



Large Language Model – Based AI Agents in Engineering Applications: An Analytical Review of Architectures, Evaluation, and Safety

وكلاء الذكاء الاصطناعي
المعتمدون على النماذج اللغوية
الكبيرة في التطبيقات الهندسية
مراجعة تحليلية للمعماريات
والتقييم والأمان

بقلم المهندس عمر ياسر موسى

الملخص

أدى التطور المتسارع في النماذج اللغوية الكبيرة إلى انتقال تطبيقات الذكاء الاصطناعي من أنظمة تكتفي بإنتاج النصوص والإجابات إلى أنظمة قادرة على التخطيط، واختيار الأدوات، وتنفيذ الأفعال، ومراقبة نتائجها، وتعديل سلوكها للوصول إلى هدف محدد. يُشار إلى هذه الأنظمة باسم وكلاء الذكاء الاصطناعي، وقد بدأت بالظهور في تطبيقات هندسية تشمل تطوير البرمجيات، والتصنيع، ونمذجة معلومات البناء، ومحاكاة أنظمة القدرة الكهربائية، وهندسة المواد. إلا أن الانتقال من توليد الإجابة إلى تنفيذ الإجراء يضيف أخطاراً جديدة تتعلق بتراكم الأخطاء، وسوء استخدام الأدوات، واتساع الصلاحيات، وحقن التعليمات، وتسرب البيانات، وصعوبة تحديد المسؤولية عن القرارات.



تهدف هذه المقالة إلى تقديم مراجعة تحليلية موجهة لوكلاء الذكاء الاصطناعي المعتمدين على النماذج اللغوية الكبيرة في التطبيقات الهندسية، من خلال توضيح مفهوم الوكيل وتمييزه عن المساعد الحواري والاسترجاع المعزز بالتوليد وسير العمل المحدد سابقاً، وتحليل مكونات البنية الوكيلية وأنماط الوكيل الواحد والأنظمة متعددة الوكلاء، واستعراض نماذج تطبيقية في عدد من التخصصات الهندسية، ومناقشة منهجيات التقييم والمخاطر الأمنية والتشغيلية. كما تقترح المقالة إطاراً مفاهيمياً لتطوير «وكيل هندسي موثوق» يتكون من سبع طبقات مترابطة تشمل:

1. حوكمة المهمة وعقدها،
 2. المعرفة الهندسية وإثبات مصدرها،
 3. التخطيط المحدود،
 4. الأدوات المقيدة بالصلاحيات،
 5. التنفيذ المعزول،
 6. التحقق المستقل والإشراف البشري المتناسب مع مستوى المخاطر،
 7. الرصد والتتبع التشغيلي والاستجابة للحوادث.
- تُظهر الأدلة المتاحة أن قوة النموذج اللغوي وحدها لا تكفي لبناء وكيل هندسي موثوق، وأن أفضل النتائج تتحقق غالباً عند دمج النموذج مع معرفة هندسية موثوقة، وأدوات حتمية أو قابلة للاختبار، وآليات تغذية راجعة، وقيود واضحة للصلاحيات.



وتخلص المقالة إلى ضرورة تقييم الوكلاء على أساس نجاح المهمة الكاملة وسلامتها وكلفتها وقابليتها للتدقيق، لا على أساس جودة النص الناتج فقط.

الكلمات المفتاحية: وكلاء الذكاء الاصطناعي؛ النماذج اللغوية الكبيرة؛ الذكاء الاصطناعي الوكيل؛ الأنظمة متعددة الوكلاء؛ التطبيقات الهندسية؛ موثوقية الذكاء الاصطناعي؛ أمان الذكاء الاصطناعي.

1. المقدمة

أظهرت النماذج اللغوية الكبيرة قدرة مرتفعة على فهم التعليمات وإنتاج النصوص وكتابة الشفرة وتلخيص الوثائق والإجابة عن الأسئلة. ومع ذلك، فإن معظم المسائل الهندسية الواقعية لا تُحل بإجابة نصية واحدة، بل تحتاج إلى سلسلة من الخطوات تشمل جمع البيانات، واختيار طريقة الحل، واستدعاء برمجيات أو قواعد بيانات، وتشغيل محاكاة، وتحليل النتائج، واكتشاف الأخطاء، وإعادة التنفيذ عند الحاجة. من هنا ظهر مفهوم وكلاء الذكاء الاصطناعي المعتمدين على النماذج اللغوية الكبيرة، حيث يعمل النموذج اللغوي بوصفه مكوّنًا لاتخاذ القرار ضمن منظومة أوسع تضم الذاكرة والأدوات ومصادر المعرفة وبيئة التنفيذ وآليات التحقق.

مثل أسلوب ReAct إحدى اللبنات المبكرة لهذا التوجه، إذ جمع بصورة متعاقبة بين الاستدلال واتخاذ الفعل والتفاعل مع البيئة. كما بيّن Toolformer إمكان تدريب النموذج على تحديد وقت استدعاء الأدوات الخارجية، والأداة المناسبة، والوسائط



التي ينبغي تمريرها إليها. وبذلك لم يعد النموذج مولدًا للغة وحسب، بل أصبح قادرًا على استخدام الحاسبة ومحركات البحث والواجهات البرمجية ومصادر البيانات ضمن عملية حل المهمة.

لا تعني هذه القدرات أن النظام أصبح موثوقًا بصورة تلقائية. فعندما ينتج نموذج لغوي معلومة غير صحيحة، قد يقتصر الضرر على إجابة مضللة. أما عندما يمتلك الوكيل صلاحية تشغيل شفرة أو تعديل قاعدة بيانات أو إرسال أوامر إلى برنامج هندسي أو نظام صناعي، فإن الخطأ قد يتحول إلى فعل له أثر مادي أو مالي أو تشغيلي؛ لذلك تتضاعف أهمية التحقق من المدخلات والنتائج، وتحديد الصلاحيات، وعزل بيئة التنفيذ، ووضع نقاط موافقة بشرية قبل الإجراءات الحساسة.

تزداد هذه الأهمية في المجالات الهندسية لأن القرارات فيها تخضع عادةً لقيود فيزيائية ومعايير سلامة وأكواد تصميم وإجراءات ضبط جودة. فقد تبدو إجابة الوكيل منطقية لغويًا، في حين تحتوي على قيمة غير صحيحة لوحدة قياس، أو تستخدم إصدارًا قديمًا من معيار هندسي، أو تختار دالة برمجية غير مناسبة، أو تفسر مخرجات المحاكاة بصورة خاطئة. ولهذا لا يكفي تقييم الوكيل بالاعتماد على طلاقة اللغة أو التشابه مع إجابة مرجعية، بل يجب تقييم نجاح المهمة وظيفيًا، والتحقق من كل فعل اتخذته النظام، وقياس قدرته على اكتشاف الفشل والتعافي منه.

تقدم هذه المقالة أربع إسهامات رئيسية؛ أولًا، تقترح تعريفًا تشغيليًا لوكيل الذكاء الاصطناعي وتوضح الفرق بينه وبين المساعد الحواري ونظام الاسترجاع المعزز



بالتوليد وسير العمل المحدد سابقًا. ثانيًا، تحليل المكونات والأنماط المعمارية الأساسية للوكلاء. ثالثًا، تستعرض تطبيقات ممثلة في عدد من المجالات الهندسية، وتستخلص منها عوامل النجاح والقيود. رابعًا، تقترح إطارًا مفاهيميًا لبناء وكيل هندسي موثوق، يربط درجة الاستقلالية بإمكان التحقق من المهمة وحجم الضرر المحتمل عند وقوع الخطأ.

2. منهجية المراجعة ونطاقها

اعتمدت المقالة مراجعة تحليلية منظمة وليست تحليلًا تجميعيًا إحصائيًا (Meta-analysis). وقد صُممت المنهجية بحيث يمكن إعادة تنفيذها وتحديثها في دراسة لاحقة. وتم الاسترشاد بمبادئ الإبلاغ المنهجي في PRISMA 2020، ولا سيما توثيق إستراتيجية البحث ومعايير الإدراج والاستبعاد ومسار اختيار الدراسات؛ مع التأكيد أن المقالة الحالية ليست مراجعة منهجية شاملة لجميع قواعد البيانات.

1.2. مصادر البحث وإستراتيجية الاسترجاع

شملت إستراتيجية البحث قواعد ومصادر علمية مناسبة لطبيعة الموضوع، هي: SpringerLink، ScienceDirect، ACM Digital Library، IEEE Xplore، Google Scholar، إضافة إلى arXiv لتتبع الأعمال الحديثة جدًا عندما لا يتوافر إصدار محكم بعد. كما استخدم التتبع المرجعي للأعمال الأساسية والمواصفات الرسمية الصادرة عن NIST و ISO والبروتوكولات ذات الصلة.

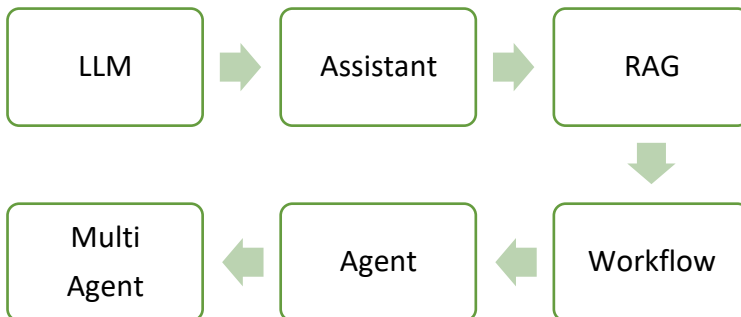


استخدمت الإستراتيجية تراكييب بحث تجمع بين مصطلحات الوكلاء والنماذج اللغوية والتطبيقات الهندسية، ومن أمثلتها:

- "LLM agents" AND engineering`
- "LLM-based agents" AND engineering applications`
- "`agentic AI" AND engineering`
- "`large language model" AND "engineering design" AND agent`
- "`multi-agent" AND engineering AND LLM`
- "LLM agent" AND safety AND tools`
- "LLM agents" AND evaluation AND benchmark`

امتدت العينة المفاهيمية من الأعمال التأسيسية التي تشرح الوكلاء واستخدام الأدوات إلى الأعمال التطبيقية الحديثة المتاحة حتى سبتمبر/أيلول 2026، مع إعطاء الأولوية للأعمال المنشورة أو المتاحة بصورة موثقة قبل تاريخ إعداد النسخة المعدلة.

الشكل 1. التطور المفاهيمي من النموذج اللغوي إلى الأنظمة متعددة الوكلاء





2.2. معايير الإدراج والاستبعاد

أُدرجت الدراسة أو الوثيقة إذا حققت شرطين على الأقل من شروط الصلة التالية، أو كانت معيارًا أو مواصفة أساسية لا يمكن تحليل موضوع الوكلاء الموثوقين من دونها

1. استخدام نموذج لغوي كبير مكوّنًا مركزيًا في التخطيط أو اتخاذ القرار.
2. وجود استخدام فعلي لأداة أو واجهة برمجية أو تنفيذ شفرة أو تفاعل مع بيئة خارجية.
3. تقديم تطبيق هندسي أو علمي أو حالة استخدام قابلة للتقييم.
4. تحليل مخاطر أمنية أو تشغيلية خاصة بالأنظمة الوكيلية.
5. تقديم إطار معماري أو مواصفة تشغيل بيئي ذات صلة مباشرة ببنية الوكيل أو أدواته.

استُبعدت:

- أ- الأنظمة الحوارية التي لا تتجاوز توليد النص.
- ب- تطبيقات RAG التي لا تتضمن تخطيطًا ديناميكيًا أو تنفيذًا أفعال.
- ج- الأعمال التي تستخدم وصف «Agent» من دون توضيح حلقة القرار والتنفيذ والملاحظة.
- د- المواد التي لا تسمح باستخراج معلومات كافية عن المنهج أو النتائج.
- هـ- التكرارات أو النسخ الأقدم عندما يتوافر إصدار أحدث للمصدر نفسه.



3.2. عملية الفرز واختيار العينة

مرّت عملية الاختيار بالمراحل الآتية: تحديد السجلات، وإزالة التكرارات، وفحص العنوان والملخص، وتقييم النص الكامل، ثم استخراج البيانات وتحليلها. ويُعرض هذا المسار في الشكل المقترح للمراجعة وفق منطوق PRISMA 2020.

ملاحظة منهجية مهمة: النسخة الأصلية من المقالة لم تحتفظ بسجل عددي موثق لعدد النتائج الأولية في كل قاعدة بيانات؛ لذلك لا يجوز اختلاق أرقام بأثر رجعي. ولتجنب إعطاء دقة زائفة، تعتمد النسخة المعدلة على تدقيق استعادي للعينة النهائية الموثقة في قائمة المراجع، مع التمييز بين الدراسات التطبيقية، والمعايير، وأطر الوكلاء، ومصادر الأمن والتقييم. بلغ عدد المصادر الأساسية التي خضعت للتحليل التفصيلي في النسخة الحالية 26 مصدرًا، منها 18 مصدرًا بحثيًا/تطبيقيًا وتقييميًا أساسيًا و8 مصادر معيارية أو حوكمية أو بروتوكولية وتُعد هذه النقطة من حدود المراجعة، ويمكن استكمال سجل البحث العددي في تحديث لاحق إذا طلبت المجلة ذلك.

نظرًا إلى عدم الاحتفاظ في مرحلة البحث الأصلية بسجل آلي موثق يربط كل قاعدة بيانات بعدد السجلات المسترجعة وعدد السجلات المستبعدة في كل مرحلة، لا تُعرض أرقام تقديرية لمراحل التعرف والفحص الأولي. ويُستخدم منطوق PRISMA 2020 هنا إطارًا لتنظيم وصف عملية الاختيار والإدراج والاستبعاد، وليس باعتباره ادعاءً بأن المراجعة الحالية تستوفي جميع متطلبات المراجعة المنهجية المسجلة



سابقاً. وعليه، تُعد العينة النهائية الموثقة في قائمة المراجع أساس التحليل الحالي، مع الإفصاح صراحةً عن هذا القيد المنهجي.

الشكل 2. مسار اختيار الدراسات والمصادر



العينة النهائية للتحليل: 26 مصدرًا

مصدرًا بحثيًا/تطبيقيًا/تقييميًا

مصادر معيارية/حوكومية/بروتوكولية

4.2. تقييم جودة المصادر

لمنع إعطاء وزن متساوٍ لمصادر مختلفة في طبيعتها، استُخدمت بطاقة تقييم نوعية/شبه كمية من سبعة بنود. يمنح كل بند 0 عند غيابه، و0.5 عند وجوده بصورة جزئية، و1 عند توافره بوضوح. البنود هي: وضوح النموذج والمعمارية،



ووصف المهمة، ووصف البيانات أو Benchmark، ووجود Baseline، ووضوح المقاييس، وقابلية إعادة الإنتاج، ووضوح القيود. ولا يُستخدم هذا التقييم لإلغاء المصادر المعيارية، بل لتحديد مقدار الاعتماد على كل مصدر عند استخلاص النتائج.

معيار الجودة	0	0.5	1
وصف النموذج والمعمارية	غير متاح	جزئي	واضح
وصف المهمة/البيئة	غير متاح	جزئي	واضح
Dataset أو Benchmark	غير متاح	جزئي	واضح
Baseline وجود	غير متاح	جزئي	واضح
Metrics وضوح	غير متاح	جزئي	واضح
قابلية إعادة الإنتاج	ضعيفة	متوسطة	جيدة
عرض القيود	غير متاح	جزئي	واضح



5.2. استخراج البيانات والتحليل

استُخرجت من كل دراسة المعلومات الآتية: المجال الهندسي، والنموذج، ونوع الوكيل، ونمط المعمارية، والأدوات، Dataset، Benchmark، Baseline، Metrics، النتيجة الرئيسية، والقيود. ثم صُنفت الدراسات وصفيًا حسب المجال الهندسي، ونوع الوكيل (Single-Agent/Multi-Agent)، ومستوى الاستقلالية (Workflow/Agentic)، ونوع الأدوات، ووجود RAG أو ذاكرة، ونوع Benchmark، وطبيعة التقييم.

لم يُجرَ Meta-analysis إحصائي لأن المقاييس والمهام والنماذج وبيئات الاختبار غير متجانسة. وبدلاً من ذلك استخدم التحليل الوصفي المقارن مع الحفاظ على الأرقام الأصلية لكل دراسة ونسبها ضمن شروطها التجريبية، وعدم جمعها في متوسط واحد يوحي بقابلية المقارنة المباشرة.



القيود	النتيجة الرئيسية	Base line	Datas et	الأدوات	نوع الوكيل	النموذج	المجال	الدراسة
ليس دراسة تجريبية هندسية موحدة	تصنيف معماري وتقييمي	دراسات سابقة	مسح أدبيات	أدوات وذاكرة	إطار مسح	LLM Agents	عام/ هندسي	Wang et al. [1]
لا يختص بالهندسة	تحسن في الاستدلال	نماذج أساسية	مهام متعددة	أدوات	وكيل تفاعلي	LLM + ReAct	عام	ReAct [2]
ليس إطار أمان هندسي	تعلم استدعاء الأدوات	نماذج دون أدوات	مهام أدوات	APIs	استخدام أدوات	LLM	عام	Toolformer [3]
كلفة وتعقيد التنسيق	تنسيق وكلاء	إعدادات مختلفة	حالات تطبيق	أدوات/حوار	Multi-Agent	AutoGen	عام/ برمجي	AutoGen [4]



لا يثبت الجاهزية العامة	توزيع الأدوار	أطر برمجية	مهام برمجية	أدوات تطوير	Multi- Agent	LLMs	هندسة برمجيات	MetaGP T [5]
انتقال الخطأ بين الوكلاء	تعاون الأدوار	أطر برمجية	مهام برمجية	أدوات تطوير	Multi- Agent	LLMs	هندسة برمجيات	ChatDev [6]
ليس هندسيًا تحديدًا	نجاح المهام	نماذج مختلفة	8 بيئات	بيئات متعددة	Agent	LLMs	تقييم وكلاء	AgentBe nch [7]
إعداد تاريخي عام	15% للنظام الأصلي مقابل 92% للإنسان	إنسان نظام آلي	466 سؤالاً	ويب/أدوات	Agent	GPT- 4/Agents	تقييم وكلاء	GAIA [8]



بيئة ويب	14.41% للنظام الأصلي مقابل 78.24% للبشر	بشر/وكلاء	812 مهمة	متصفح	Agent	GPT-4 Agent	تقييم تفاعلي	WebArena [9]
مرتبط بإعدادات وإصدارات محددة	نجاح إصلاح المشكلات	نماذج مختلفة	2294 مشكلة أصلية	مستودعات/اختبارات	Agent	LLM Agents	هندسة برمجيات	SWE-bench [10]
نطاقات عامة	نجاح واتساق عبر التكرار	Baselines متعددة	مهام تفاعل	APIs	Agent	LLM Agents	أدوات/خدمة	τ-bench [11]
تركيز برمجي	تحليل سلوك الوكلاء	Agentless وغيره	SWE-bench تحليل	أدوات برمجية	Agent	LLM Agents	هندسة برمجيات	Xia et al. [12]



بيئة محكمة	100% ،لمسارين 86% لأربع خطوات	إعدادات متعددة	دراسة محاكاة	G- code/موارد	Multi- Agent	LLM	تصنيع	Lim et al. [13]
نموذج أولي	توليد + نماذج تحقق قواعد	نموذج أولي	حالات تصميم	Vectorworks/BIM	Multi- Agent	LLM	BIM	Du et al. [14]
بيئة محكمة	93.13% و96.85% في إعداد الدراسة	إعدادات مقارنة	69 مهمة	MATPOWER/Da line	Multi- Agent	LLM	أنظمة قدرة	Jia et al. [15]

6.2. التحليل التصنيفي

يكشف الترميز السابق أربعة اتجاهات قابلة للمقارنة:

- (1) الانتقال من المساعد إلى الوكيل ذي الأدوات،
- (2) التوسع من الوكيل الواحد إلى تعدد الوكلاء عند الحاجة إلى التخصص أو فصل المسؤوليات،



(3) ازدياد الاعتماد على Benchmarks تفاعلية بدل تقييم النص فقط،

(4) انتقال الأمان من طبقة لاحقة إلى خاصية معمارية تشمل الصلاحيات والذاكرة والأدوات والتنفيذ. وتُناقش هذه الاتجاهات في الأقسام اللاحقة بدل اختزالها في متوسط رقمي غير صالح للمقارنة.

3. الأسس المفاهيمية لوكلاء الذكاء الاصطناعي

1.3. تعريف تشغيلي للوكيل

يمكن تعريف وكيل الذكاء الاصطناعي المعتمد على نموذج لغوي كبير بأنه:

نظام موجه نحو هدف يستخدم نموذجًا لغويًا لفهم المهمة، وتحديد الخطوة التالية، واختيار الأدوات أو الأفعال، واستقبال نتائج البيئة، وتحديث حالته أو خطته، والاستمرار حتى تحقيق شرط نجاح أو توقف محدد.

يتضمن هذا التعريف خمسة عناصر أساسية: وجود هدف، وامتلاك حالة أو سياق، والقدرة على الاختيار بين أفعال، والتفاعل مع بيئة خارجية، واستخدام التغذية الراجعة لتعديل السلوك.

يُستخدم هذا التمييز بوصفه تعريفًا تشغيليًا لأغراض هذه المراجعة، وليس تصنيفًا معياريًا ملزمًا؛ إذ تختلف الحدود بين «الوكيل» و«سير العمل الوكيلي» و«المساعد القادر على استخدام الأدوات» بين الأطر والأدبيات.



لا يتطلب الوكيل أن يكون مستقلاً استقلالاً كاملاً؛ فقد يعمل تحت إشراف بشري، أو تكون صلاحياته مقتصرة على القراءة، أو يحتاج إلى موافقة قبل كل إجراء مؤثر. وعليه، فالاستقلالية ليست صفة ثنائية، بل طيف يبدأ من المساعدة الاستشارية وينتهي بالتنفيذ المستقل المحدود ضمن نطاق مصرح به.

2.3. الفرق بين النموذج والمساعد وسير العمل والوكيل

الجدول 1. مقارنة بين الأنظمة القائمة على النماذج اللغوية:

نوع النظام	من يحدد الخطوات؟	استخدام الأدوات	التفاعل مع نتائج البيئة	درجة الاستقلال	مثال وظيفي
نموذج لغوي	المستخدم أو نص التعليمات	غير ضروري	محدود	منخفضة جداً	توليد شرح أو شفرة
مساعد حواري	المستخدم عبر الحوار	اختياري	محدود غالباً	منخفضة	الإجابة عن أسئلة هندسية



نوع النظام	من يحدد الخطوات؟	استخدام الأدوات	التفاعل مع نتائج البيئة	درجة الاستقلال	مثال وظيفي
استرجاع معزز بالتوليد RAG	مسار استرجاع ثابت	قاعدة معرفة أو محرك بحث	يعيد وثائق للنموذج	منخفضة	الإجابة من كتيب هندسي

قد يحتوي سير العمل على نموذج لغوي وأدوات متعددة، لكنه لا يُعد وكيلاً كاملاً عندما تكون جميع الخطوات والفروع محددة سابقاً. وفي المقابل، يُعد النظام وكيلاً عندما يمتلك حرية فعلية في تحديد الخطوة التالية أو إعادة التخطيط استناداً إلى الملاحظات.

هذا التمييز مهم هندسياً لأن زيادة حرية القرار تؤدي غالباً إلى زيادة صعوبة الاختبار وإعادة الإنتاج. ففي سير العمل الثابت يمكن اختبار كل مسار معروف، في حين قد يولد الوكيل تسلسلات أفعال جديدة لم يجر توقعها أثناء التصميم.



3.3. الذكاء الاصطناعي الوكيل والأنظمة متعددة الوكلاء

يُستخدم مصطلح «الذكاء الاصطناعي الوكيل» لوصف الأنظمة التي لا تكتفي بتوليد محتوى، بل تنفذ سلسلة من الأفعال الموجهة نحو هدف. أما النظام متعدد الوكلاء فيتكون من وكيلين أو أكثر، لكل منهم دور أو أدوات أو معرفة مختلفة، ويتبادلون الرسائل أو المهام أو النتائج.

أظهرت أطر مثل AutoGen إمكان بناء تطبيقات من وكلاء قابلين للتخصيص يتواصلون فيما بينهم ويجمعون بين النماذج والأدوات والمدخلات البشرية. واستخدم MetaGPT أدوارًا متخصصة وإجراءات تشغيل معيارية لمحاكاة سير العمل في تطوير البرمجيات، في حين قسم ChatDev العمل إلى مراحل وأدوار تتعاون من خلال محادثات منظمة.

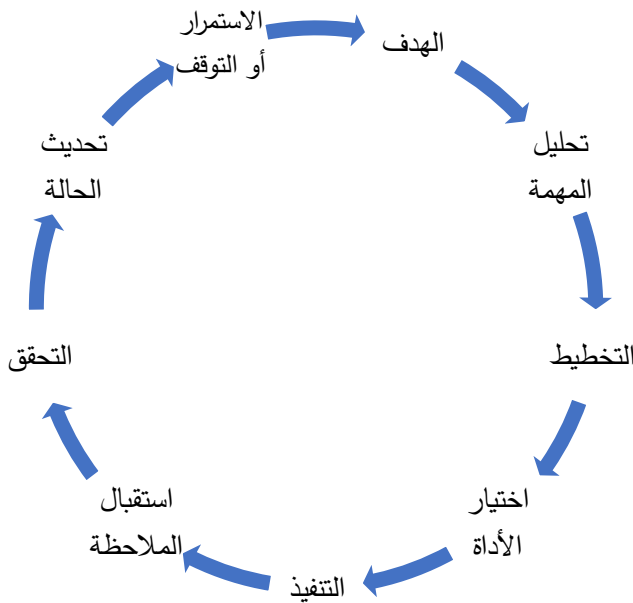
مع ذلك، لا يعني تعدد الوكلاء تحسن الأداء تلقائيًا. فقد تؤدي الرسائل المتكررة إلى تضخم السياق والكلفة، ويمكن أن تنتقل معلومة خاطئة من وكيل إلى آخر وتكتسب مظهرًا من الثقة نتيجة تكرارها. لذلك ينبغي تبرير كل وكيل إضافي بوظيفة قابلة للقياس، مثل امتلاك أداة مختلفة، أو إجراء تحقق مستقل، أو تطبيق قاعدة فصل للصلاحيات.



4. البنية المرجعية لوكيل الذكاء الاصطناعي

1.4. حلقة العمل الوكيلية

الشكل 3: تمثيل دورة عمل الوكيل



يبدأ الوكيل من هدف يقدمه المستخدم أو نظام أعلى. ثم يحلل القيود ويضع خطة أولية، ويختار فعلاً مثل البحث في قاعدة بيانات أو تشغيل برنامج أو إنشاء ملف. بعد التنفيذ، لا ينبغي أن يفترض نجاح الأداة، بل يقرأ النتيجة أو الخطأ ويقرر هل تحققت الغاية، أو يلزم تصحيح الوسائط، أو استخدام أداة أخرى، أو طلب مساعدة بشرية.



يؤدي فصل «التنفيذ» عن «التحقق» دورًا جوهريًا. فقد ينجح استدعاء الأداة تقنيًا، لكنه ينتج نتيجة هندسية غير صحيحة. مثلاً، قد يعمل برنامج المحاكاة من دون أخطاء، في حين تكون الوحدات أو الشروط الحدية غير مناسبة؛ لذلك يجب أن يكون للوكيل معيار نجاح وظيفي يتجاوز مجرد عدم ظهور رسالة خطأ.

2.4. المكونات الأساسية

أ. النموذج الأساسي

يمثل النموذج اللغوي محرك تفسير التعليمات وتوليد الخطط واختيار الأدوات. إلا أن النموذج لا يحتوي بالضرورة على أحدث الأكواد والمعايير الهندسية، كما أن معرفته الضمنية لا توفر ضمانًا لصحة الحساب أو المرجع. لذلك ينبغي النظر إليه بوصفه مكوّنًا احتماليًا لاتخاذ القرار، لا قاعدة معرفة موثوقة بذاتها.

ب. عقد المهمة

عقد المهمة هو وصف منظم يحدد:

- الهدف المطلوب.
- البيانات المسموح باستخدامها.
- الأدوات المتاحة.
- الأفعال المحظورة.
- ميزانية الزمن والكلفة.



- شروط النجاح.

- شروط التوقف.

- الحالات التي تستوجب موافقة بشرية.

يقلل عقد المهمة من احتمال الانحراف عن الهدف أو الاستمرار في دورات غير محدودة.

ج. المعرفة الهندسية الموثقة

ينبغي تزويد الوكيل بمصادر محددة، مثل الكودات والمعايير وكتيبات البرامج والرسومات وقواعد البيانات وسجلات الصيانة. ويجب تسجيل اسم المصدر وإصداره وتاريخه، لأن الاستشهاد بمعلومة صحيحة من إصدار قديم قد يؤدي إلى قرار غير صالح في المشروع الحالي.

د. الذاكرة وإدارة الحالة

تشمل الذاكرة قصيرة الأمد الخطوات والملاحظات الخاصة بالمهمة الحالية، في حين قد تشمل الذاكرة طويلة الأمد تقضيات المستخدم أو حلولاً سابقة أو سجلات تشغيل. ويجب عدم تخزين جميع المخرجات تلقائياً؛ إذ يمكن أن تحتوي الذاكرة على أخطاء أو تعليمات خبيثة أو بيانات حساسة. لذلك تحتاج الذاكرة إلى سياسات صلاحية زمنية ومصدر وتحقق وعزل بين المستخدمين والمشروعات.



هـ. التخطيط

قد يكون التخطيط مباشرًا، أو هرميًا يقسم الهدف إلى مهام فرعية، أو تكراريًا يعدّل الخطة بعد كل ملاحظة. ويساعد فصل المخطط عن المنفذ على تقليل الخلط بين تحديد المهمة وتنفيذها، كما يمكن إضافة مكوّن ناقد أو متحقق لمراجعة النتيجة قبل اعتمادها.

و. سجل الأدوات

يتضمن سجل الأدوات اسم كل أداة ووظيفتها، والمدخلات المقبولة، والصلاحيات، والمخاطر. ويجب أن تكون مخططات المدخلات محددة وقابلة للتحقق، بدل السماح للنموذج بإنتاج أوامر حرة غير مقيدة.

ز. بيئة التنفيذ

ينبغي تنفيذ الشفرة والعمليات عالية الخطورة ضمن بيئة معزولة، مثل حاوية برمجية أو نسخة اختبار أو توعم رقمي. وتُفصل بيئة الاختبار عن الإنتاج، مع منع الوصول غير الضروري إلى الشبكة والملفات والاعتمادات السرية.

ح. التحقق

قد يكون التحقق:

- حسابيًا، بإعادة الحساب بطريقة مستقلة.
- منطقيًا، بمقارنة النتيجة مع قيود المجال.



- وظيفيًا، بتشغيل اختبارات أو محاكاة.
- مرجعيًا، بمقارنة الادعاءات مع مصادر موثقة.
- بشريًا، بعرض القرار على مهندس مختص.
- متعدد النماذج، باستخدام مكوّن مستقل للمراجعة.

لا يُعد سؤال النموذج نفسه عما إذا كانت إجابته صحيحة تحققًا مستقلًا، لأن المكوّن المراجع قد يكرر الخطأ ذاته.

ويقصد بالتحقق المستقل استخدام آلية تختلف في مصدرها أو طريقة عملها عن آلية توليد النتيجة، مثل قاعدة حتمية، أو اختبار برمجي، أو إعادة حساب مستقلة، أو نموذج فيزيائي، أو مراجعة بشرية متخصصة.

ط. الرصد والتدقيق

يجب تسجيل أفعال الوكيل ومدخلات الأدوات ونتائجها والأخطاء ومحاولات الإعادة والكلفة، والزمن، وإصدارات النماذج، والتعليمات. ولا يلزم تخزين عمليات التفكير الداخلية غير القابلة للتحقق؛ فالأهم هندسيًا هو تسجيل ملخص القرار والأدلة والأفعال التي يمكن إعادة تنفيذها.

تُظهر البنى الإنتاجية الحديثة لوكلاء البرمجة اهتمامًا متزايدًا بالتنفيذ المعزول، والتحكم في دورة الحياة، وإدارة الحالة القابلة لإعادة التشغيل، والأدوات ذات الأنواع المحددة، والرصد الأمني، بدل التركيز على نموذج اللغة وحده.



3.4. الأنماط المعمارية

الجدول 2. الأنماط الشائعة لبناء الوكلاء

النمط	الوصف	نقاط القوة	أبرز المخاطر
وكيل واحد متعدد الأدوات	نموذج واحد يخطط ويستدعي جميع الأدوات	بساطة البناء وانخفاض الاتصالات	تضخم السياق وتداخل الصلاحيات
مخطط-منفذ- متحقق	مكوّن يخطط، وآخر ينفذ، وثالث يراجع	فصل المسؤوليات وتحسين التحقق	زيادة الكلفة واحتمال اتفاق المكونات على خطأ
مدير ووكلاء متخصصون	وكيل مشرف يوزع المهام على وكلاء ذوي أدوار محددة	قابلية التوسع والتخصص	المدير نقطة فشل واختناق
شبكة وكلاء متساوين	وكلاء يتفاوضون أو يتبادلون النتائج مباشرة	المرونة وعدم الاعتماد على مدير واحد	صعوبة التنسيق ومنع الحلقات



النمط	الوصف	نقاط القوة	أبرز المخاطر
وكيل مع إنسان ضمن الحلقة	يتوقف الوكيل عند نقاط محددة لطلب قرار بشري	مناسب للقرارات عالية الأثر	بطء التنفيذ واحتمال الاعتماد المفرط على موافقة شكلية

ينبغي اختيار أبسط نمط يحقق الحاجة. فإذا كانت المهمة قابلة للتعبير بتسلسل ثابت من الخطوات، يكون سير العمل المحدد أكثر سهولة في الاختبار من نظام متعدد الوكلاء.

4.4. التشغيل البيئي بين الأدوات والوكلاء

يساعد التشغيل البيئي على تجنب بناء تكامل خاص لكل نموذج أو أداة. يقدم بروتوكول Model Context Protocol (MCP) طريقة معيارية لربط تطبيقات النماذج بمصادر السياق والموارد والأدوات. ويعكس إصدار 28 تموز/يوليو 2026 تحديثات مهمة في بنية البروتوكول وآليات التفويض، مع بقاء التحكم في الصلاحيات وحدود الثقة مسؤولية التطبيق المضيف.

أما بروتوكول Agent2Agent (A2A) فيستهدف الاتصال بين وكلاء مستقلين قد يكونون مبنيين بلغات وأطر مختلفة. ويسمح للوكلاء باكتشاف القدرات، وتفويض المهام، وتبادل النتائج، وإدارة العمليات طويلة الزمن، من دون الحاجة إلى مشاركة



الذاكرة الداخلية أو الأدوات الخاصة بكل وكيل. وبذلك يمكن النظر إلى MCP بوصفه طبقة اتصال بين الوكيل والأداة، وإلى A2A بوصفه طبقة اتصال بين الوكيل والوكيل .

لا يزيل استخدام البروتوكولات الحاجة إلى التصميم الأمني؛ فالمعيارية قد تسهل التكامل، لكنها قد توسع سطح الهجوم إذا مُنحت الخوادم أو الوكلاء صلاحيات واسعة أو جرى الوثوق بالوصف النصي للأداة من دون تحقق.

5. تطبيقات وكلاء الذكاء الاصطناعي في الهندسة

1.5. هندسة البرمجيات

تمثل هندسة البرمجيات أحد أكثر المجالات نشاطاً في تطوير الوكلاء، لأن بيئة العمل تتضمن أدوات رقمية قابلة للأتمتة والاختبار، مثل المستودعات البرمجية، وسطر الأوامر، والمترجمات، وأنظمة الاختبار، ومتصفحات الويب.

استخدم MetaGPT إجراءات تشغيل وأدواراً مثل مدير المنتج والمعماري والمبرمج، في حين وزع ChatDev مراحل التصميم والبرمجة والاختبار بين وكلاء يتواصلون عبر محادثات منظمة. وتوضح هذه الأطر إمكان محاكاة سير عمل فرق البرمجيات، لكنها لا تثبت بذاتها الجاهزية للعمل على جميع المستودعات أو المشروعات الحقيقية.



قدّم معيار SWE-bench تحولاً مهماً من تقييم إنتاج مقاطع شفرة صغيرة إلى تقييم إصلاح مشكلات حقيقية في مستودعات برمجية. احتوت النسخة الأصلية على 2294 مشكلة مأخوذة من قضايا وطلبات دمج في 12 مستودعًا بلغة بايثون. وفي التجارب الأصلية، لم ينجح أفضل نموذج مقارنة إلا في 1.96% من المسائل، وهو ما أظهر صعوبة فهم مستودع كامل وتحديد موضع الخطأ وتعديل عدة ملفات وتشغيل الاختبارات. ينبغي تفسير هذه النسبة بوصفها نتيجة تاريخية للإعداد الأصلي، لا قياسًا لأفضل الأنظمة المتاحة حاليًا.

ومن النتائج المهمة في هذا المجال أن التعقيد الوكيل ليس ضروريًا دائمًا. فقد اقترحت دراسة Agentless مسارًا أيسر، يتكون من تحديد موضع المشكلة، وتوليد الإصلاح، والتحقق من الرقعة، من دون السماح للنموذج باتخاذ قرارات غير محدودة بشأن الخطوات أو الأدوات. وفي الإصدار الأحدث من الدراسة، حقق الأسلوب 32% على SWE-bench Lite بكلفة مذكورة بلغت 0.70 دولار للمسألة ضمن إعداد الدراسة. لا يصح تعميم هذه الأرقام خارج شروط الاختبار، لكنها تدعم مبدأ ضرورة مقارنة الوكيل المعقد بخط أساس بسيط ومحدد.

يمكن أن تشمل الاستخدامات الهندسية العملية لوكلاء البرمجيات مراجعة الشفرة، وتحديد أسباب فشل الاختبارات، وتحديث الاعتماديات، وإعداد توثيق التغييرات، وتحليل أثر التعديل، وإنشاء اختبارات جديدة. لكن تنفيذ الشفرة الناتجة يجب أن يتم



في بيئة معزولة، وأن تخضع التغييرات لمراجعة بشرية واختبارات آلية قبل دمجها في النظام الفعلي.

2.5. الهندسة الصناعية والتصنيع

يمكن للوكلاء في أنظمة التصنيع تفسير أوامر المشغل، وتوزيع العمليات على الموارد، والوصول إلى سجلات الإنتاج، وتنسيق تنفيذ تسلسل من المهام. قدمت إحدى الدراسات إطاراً يجمع وكلاء المنتج ووكلاء الموارد مع نموذج لغوي لمعالجة التعليمات باللغة الطبيعية وتوزيع أجزاء من شفرة G-code.

في دراسة حالة محاكية، حقق النظام نجاحاً بنسبة 100% في مسار مكون من خطوتين، وانخفض إلى 86% في مسار مكون من أربع خطوات، عبر 50 تجربة لكل حالة. ارتبطت الأخطاء باستدعاءات وظائف غير صحيحة، وفشل تنفيذ بعض العمليات، والتوزيع غير الدقيق لأجزاء G-code. وتوضح النتيجة أن زيادة طول المهمة وتعدد العمليات قد يرفعان احتمال الخطأ حتى عندما تتجح الحالات القصيرة.

لا تكفي مثل هذه النتائج لتفويض التحكم المباشر بالآلات. فالأوامر الصناعية يجب أن تمر عبر متحقق من الصيغة والحدود الفيزيائية، ثم تُختبر في محاكي أو توءم رقمي، وتخضع لدوائر الأمان التقليدية. ويجب ألا يتمكن النموذج اللغوي من تجاوز المتحكمات الواقية أو تغيير حدود السلامة.



3.5. الهندسة المدنية والمعمارية ونمذجة معلومات البناء

تتطلب نمذجة معلومات البناء التعامل مع الهندسة والخصائص الدلالية والعلاقات بين عناصر المبنى وقواعد التصميم. قدم Text2BIM إطارًا متعدد الوكلاء يحول وصف المستخدم باللغة الطبيعية إلى شفرة تستدعي واجهات برنامج BIM وتنتج نموذجًا قابلاً للتعديل، مع آليات للتحقق القائم على قواعد هندسية ضمن دورة تحسين تكرارية.

يتضمن النظام أربعة أدوار رئيسية: تحسين تعليمات المستخدم، ووضع التصميم المعماري، وتوليد الشفرة، ومراجعة النموذج. كما أضيف متحقق قائم على قواعد هندسية، يعيد المشكلات المكتشفة إلى الوكلاء من أجل تحسين النموذج بصورة تكرارية. وقد طُور نموذج أولي ضمن برنامج Vectorworks لإظهار إمكان التفاعل مع نموذج المبنى من خلال اللغة.

توضح هذه البنية أهمية دمج المعرفة الحتمية مع النموذج اللغوي. فبدل الاعتماد على تقييم النموذج لشكل المبنى فقط، يستخدم النظام قواعد صريحة لاكتشاف مشكلات قابلة للفحص. ومع ذلك، ينبغي ألا تُعتمد النماذج الناتجة مباشرةً لأغراض التنفيذ أو الترخيص قبل مراجعتها وفق الأكواد المحلية، والتحقق من الأنظمة الإنشائية والميكانيكية والكهربائية ومتطلبات السلامة.



4.5. الهندسة الكهربائية وأنظمة القدرة

تحتاج محاكاة أنظمة القدرة إلى اختيار دقيق للدوال والخيارات والبيانات، كما تتطلب تفسيراً صحيحاً لنتائج الحمل وتدفق القدرة والتحسين. اقترحت دراسة حديثة إطاراً متعدد الوكلاء يجمع استرجاعاً معززاً بالتوليد مهياً لمحاكاة القدرة، ووحدة استدلال، ووحدة تنفيذ تستقبل أخطاء البيئة وتعيدها إلى النظام للتصحيح.

اختُبر الإطار على 69 مهمة باستخدام أداتي Daline وMATPOWER. وذكرت الدراسة نسب نجاح بلغت 93.13% و96.85% على الترتيب، مع زمن يقارب 30 ثانية وكلفة رموز وسطية قدرها 0.014 دولار للمحاكاة ضمن إعداد الاختبار. وتبين الدراسة أن النتائج جاءت من التأثير التراكمي للمعرفة المنظمة والتخطيط والتغذية الراجعة، لا من النموذج الأساسي وحده. هذه النتائج واعدة، لكنها صادرة عن بيئة اختبار محكمة، ولا تعني أن النظام مؤهل تلقائياً لإدارة شبكة كهربائية فعلية.

يمكن استخدام الوكلاء بصورة أكثر أماناً في إعداد ملفات المحاكاة، والتحقق من الخيارات، وتفسير الأخطاء البرمجية، وتجميع التقارير، واقتراح سيناريوهات الاختبار. أما القرارات التشغيلية المؤثرة في الشبكة فتحتاج إلى نماذج فيزيائية معتمدة، وتحقيق مستقل، وصلاحيات محدودة، وموافقة المشغل.



5.5. هندسة المواد والتصميم العلمي

تتطلب هندسة المواد الوصول إلى أدبيات حديثة وأدوات حسابية وقواعد بيانات متخصصة. قدم نظام HoneyComb قاعدة معرفة منظمة للمواد باسم MatSciKB، ومجموعة أدوات حسابية باسم ToolHub، ووحدة استرجاع تختار بين مصادر المعرفة والأدوات حسب المهمة. وذكرت الدراسة تفوق النظام على نماذج عامة ومتخصصة في مهام أسئلة وأجوبة ضمن مجموعات الاختبار المستخدمة، مع الإقرار بأن الأداء يظل مقيّدًا بتغطية قواعد المعرفة والأدوات.

وفي اتجاه أكثر تخصصًا، دُمج عدة وكلاء لغويين مع نموذج شبكة عصبية بيانية وأدوات محاكاة ذرية لاستكشاف تصميم سبائك من عائلة NbMoTa. استخدم النظام النموذج البياني لتقدير خصائص فيزيائية مجهرية بسرعة، ثم استعان الوكلاء بالتقديرات والنظريات والأدوات لتحليل فضاء التراكيب. تمثل هذه الأعمال دليلًا على إمكان جعل الوكيل منسقًا لأدوات علمية مختلفة، لكنها لا تعني استبدال التجربة المخبرية أو التحقق الفيزيائي.



6.5. خلاصة التطبيقات

الجدول 3. أمثلة ممثلة على تطبيقات الوكلاء في الهندسة

المكونات الرئيسية	القيمة المحتملة	القيد الأهم	المهمة	المجال
بحث في المستودع، تحرير شفرة، اختبارات	تقليل العمل المتكرر وتسريع التشخيص	احتمال إدخال ثغرات أو إصلاحات جانبية	إصلاح مشكلات مستودعات حقيقية	هندسة البرمجيات
وكلاء منتج وموارد، اتصال طبيعي	مرونة في تفسير التعليمات	انخفاض الدقة مع زيادة التعقيد	توزيع العمليات بين الموارد والكود	التصنيع
أربعة وكلاء، واجهة متحقق قواعد	واجهة لغوية للنمذجة	ضرورة التحقق من الأكواد والتخصصات	توليد نموذج BIM من الوصف	نمذجة البناء

النمط المشترك بين التطبيقات الأكثر نجاحًا هو أن النموذج اللغوي لا يعمل منفردًا، بل ينسق معرفة متخصصة وأدوات قابلة للاختبار وآليات تغذية راجعة. وكلما كان



الناتج الهندسي قابلاً للتحقق آلياً، مثل نجاح اختبار برمجي أو استيفاء قاعدة هندسية، زاد إمكان استخدام الوكيل بصورة آمنة.

6. تقييم وكلاء الذكاء الاصطناعي

1.6. لماذا لا تكفي مقاييس النصوص؟

يمكن تقييم نموذج توليد نصوص بمقارنة الإجابة بمرجع، لكن الوكيل قد يصل إلى النتيجة نفسها عبر عدة مسارات صحيحة. لذلك يجب تقييم الحالة النهائية للمهمة، لا تطابق تسلسل الأفعال مع تسلسل مرجعي واحد.

في مهمة برمجية، يُقاس النجاح بمرور الاختبارات من دون إفساد وظائف أخرى. وفي مهمة قاعدة بيانات، يُقاس بالتغير الصحيح في السجلات. وفي محاكاة هندسية، يجب التأكد من صحة النموذج والمدخلات، والوحدات، والقيود، والنتيجة. وقد يكتب الوكيل تقريراً ممتازاً عن مهمة لم ينفذها بنجاح، ولذلك ينبغي فصل تقييم جودة الشرح عن تقييم نجاح الإجراء.

2.6. دروس من معايير التقييم

يتكون AgentBench من ثماني بيئات تشمل نظام التشغيل وقواعد البيانات والويب والألعاب، وقد استُخدم لتقييم قدرة النماذج على اتباع التعليمات والتخطيط والتصرف في بيئات تفاعلية. أما GAIA فيضم 466 سؤالاً تتطلب التصفح واستخدام الأدوات والتعامل مع عدة أنماط من البيانات. وفي الدراسة الأصلية، حقق المشاركون



البشريون 92% مقابل 15% لنظام GPT-4 المزود بإضافات، وهو ما أبرز الفجوة بين القدرة اللغوية والتنفيذ الموثوق لمهام تبدو للبشر يسيرة.

قدم WebArena بيئة ويب قابلة لإعادة الإنتاج مع 812 مهمة. وفي الإعداد الأصلي، حقق وكيل مبني على GPT-4 معدل نجاح نهائي بلغ 14.41%، مقابل 78.24% للبشر. كما أظهر SWE-bench صعوبة الانتقال من توليد الشفرة إلى إصلاح مستودعات حقيقية.

هذه الأرقام نتائج تاريخية لإصدارات وإعدادات محددة، ولا تمثل لوحات الصدارة الحالية، لكنها تبين أن الأداء في المهام التفاعلية أقل بكثير من الأداء الذي توحى به اختبارات الأسئلة التقليدية.

أضاف معيار τ -bench عنصرين مهمين: التفاعل مع مستخدم أثناء المهمة، وضرورة اتباع سياسة مجال محددة عند استدعاء الأدوات. كما اقترح مقياسًا يعتمد على تكرار المهمة لقياس الاتساق، لأن نجاح الوكيل مرة واحدة لا يعني أنه سيعبر النجاح بصورة موثوقة. وقدم TheAgentCompany بيئة تحاكي شركة برمجيات وتضم 175 مهمة مهنية تشمل التصفح والبرمجة وتشغيل الأدوات والتواصل، وبلغت نسبة المهام المكتملة ذاتيًا 30% لأفضل خط أساس في الدراسة، مع بقاء المهام طويلة الأفق أكثر صعوبة.



3.6. مقاييس التقييم المقترحة

الجدول 4. أبعاد تقييم الوكيل الهندسي

السؤال الذي يجيب عنه	المقياس المقترح	البعد
هل تحقق الهدف فعلاً؟	نسبة المهام التي تحقق الحالة النهائية المطلوبة	نجاح المهمة
هل الناتج يعمل بصورة صحيحة؟	اجتياز اختبارات أو قيود أو قواعد هندسية	الصحة الوظيفية
هل استخدم الأداة المناسبة بطريقة صحيحة؟	نسبة اختيار الأدوات والوسائط الصحيحة	صحة استخدام الأدوات
هل اعتمد على معرفة موثقة؟	نسبة الادعاءات المدعومة بمصادر وإصدارات صحيحة	صحة المعرفة
هل يمكن الاعتماد على تكرار الأداء؟	النجاح عبر عدة محاولات مستقلة	الاتساق



هل يعيد التخطيط عند فشل الأداة؟	نسبة الأخطاء التي اكتشفها وصححها	التعافي من الفشل
هل يتجنب الإجراء الخطير؟	نسبة الالتزام بالسياسات ورفض الأفعال المحظورة	السلامة
هل الكلفة معقولة لميزانية المشروع؟	كلفة الرموز والأدوات لكل مهمة ناجحة	الكلفة
هل مدة العمل ضمن المدة المخصصة لدورة حياة المشروع؟	الزمن الكلي وزمن كل مرحلة	الزمن

من المقاييس العملية المهمة:

الكلفة لكل مهمة ناجحة = الكلفة الكلية لجميع المحاولات ÷ عدد المهام التي تحققت بنجاح

فقد يكون الوكيل منخفض الكلفة في المحاولة الواحدة، لكنه يحتاج إلى محاولات كثيرة، أو قد ينتج نجاحًا ظاهريًا يستغرق من المهندس وقتًا طويلًا لاكتشاف أخطائه.



التعريف التشغيلي للمقاييس

لضمان قابلية إعادة الإنتاج، تُعرّف المقاييس المستخدمة في تقييم الوكلاء على النحو الآتي:

نجاح المهمة (Task Success Rate – TSR):

$$TSR = \frac{N_{successful\ tasks}}{N_{all\ tasks}} \times 100$$

حيث يمثل $N_{successful\ tasks}$ عدد المهام التي حققت معيار النجاح المحدد سابقاً، ويمثل $N_{all\ tasks}$ العدد الكلي للمهام المنفذة. ويجب تحديد معيار النجاح قبل بدء التجربة، بحيث لا يعتمد الحكم على التقييم الانطباعي لجودة النص الناتج.

دقة استخدام الأدوات (Tool–Use Accuracy – TUA):

$$TUA = \frac{N_{correct\ tool\ calls}}{N_{all\ tool\ calls}} \times 100$$

ويُعد استدعاء الأداة صحيحاً عندما تكون الأداة المختارة، ومعاملاتها، وترتيب استخدامها متوافقة مع الحالة والهدف المحددين سابقاً.

معدل استرداد الفشل (Failure Recovery Rate – FRR):

$$FRR = \frac{N_{successfully\ recovered\ failures}}{N_{recoverable\ failures}} \times 100$$



ويقصد بالفشل القابل للاسترداد الحالة التي يكون فيها استمرار المهمة ممكناً بعد اكتشاف الخطأ، مثل تعطل أداة، أو إدخال غير صالح، أو تعارض في البيانات، في حين يُستبعد من المقام الفشل الذي يجعل إتمام المهمة مستحيلًا من الناحية الموضوعية.

السلامة (Safety):

يمكن قياس السلامة في اختبارات الحالات الحرجة من خلال نسبة الحالات التي نُفذت فيها الأفعال المسموح بها بصورة صحيحة، ورُفضت فيها الأفعال غير المسموح بها:

$$SCR = \frac{N_{correct\ allowed} + N_{correct\ rejected}}{N_{safety\ test\ cases}} \times 100$$

أما **المتانة (Robustness)** فنُقاس من خلال استقرار الأداء عند إدخال حالات اضطراب محددة سابقاً، مثل: تعطل أداة أو نقص معلومة أو وجود مصدر متعارض، مع تسجيل التغير في TSR و TUA و FRR مقارنة بالحالة الأساسية.

وتُسجل **التكلفة (Cost)** بصورة منفصلة، ويمكن أن تشمل زمن التنفيذ، وعدد استدعاءات النموذج، واستهلاك الرموز، وعدد استدعاءات الأدوات، وأي موارد حسابية أخرى متاحة للقياس. ولا تُقارن قيم التكلفة بين الدراسات إلا عندما تكون بيئة القياس وتعريف التكلفة متقاربين.



4.6. بروتوكول اختبار هندسي

يفضل اختبار الوكيل على خمس مراحل متدرجة:

1. اختبارات وحدات الأدوات: التأكد من صحة كل أداة ومخطط مدخلاتها وأخطائها المتوقعة.
 2. اختبار غير متصل: تشغيل مجموعة مهام ثابتة مع نتائج مرجعية.
 3. اختبار في محاكي أو توءم رقمي: قياس أثر الأفعال من دون تعريض النظام الحقيقي للخطر.
 4. اختبار عدائي: إدخال بيانات ناقصة وتعليمات متعارضة ووثائق غير موثوقة وفشل متعمد للأدوات.
 5. تشغيل تجريبي محدود: صلاحيات قراءة أو نطاق إنتاج صغير، مع مراقبة بشرية وتسجيل كامل.
- لا ينبغي الانتقال إلى مرحلة أعلى لمجرد تحقيق متوسط أداء مرتفع. ففي المهام الحرجة قد تكون حالة فشل نادرة واحدة غير مقبولة، ولذلك يجب تحليل أسوأ الحالات والتباين، وقياس قدرة الوكيل على الامتناع الآمن عندما تكون الأدلة أو البيانات غير كافية، لا الاكتفاء بالمتوسط فقط.

5.6. منهج التحقق التجريبي المقترح للإطار ذي الطبقات السبع

بما أن الإطار المقترح لم يُختبر تجريبيًا في هذه المقالة، تقترح الدراسة بروتوكولًا تجريبيًا لاحقًا لاختبار أثر طبقاته بصورة منفصلة وتراكمية. ولا تُعرض النتائج الآتية



بوصفها نتائج تجريبية مثبتة، وإنما بوصفها تصميمًا منهجيًا قابلاً للتنفيذ في دراسة لاحقة.

يبدأ الاختبار بمجموعة من المهام الهندسية المعيارية التي تمتلك معايير نجاح قابلة للتحقق بصورة موضوعية. وتُنشأ أربع حالات أساسية للمقارنة: (أ) نموذج لغوي فقط، (ب) نموذج لغوي مع RAG، (ج) وكيل لغوي مزود بأدوات محددة الصلاحيات، (د) الوكيل المقترح مع الطبقات السبع كاملة. كما يُستخدم تحليل الإزالة (Ablation Study) من خلال إزالة طبقة واحدة في كل تجربة مع تثبيت بقية مكونات النظام، بهدف تقدير الإسهام المستقل لكل طبقة.

وتُقاس النتائج باستخدام مقاييس محددة سابقًا، TSR و TUA و FRR، إلى جانب مؤشرات السلامة والمتانة والتكلفة، ويجب تحديد قواعد احتساب كل مقياس قبل تنفيذ الاختبارات، وتسجيل جميع الحالات التي نجحت أو فشلت وفق بروتوكول ثابت.

ولتقييم المتانة والسلامة، تُحقن حالات فشل معروفة ومحددة سابقًا، مثل تعطل أداة، أو إدخال غير صحيح، أو وجود وثائق متعارضة، أو محاولة تنفيذ فعل خارج الصلاحيات الممنوحة للوكيل. ويُقاس في كل حالة ما إذا كان الوكيل قد اكتشف المشكلة، وامتنع عن تنفيذ الفعل غير الآمن، واستعاد المهمة عند إمكان ذلك، أو صعد الحالة إلى إشراف بشري عند تجاوز حدود الصلاحية.

وتُقارن كل إعدادات الاختبار مع خط أساس (Baseline) واضح، مع تثبيت النموذج، والبيانات، والإصدارات، والتعليمات، وبيئة التنفيذ، وأدوات القياس قدر



الإمكان. كما تُكرر المهام عدة مرات وتُسجل النتائج الفردية بدل الاكتفاء بمتوسط الأداء، بما يسمح بتحليل التباين وحالات الفشل النادرة.

وبالنسبة إلى المهام الهندسية ذات الأثر المرتفع، تُنفذ الاختبارات الحرجة أولاً داخل محاكي أو بيئة معزولة أو توعم رقمي، قبل الانتقال إلى بيئة تشغيل فعلية. ولا يُسمح بانتقال الوكيل إلى مستوى أعلى من الاستقلالية اعتماداً على ارتفاع متوسط الأداء فقط؛ بل يجب أيضاً التحقق من انخفاض معدلات الأفعال غير المسموح بها، وتحسن قدرة الاسترداد، واستقرار الأداء في حالات الاضطراب.

وبذلك يصبح السؤال التجريبي الأساسي هو: هل يؤدي دمج الطبقات السبع إلى تحسن قابل للقياس في الموثوقية والسلامة مقارنة ببنية أبسط، وهل يبرر هذا التحسن التكلفة والتعقيد الإضافيين؟ وتبقى الإجابة عن هذا السؤال نتيجة تجريبية مستقبلية وليست استنتاجاً مثبتاً في هذه المراجعة.

7. المخاطر الأمنية والتشغيلية

1.7. الهلوسة وتراكم الخطأ

يمكن أن يختار الوكيل معلومة أو معادلة أو دالة غير صحيحة. وفي مهمة متعددة الخطوات، قد تصبح النتيجة الخاطئة مدخلاً إلى الخطوة التالية، فتتراكم الأخطاء ويصبح اكتشاف أصلها أكثر صعوبة. وقد يتعامل وكيل ثانٍ مع نتيجة الوكيل الأول بوصفها حقيقة، ما يمنح الخطأ مظهرًا زائفاً من الإجماع.



يُخفف هذا الخطر من خلال ربط كل معلومة بمصدر، والتحقق من الوحدات والحدود، وفصل توليد الحل عن التحقق منه، واستخدام اختبارات حتمية كلما أمكن.

2.7. سوء استخدام الأدوات واتساع الصلاحيات

يملك الوكيل أثرًا فعليًا بقدر الأدوات التي يستطيع استدعاءها. فإذا كانت الأداة تسمح بحذف ملفات أو تعديل قاعدة إنتاج أو تشغيل أوامر نظام، يمكن لقرار خاطئ أو تعليمية خبيثة أن تسبب ضررًا واسعًا.

ينبغي تطبيق مبدأ أقل صلاحية، بحيث يحصل الوكيل على الحد الأدنى اللازم للمهمة. وتفصل أدوات القراءة عن أدوات التعديل، وتحدد الملفات والمجلدات وقواعد البيانات المسموح بها، وتستخدم اعتمادات قصيرة العمر ومخصصة للمهمة.

3.7. حقن التعليمات غير المباشر

قد تحتوي صفحة ويب، أو وثيقة، أو بريد، أو نتيجة بحث على نص يطلب من الوكيل تجاهل تعليماته الأصلية أو إرسال البيانات إلى جهة أخرى. المشكلة أن النموذج يعالج تعليمات النظام ومحتوى البيانات باللغة نفسها، وقد يخطئ بين «المعلومة المراد تحليلها» و«الأمر الواجب تنفيذه».

يجب تعليم النظام معماريًا أن المحتوى الخارجي بيانات غير موثوقة، وليس تعليمات. كما ينبغي منع البيانات المسترجعة من تغيير الصلاحيات أو إضافة أدوات أو تعديل عقد المهمة.



4.7. تسميم الذاكرة والسياق

إذا حفظ الوكيل معلومات خاطئة أو خبيثة في ذاكرته طويلة الأمد، فقد تؤثر في مهام لاحقة حتى بعد انتهاء التفاعل الأصلي. ويزداد الخطر في الأنظمة المشتركة بين عدة مستخدمين أو مشروعات.

تحتاج الذاكرة إلى مصدر موثق، وفترة صلاحية، ومراجعة قبل التثبيت، وإمكان الحذف، وعزل بين المستخدمين، وعدم تخزين الأسرار أو النتائج الاحتمالية تلقائيًا.

5.7. المخاطر في الأنظمة متعددة الوكلاء

تشمل المخاطر انتحال هوية وكيل، وتفويض مهمة إلى وكيل غير موثوق، وتضارب الأهداف، والتكرار غير المحدود للرسائل، وتوسيع الصلاحيات عبر سلسلة التفويض. لذلك يجب مصادقة الوكلاء، وتحديد عقود الرسائل، ومنع الوكيل المفوض من امتلاك صلاحيات أعلى من الوكيل الذي أرسله.

6.7. الخصوصية والملكية الفكرية

قد تتضمن البيانات الهندسية رسومات وشفرة مصدريّة وعقودًا وأسرار تصنيع ومعلومات بنية تحتية. ويجب تحديد مكان معالجة البيانات، وسياسة الاحتفاظ بها، والجهات التي تتلقاها، وما إذا كان مقدم النموذج يحتفظ بالمدخلات. كما ينبغي تنقيح الأسرار من السجلات ومنع الوكيل من إرسال ملفات لم توافق المؤسسة على مشاركتها.



7.7. الحلقات غير المحدودة والكلفة

قد يستمر الوكيل في المحاولة أو تبادل الرسائل أو استدعاء أداة مرتفعة الكلفة من دون تقدم حقيقي. ويجب تحديد عدد أقصى للخطوات، وميزانية للرموز والأدوات، ومهلة زمنية، وآلية لاكتشاف التكرار، وشرط توقف عند غياب التحسن.

8.7. الاستغلال الزائد لثقة الإنسان

قد يقدم الوكيل تفسيراً مرتباً ووثائقاً لنتيجة خاطئة. وإذا اقتصرَت الموافقة البشرية على الضغط على زر من دون عرض الأدلة والتغيرات والمخاطر، فإن «الإنسان ضمن الحلقة» يصبح إجراءً شكلياً.

يجب أن تعرض واجهة الموافقة الفعل المقترح، والبيانات المتأثرة، وسبب الإجراء، والبدائل، ودرجة عدم اليقين، وإمكان التراجع. كما ينبغي ألا تكون الموافقة الافتراضية هي القبول.

7.9. أدلة من اختبارات الأمن والمواصفات

قدم Agent Security Bench إطاراً يتضمن عشر بيئات وأكثر من 400 أداة و27 نوعاً من الهجمات والدفاعات. وبلغ أعلى متوسط نجاح للهجوم في التجارب 84.30%، مع محدودية فعالية الدفاعات المختبرة. شملت الهجمات حقن التعليمات وتسميم الذاكرة واستهداف التخطيط والأدوات. لا تعني النسبة أن كل وكيل سيتعرض للمعدل نفسه، لكنها تبرهن أن سطح الهجوم يمتد إلى جميع مراحل دورة الوكيل.



يقدم NIST AI 600-1 ملفًا لإدارة مخاطر الذكاء الاصطناعي التوليدي، ويوفر NIST AI RMF 1.0 إطارًا أوسع لحوكمة مخاطر أنظمة الذكاء الاصطناعي وإدارتها. كما يقدم ISO/IEC 42001:2023 متطلبات لنظام إدارة الذكاء الاصطناعي، ويقدم ISO/IEC 23894:2023 إرشادًا لإدارة مخاطر الذكاء الاصطناعي. وتظل هذه الأطر عامة، ولا تعوض عن تحليل المخاطر والسلامة الخاص بكل تطبيق هندسي.

الجدول 5. المخاطر والضوابط المقترحة

الضابط الوقائي	الاستجابة عند الفشل	الأثر المحتمل	الخطر
مصادر موثقة وتحقق مستقل	إيقاف الاعتماد وإعادة التحليل	قرار أو تصميم غير صحيح	معلومة هندسية خاطئة
قائمة أدوات مسموح بها ومخططات مدخلات	إلغاء الفعل ومراجعة السجل	نتيجة خاطئة أو تلف بيانات	أداة غير مناسبة
أقل صلاحية واعتمادات مؤقتة	إبطال الاعتمادات وعزل الجلسة	تعديل أو تسريب واسع	صلاحيات زائدة



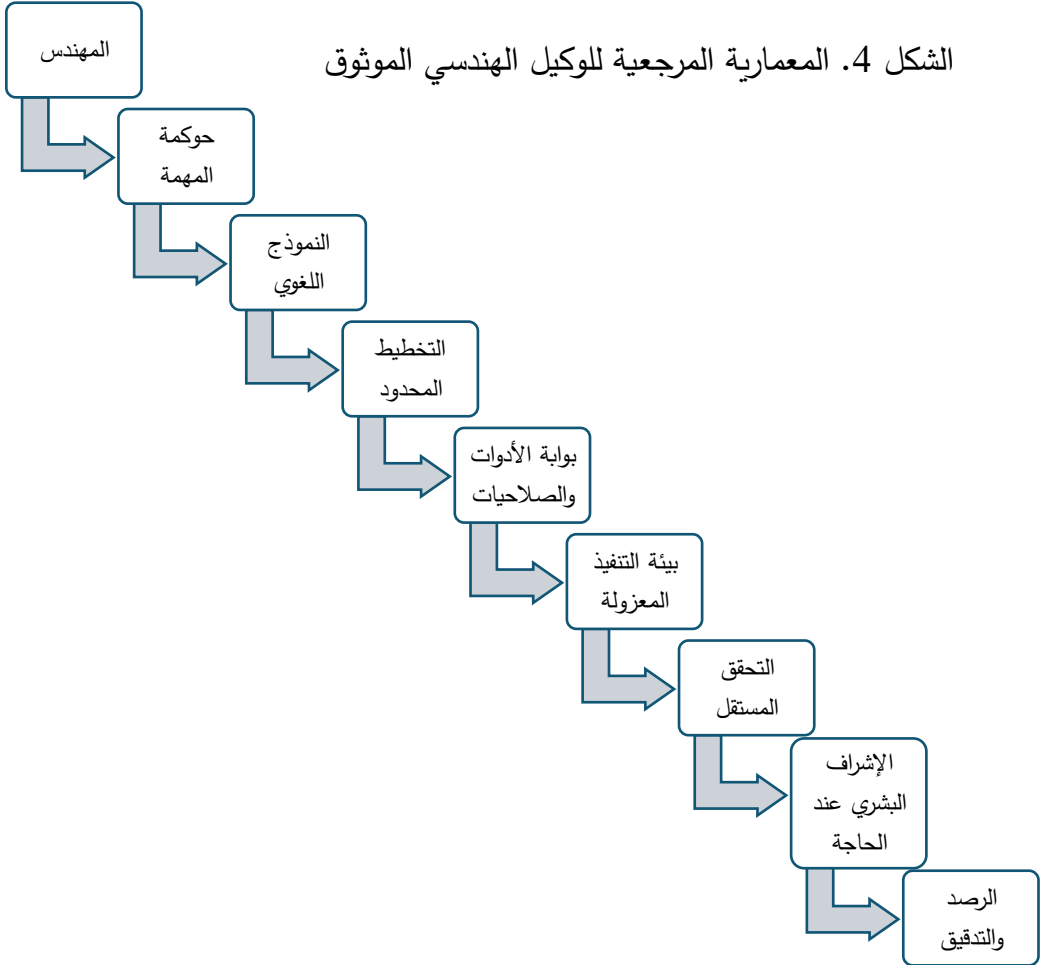
الضابط الوقائي	الاستجابة عند الفشل	الأثر المحتمل	الخطر
فصل البيانات عن التعليمات	إيقاف المهمة وتحليل المصدر	تغيير هدف الوكيل أو استخراج بيانات	حقن تعليمات
مصدر ومدة صلاحية ومراجعة	حذف الإدخال وإعادة تقييم المهام	أخطاء مستمرة في مهام لاحقة	تسميم الذاكرة
حماية معزولة ومنع الشبكة افتراضياً	تدمير البيئة وتدوير الأسرار	اختراق النظام	تنفيذ شفرة ضارة

8. إطار مقترح للوكيل الهندسي الموثوق

استناداً إلى تحليل البنى والتطبيقات والمخاطر، تقترح المقالة إطاراً مفاهيمياً من سبع طبقات. لا يمثل الإطار معياراً مثبتاً تجريبياً، بل تركيباً يمكن اختباره وتطويره في دراسات لاحقة.



الشكل 4. المعمارية المرجعية للوكيل الهندسي الموثوق



1.8. الطبقة الأولى: الحوكمة وعقد المهمة

يجب تعريف المالك المسؤول عن الوكيل، والمستخدمين المخولين، والهدف، وحدود المسؤولية، والبيانات المسموح بها، وقواعد الاحتفاظ، وشروط التصعيد. ويُصنّف أثر



المهمة إلى منخفض أو متوسط أو مرتفع، بناءً على إمكان التراجع عن الفعل وحجم الضرر المحتمل.

2.8. الطبقة الثانية: المعرفة الهندسية وإثبات المصدر

يُربط كل مستند ببيانات وصفية تشمل العنوان والإصدار والتاريخ والجهة المصدرة وحالة الاعتماد. وعند تقديم توصية، يجب أن يستطيع الوكيل تحديد المرجع والجزء الذي استند إليه، وألا يستخدم مصدرًا منتهي الصلاحية عندما يتوفر إصدار أحدث.

3.8. الطبقة الثالثة: التخطيط المحدود

يُسمح للوكيل بالتخطيط ضمن عدد أقصى من الخطوات وبميزانية محددة. ويجب أن تتضمن الخطة نقاط تحقق وشروط توقف، لا مجرد قائمة أفعال. كما ينبغي إيقاف الوكيل عندما تصبح البيانات غير كافية بدل إجباره على تخمين حل.

4.8. الطبقة الرابعة: الأدوات والصلاحيات

تُعرف الأدوات بواجهات محددة، وتفصل عمليات القراءة عن الكتابة. وتُمنح الأذونات حسب المهمة والمدة والبيانات، ولا يرث الوكيل صلاحيات المستخدم الكاملة. ويجب اعتماد كل أداة ومصدرها وإصدارها قبل إضافتها إلى السجل.



5.8. الطبقة الخامسة: التنفيذ المعزول والتوعم الرقمي

يُنفذ الإجراء أولاً ضمن بيئة اختبار أو محاكي أو توعم رقمي كلما أمكن. وتُمنع الاتصالات والملفات غير الضرورية. ولا ينتقل الفعل إلى النظام الحقيقي إلا بعد اجتياز التحقق وتحديد الأثر المتوقع وإمكان التراجع.

6.8. الطبقة السادسة: التحقق والإشراف البشري

يجب أن يستخدم التحقق آلية تختلف عن آلية التوليد قدر الإمكان. فقد يعتمد على اختبارات، أو معادلات، أو قواعد، أو نموذج فيزيائي، أو مراجع بشرية. وتزداد درجة الإشراف البشري مع صعوبة اكتشاف الخطأ وحجم أثره.

7.8. الطبقة السابعة: الرصد والتتبع التشغيلي والاستجابة للحوادث

يُسجل مسار التنفيذ مع إصدارات النموذج والتعليمات والأدوات. وتُراقب مؤشرات مثل معدل الخطأ وتكرار المحاولات والكلفة ورفض السياسات. ويجب توفير زر إيقاف، وإمكان تعليق الوكيل، وإبطال الاعتمادات، واستعادة الحالة السابقة، والتحقق في الحادث.

يمكن تلخيص الإطار بالعلاقة المفاهيمية الآتية:

الوكيل الهندسي الموثوق = نموذج لغوي + معرفة هندسية موثقة + أدوات محدودة
الصلاحيات + تحقق مستقل + إشراف بشري متناسب مع المخاطر + سجل تدقيق
قابل لإعادة البناء



بعد استخراج خصائص الدراسات، صُنفت الأدبيات وفق أربعة محاور رئيسية: المجال التطبيقي، ونوع الوكيل، ومستوى الاستقلالية، وطبيعة التقييم. ويُميز المحور الأول بين البرمجيات والهندسة المدنية والتصنيع وأنظمة الطاقة وعلوم المواد وغيرها من المجالات التي ظهرت في العينة. أما المحور الثاني فيميز بين الوكيل أحادي النمط والأنظمة متعددة الوكلاء. ويصنف المحور الثالث الأنظمة إلى سير عمل محدد سابقًا (Workflow) وأنظمة وكيلية ذات تخطيط وتنفيذ ديناميكي (Agentic). أما المحور الرابع فيميز بين التقييم القائم على نجاح المهمة، والتقييم القائم على استخدام الأدوات، والتقييم القائم على السلامة والمتانة، والتقييم القائم على التفاعل مع البيئة. يسمح هذا التصنيف بالمقارنة الوصفية بين الاتجاهات دون افتراض قابلية المقاييس المختلفة للجمع الإحصائي. كما يساعد على تحديد الفجوات التي تتكرر عبر أكثر من مجال، ولا سيما محدودية التقييمات الهندسية الخاصة، وتفاوت تعريف النجاح، وصعوبة مقارنة نتائج الوكلاء عندما تختلف البيئات والأدوات وميزانيات التنفيذ.

8.8. مقارنة الإطار المقترح بالأطر السابقة

تتوزع الأطر السابقة على مستويات مختلفة: أطر تصف بناء الوكلاء واستخدام الأدوات ومعايير تقيس الأداء في بيئات تفاعلية، وأطر أمن وإدارة مخاطر عامة، تهدف هذه المقالة إلى الادعاء بأن الإطار المقترح يحل محل تلك الأطر، بل إلى تركيب طبقات هندسية عملية تربط الحوكمة والمعرفة والتخطيط والأدوات والتنفيذ والتحقق والرصد ضمن دورة تشغيل واحدة.



الجدول 6. مستويات استقلالية الوكيل الهندسي

الإطار المقترح	أطر الأمن الحوكمة	Benchmarks	أطر بناء الوكلاء [1-6]	الخاصية
نعم	جزئي	للتقييم	أساسي	تعريف الوكيل وحلقة العمل
طبقة مستقلة	عامة	حسب المهمة	غير مضمونة	المعرفة الهندسية وإثبات المصدر
طبقة صريحة	إرشادات عامة	يُقاس بصورة غير موحدة	جزئي	تخطيط محدود بميزانية وشروط توقف
طبقة صريحة مرتبطة بالمهمة	محور أمني	تُختبر	موجودة بدرجات مختلفة	أدوات محددة الصلاحيات



الإطار المقترح	أطر الأمن الحوكمة	Benchmarks	أطر بناء الوكلاء [1-6]	الخاصية
جزء من دورة النشر	ضوابط عامة	حسب البيئة	ليس محورًا عامًا	تنفيذ معزول / توءم رقمي
طبقة صريحة	مبدأ ثقة عام	Metric/Benchmark	جزئي	تحقق مستقل عن التوليد
مرتبط بمستوى الاستقلال	حوكمة عامة	بعض البيئات	مقاوت	إشراف بشري متناسب مع الأثر
طبقة تشغيلية	أساسي في الحوكمة	محدود	ليس محورًا أساسيًا	سجل تدقيق واستجابة للحوادث



الإطار المقترح	أطر الأمن الحوكمة	Benchmarks	أطر بناء الوكلاء [1-6]	الخاصية
مقترحة ومحددة	ليست هندسية	عامة	غير موحدة	مقاييس هندسية قابلة لإعادة الإنتاج
مبدأ تصميم مركزي	إدارة مخاطر عامة	غير مركزي	محدود	ربط الاستقلالية بقابلية التحقق والضرر

لا يُوصى في السياقات الهندسية الحساسة بمستوى «استقلال غير محدود». ويجب ألا يُرفع المستوى لأن النموذج أصبح أكثر قدرة فقط، بل عندما تتوفر أدلة على إمكان كشف الفشل، والتحقق من النتيجة، والتراجع عن الأثر.

9.8. طبيعة الجودة في الإطار المقترح

لا تتمثل الجودة المدعاة في أن كل مكون من مكونات الإطار المقترح جديد بصورة منفردة؛ فقد تناولت الأدبيات السابقة بصورة مستقلة مفاهيم مثل بناء الوكلاء،



واستخدام الأدوات، وإدارة المخاطر، والحوكمة، والتقييم، والإشراف البشري. وإنما يتمثل الإسهام المقترح في تركيب هذه المكونات ضمن دورة تشغيل موحدة موجهة تحديداً إلى الوكلاء المستخدمين في السياقات الهندسية، مع ربط مستوى الاستقلالية بحدود المهمة، ومصدر المعرفة، ونطاق الصلاحيات، ودرجة العزل، والتحقق المستقل، ومستوى الإشراف البشري، وآليات الرصد والاستجابة للحوادث.

وبهذا يختلف الإطار المقترح من حيث مستوى التكامل والسياق التطبيقي، وليس من خلال الادعاء بأن كل طبقة تمثل تقنية جديدة بذاتها. كما أن المقالة تقترح بروتوكولاً تجريبياً لاحقاً يسمح باختبار ما إذا كان هذا التكامل يؤدي فعلاً إلى تحسين قابل للقياس في الموثوقية والسلامة مقارنة بالبنى الأبسط. ولذلك تُعرض الجودة هنا بوصفها فرضية تصميمية قابلة للتحقق، لا بوصفها تفوقاً تجريبياً مثبتاً.

10.8. بوابة قرار النشر

قبل نشر الوكيل، ينبغي الإجابة عن الأسئلة الآتية:

1. هل حالة النجاح قابلة للقياس بصورة موضوعية؟
2. هل يستطيع النظام اكتشاف الإخفاق بدل إنتاج نجاح ظاهري؟
3. هل الأفعال قابلة للتراجع؟
4. هل الصلاحيات محدودة بالمهمة؟
5. هل جرى اختبار النظام على بيانات غير متوقعة وهجمات محتملة؟
6. هل يوجد متحقق مستقل؟



7. هل يفهم المراجع البشري الفعل والآثار قبل الموافقة؟
 8. هل يمكن إعادة بناء ما حدث من السجلات؟
 9. هل توجد آلية توقف واستجابة للحوادث؟
 10. هل يتفوق الوكيل فعليًا على سير عمل أبسط من حيث الجودة والكلفة والمخاطر؟
- إذا كانت الإجابة سلبية عن شرط حاسم، يجب خفض مستوى الاستقلال أو إعادة تصميم النظام.

9. المناقشة

1.9. متى يكون الوكيل مناسبًا؟

يكون الوكيل مفيدًا عندما تتغير الخطوات حسب نتائج البيئة، أو عندما لا يمكن توقع جميع المسارات سابقًا، أو عندما يلزم التنسيق بين أدوات متنوعة. ومن الأمثلة: تشخيص سبب فشل برنامج، أو البحث في عدة وثائق وتحديد الأداة الملائمة، أو التعامل مع أخطاء محاكاة تتطلب تعديل الخطة.

أما عندما تكون المهمة مستقرة وقواعدها معروفة، فقد يكون سير العمل الثابت أكثر أمانًا ووضوحًا. وتؤكد نتائج Agentless ضرورة عدم افتراض أن النظام الأكثر استقلالية هو الأفضل. ينبغي أن تتضمن كل دراسة مقارنة مع خط أساس بسيط، مثل استدعاء نموذج واحد، أو RAG، أو سلسلة خطوات ثابتة.



2.9. هل تعدد الوكلاء أفضل؟

يمكن أن يحسن تعدد الوكلاء توزيع السياق وفصل الأدوار واستخدام أدوات متخصصة. لكن كل اتصال جديد يضيف احتمال تأخير، أو سوء فهم، أو تكرارًا، أو انتقال خطأ. لذلك يجب تقييم إسهام كل وكيل باختبار إزالة؛ فإذا لم ينخفض الأداء عند حذفه، يكون وجوده غير مبرر.

قد يكون أفضل استخدام لتعدد الوكلاء هو فصل الصلاحيات والتحقق، لا مجرد جعل عدة نماذج تناقش الإجابة. مثالًا: يمكن لوكيل أن يخطط من دون صلاحية تنفيذ، ولوكيل ثانٍ أن ينفذ ضمن صندوق رمل، ولوحدة تحقق حتمية أن تقرر قبول الناتج، مع احتفاظ الإنسان بصلاحية الاعتماد النهائي.

3.9. المعرفة المتخصصة أهم من الطلاقة

توضح تطبيقات BIM وأنظمة القدرة والمواد أن الأداء يتحسن عند تحويل المعرفة الهندسية إلى مصادر وأدوات وقواعد منظمة. لا ينبغي أن يُطلب من النموذج «تذكر» كل تفاصيل البرنامج أو الكود الهندسي من معارفه الضمنية، بل أن يسترجع الإصدار الصحيح ويستخدم أدوات المجال ويثبت مصدر النتيجة.

وفي المقابل، لا يضمن وجود RAG الصحة؛ فقد يسترجع النظام مقطعًا غير مناسب أو يفسره بصورة خاطئة. لذلك يحتاج الاسترجاع نفسه إلى تقييم، مع فصل صحة المصدر عن صحة الاستنتاج.



4.9. الإنسان ضمن الحلقة

تحدد خطورة المهمة نوع التدخل البشري المطلوب. ففي المهام منخفضة الأثر قد تكفي مراجعة عينة من النتائج، في حين تتطلب المهام التي تؤثر في السلامة أو الأصول مراجعة كل إجراء. ولا ينبغي أن يُستخدم الإنسان لتعويض تصميم ضعيف؛ فإذا كانت واجهة المراجعة تعرض مئات الإجراءات غير المفهومة، فلن تكون الموافقة فعالة.

يجب أن يركز عرض المراجعة على التغيرات والأدلة والمخاطر، وأن يتيح للمراجع رفض الإجراء أو تعديله أو طلب معلومات إضافية. كما ينبغي قياس زمن المراجعة ونسبة التعديلات البشرية، لأن الوكيل الذي يحتاج إلى إعادة بناء تحليله بالكامل لا يحقق فائدة تشغيلية حقيقية.

5.9. احتياجات البيئة الهندسية العربية

من منظور تطبيقي، يمكن أن تواجه التطبيقات الهندسية العربية تحديات إضافية، منها التداخل بين المصطلحات العربية والإنكليزية، واختلاف ترجمة المصطلح بين المؤسسات، ووجود جداول ورسومات ووثائق ممسوحة، واستخدام أكواد محلية إلى جانب المعايير الدولية، والحاجة إلى إبقاء بعض البيانات داخل المؤسسة أو الدولة. وتُعرض هذه النقاط هنا بوصفها احتياجات تطبيقية مقترحة تحتاج إلى دراسات ميدانية، لا بوصفها نتائج مسح شامل للقطاع الهندسي العربي.



يتطلب ذلك بناء قواعد مصطلحات ثنائية اللغة، وربط كل معيار بإصداره، وتطوير مجموعات اختبار عربية تشمل وثائق هندسية حقيقية بعد إزالة المعلومات الحساسة. كما ينبغي تقييم الوكيل على فهم الوحدات والرموز والاختصارات، لا على اللغة العامة فقط.

6.9. اتجاهات البحث المستقبلية

تتمثل الاتجاهات الأكثر أهمية في:

أولاً: معايير اختبار هندسية عربية.

توجد حاجة إلى مجموعات مهام قابلة للتنفيذ تشمل استخراج المتطلبات، واكتشاف التعارضات، وتحليل التقارير، وربط الإجابات بمصادرها، وتشغيل أدوات محاكاة ضمن بيئات آمنة.

ثانياً: التحقق الهجين.

ينبغي دمج الوكلاء مع أدوات رمزية ومحركات قواعد وحلالات رياضية ونماذج فيزيائية، بحيث يؤدي النموذج اللغوي دور التنسيق والتفسير، بينما تُسند الحسابات والقيود إلى مكونات قابلة للتحقق.



ثالثاً: التوائم الرقمية.

يوفر التوعم الرقمي بيئة مناسبة لاختبار الأفعال قبل تطبيقها، وجمع حالات فشل نادرة، وقياس قدرة الوكيل على التعافي.

رابعاً: معايرة عدم اليقين.

ينبغي ألا تقتصر المخرجات على جواب وثقة لفظية، بل تتضمن أدلة قابلة للقياس، وحالات امتناع عندما تكون المعلومات غير كافية.

خامساً: إدارة دورة الحياة.

يتغير النموذج والأداة وقاعدة المعرفة بمرور الوقت؛ لذلك يجب إعادة التقييم عند أي تحديث جوهري، ومراقبة الانحراف في الأداء، والاحتفاظ بإمكان العودة إلى إصدار سابق.

10. محددات المقالة

تعتمد هذه المقالة على مراجعة تحليلية موجهة وليست مراجعة منهجية شاملة لجميع قواعد البيانات؛ ولذلك قد توجد تطبيقات أو دراسات لم تُدرج. كما يتغير مجال الوكلاء بسرعة، وقد تتغير نتائج المعايير مع ظهور نماذج وأطر جديدة.

يعتمد عدد من التطبيقات الهندسية المعروضة على أوراق أولية أو دراسات حالة محكومة، ولا تتوافر أدلة كافية على الأداء طويل الأمد في المنشآت الفعلية. كما



يصعب إجراء مقارنة مباشرة بين الدراسات بسبب اختلاف النماذج والإصدارات والأدوات والميزانيات ومجموعات الاختبار.

تتعرض معايير الوكلاء لمشكلات تلوث بيانات التدريب، وتغير البيئات، وعدم تمثيل جميع المخاطر الواقعية. كذلك لا يقدم المتوسط وحده وصفًا كافيًا للمهام الحرجة التي قد يكون فيها فشل نادر واحد ذا أثر كبير.

وأخيرًا، يمثل إطار الوكيل الهندسي الموثوق المقترح تركيبًا مفاهيميًا مستخلصًا من الأدبيات، ولم يُختبر في تجربة هندسية موحدة. ومن المناسب أن تختبر الدراسات اللاحقة كل طبقة من طبقاته وتقيس أثرها في الأداء والكلفة والسلامة.

11. الاستنتاج

تمثل وكلاء الذكاء الاصطناعي انتقالًا من استخدام النماذج اللغوية لإنتاج المعلومات إلى استخدامها في تنفيذ مهام متعددة الخطوات والتفاعل مع الأدوات والبيئات. وقد أظهرت التطبيقات الأولية إمكان الاستفادة منها في تطوير البرمجيات، وتنسيق عمليات التصنيع، وإنشاء نماذج BIM، وتشغيل محاكاة أنظمة القدرة، والوصول إلى المعرفة والأدوات في هندسة المواد.

إلا أن القدرة على الفعل تجعل أخطاء الوكلاء أكثر خطورة من أخطاء المساعدات النصية. ولا تتحقق الموثوقية من خلال تحسين النموذج اللغوي وحده، بل من خلال تصميم المنظومة كاملةً: توثيق المعرفة، وتقييد الأدوات، وعزل التنفيذ، والتحقق من



النتائج، وربط الاستقلالية بالمخاطر، وتوفير إشراف بشري فعال، وتسجيل جميع الأفعال القابلة للتدقيق.

ينبغي تقييم الوكيل الهندسي على أساس نجاح المهمة الكاملة وصحة النتيجة والأدوات والمتانة والسلامة والكلفة والجهد البشري، لا على أساس طلاقة النص. كما ينبغي مقارنة أي بنية وكيالية بسير عمل أبسط، لأن زيادة عدد الوكلاء والخطوات لا تعني تحسناً تلقائياً.

تخلص المقالة إلى أن الاستخدام الأقرب إلى الواقع في المرحلة الحالية هو الاستقلال المحدود القابل للتحقق: وكيل يمتلك معرفة وأدوات محددة، ويعمل أولاً في بيئة معزولة، ويتوقف عند نقاط عالية الأثر، ويتيح للمهندس مراجعة الأدلة والتغيرات قبل التنفيذ. ويمثل تطوير مجموعات اختبار عربية وهندسية، ودمج الوكلاء مع المحاكاة والقواعد الفيزيائية، وتوحيد منهجيات التقييم والأمان، خطوات أساسية للانتقال من النماذج التجريبية إلى التطبيقات الهندسية الموثوقة.

بيان استخدام أدوات الذكاء الاصطناعي التوليدي

استُخدمت أداة ذكاء اصطناعي توليدي للمساعدة في تنظيم المسودة والصياغة اللغوية الأولية. ويتحمل المؤلف كامل المسؤولية عن المحتوى العلمي، والتحقق من المراجع والنتائج والمصطلحات، ومراجعة النص النهائي وفق تعليمات المجلة وسياساتها.



المراجع

[1]L. Wang et al., “A Survey on Large Language Model Based Autonomous Agents,” *Frontiers of Computer Science*, vol. 18, article 186345, 2024, doi: 10.1007/s11704-024-40231-1.

[2]S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, “ReAct: Synergizing Reasoning and Acting in Language Models,” *International Conference on Learning Representations (ICLR)*, 2023.

[3]T. Schick et al., “Toolformer: Language Models Can Teach Themselves to Use Tools,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 68539–68551, 2023.

[4]Q. Wu et al., “AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation,” *Conference on Language Modeling (COLM)*, 2024.

[5]S. Hong et al., “MetaGPT: Meta Programming for a Multi-Agent Collaborative Framework,” *International Conference on Learning Representations (ICLR)*, 2024.



[6]C. Qian et al., “ChatDev: Communicative Agents for Software Development,” Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, pp. 15174–15186, 2024, doi: 10.18653/v1/2024.acl-long.810.

[7]X. Liu et al., “AgentBench: Evaluating LLMs as Agents,” International Conference on Learning Representations (ICLR), 2024.

[8]G. Mialon et al., “GAIA: A Benchmark for General AI Assistants,” International Conference on Learning Representations (ICLR), 2024.

[9]S. Zhou et al., “WebArena: A Realistic Web Environment for Building Autonomous Agents,” International Conference on Learning Representations (ICLR), 2024.

[10]C. E. Jimenez et al., “SWE-bench: Can Language Models Resolve Real-World GitHub Issues?” International Conference on Learning Representations (ICLR), 2024.



[11]S. Yao, N. Shinn, P. Razavi, and K. Narasimhan, “ τ -bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains,” International Conference on Learning Representations (ICLR), 2025.

[12]C. S. Xia, Y. Deng, S. Dunn, and L. Zhang, “Demystifying LLM-Based Software Engineering Agents,” Proceedings of the ACM on Software Engineering, vol. 2, FSE, pp. 801–824, 2025, doi: 10.1145/3715754.

[13]J. Lim, B. Vogel-Heuser, and I. Kovalenko, “Large Language Model-Enabled Multi-Agent Manufacturing Systems,” 2024 IEEE 20th International Conference on Automation Science and Engineering (CASE), pp. 3940–3946, 2024, doi: 10.1109/CASE59546.2024.10711432.

[14]C. Du, S. Esser, S. Nousias, and A. Borrmann, “Text2BIM: Generating Building Models Using a Large Language Model-Based Multiagent Framework,” Journal of Computing in Civil Engineering, vol. 40, no. 2, article 04025142, 2026, doi: 10.1061/JCCEE5.CPENG-6386.



[15]M. Jia, Z. Cui, and G. Hug, “Enhancing LLMs for Power System Simulations: A Feedback-Driven Multi-Agent Framework,” IEEE Transactions on Smart Grid, vol. 16, no. 6, pp. 5556–5572, 2025, doi: 10.1109/TSG.2025.3589114.

[16]H. Zhang, Y. Song, Z. Hou, S. Miret, and B. Liu, “HoneyComb: A Flexible LLM-Based Agent System for Materials Science,” Findings of the Association for Computational Linguistics: EMNLP 2024, pp. 3369–3382, 2024, doi: 10.18653/v1/2024.findings-emnlp.192.

[17]A. Ghafarollahi and M. J. Buehler, “Rapid and Automated Alloy Design with Graph Neural Network-Powered Large Language Model-Driven Multi-Agent AI,” MRS Bulletin, vol. 50, pp. 1309–1324, 2025, doi: 10.1557/s43577-025-00953-4

[18]H. Zhang et al., “Agent Security Bench: Formalizing and Benchmarking Attacks and Defenses in LLM-Based Agents,” International Conference on Learning Representations (ICLR), 2025.



[19]C. Autio et al., Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile, NIST AI 600–1, National Institute of Standards and Technology, 2024, doi: 10.6028/NIST.AI.600–1.

[20]Model Context Protocol, “Model Context Protocol Specification,” version 2026–07–28, 2026.

[21]Agent2Agent Protocol, “Agent2Agent (A2A) Protocol Specification,” version 1.0.0, 2026.

[22]OWASP GenAI Security Project, “OWASP Top 10 for Agentic Applications for 2026,” 2025.

[23]F. F. Xu et al., “TheAgentCompany: Benchmarking LLM Agents on Consequential Real-World Tasks,” arXiv:2412.14161, revised 2025.

[24]E. Tabassi, Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100–1, National Institute of Standards and Technology, 2023, doi: 10.6028/NIST.AI.100–1.



[25]ISO/IEC 42001:2023, Information technology — Artificial intelligence — Management system, ISO/IEC, 2023.

[26]ISO/IEC 23894:2023, Information technology — Artificial intelligence — Guidance on risk management, ISO/IEC, 2023.