

[7] Yao, S. et al. (2023). Tree of Thoughts: Deliberate Problem Solving with Large Language Models. arXiv:2305.10601.

[8] Wang, G. et al. (2023). Voyager: An Open-Ended Embodied Agent with Large Language Models. arXiv:2305.16291.

[9] Hafner, D., Pasukonis, J., Ba, J., Lillicrap, T. (2025). Mastering diverse control tasks through world models. Nature 640, 647-653.

[10] Sajid, N., Da Costa, L., Parr, T., Friston, K. (2021). Active inference, Bayesian optimal design, and expected utility. arXiv:2110.04074.

[11] Wei, R. (2024). Value of Information and Reward Specification in Active Inference and POMDPs. arXiv:2408.06542.

[12] Debenedetti, E. et al. (2024). AgentDojo: A Dynamic Environment to Evaluate Prompt Injection Attacks and Defenses for LLM Agents. arXiv:2406.13352.

[13] Doshi, A. et al. (2026). Towards Verifiably Safe Tool Use for LLM Agents. arXiv:2601.08012.

[14] Dantas, P., Cordeiro, L., Nowroozi, E., Norbert, T. (2026). Toward Safe LLM Agents: A Survey of Specification, Verification, and Enforcement. arXiv:2608.14590.

[15] Mialon, G. et al. (2023). GAIA: a benchmark for General AI Assistants. arXiv:2311.12983.

[16] de Vries, B. et al. (2025). Expected Free Energy-based Planning as Variational Inference. arXiv:2504.14898.

[17] Rawles, C. et al. (2024). AndroidWorld: A Dynamic Benchmarking Environment for Autonomous Agents. arXiv:2405.14573.

[18] Akshathala, S. et al. (2025). Beyond Task Completion: An Assessment Framework for Evaluating Agentic AI Systems. arXiv:2512.12791.

ملحق A - بطاقة تنفيذ بحثي سريعة

MVP تجريبي

ابدأ بوكيل أدوات عادي. أضف Claim Ledger بسيطًا، ثم Epistemic Attack Scorer، ثم بوابة تمنع tool calls عالية الأثر إذا كان critical claim غير مختبر، ثم Effect Auditor يقارن الحالة قبل/بعد. اختبر أولًا على بيئة sandbox قابلة للإعادة. لا تبدأ بmulti-agent؛ أثبت قيمة كل مكوّن بablation.

B - Minimal Data Schema ملحق

$$\text{Claim} = \{\text{id, proposition, source, timestamp, confidence, impact, dependencies, counterevidence, status}\} \quad \text{(A1)}$$

$$\text{Action} = \{\text{intent, preconditions, expected_delta, forbidden_delta, reversibility, verifier, rollback}\} \quad \text{(A2)}$$

$$\text{FailureSignature} = \{\text{trigger, violated_assumption, observed_delta, root_cause, repair, transfer_scope}\} \quad \text{(A3)}$$

ملحق **C** - معيار قرار مبسط

قبل أي action مهم، اسأل خوارزميًا: (1) ما critical claim الذي إذا كان خاطئًا يغيّر القرار؟ (2) هل توجد observation منخفضة التكلفة تميّز بين الفرضيات؟ (3) هل الفعل reversible؟ (4) هل لدينا verifier يقيس state لا النص؟ (5) هل confidence calibrated لهذا النوع من المهام؟ إذا فشل اثنان أو أكثر من هذه الشروط، يجب خفض الصلاحية أو زيادة الاختبار قبل التنفيذ.

— نهاية النسخة التأسيسية —