45 CANONICAL RESEARCH ROOTS · SCIENTIFIC COMMUNITY + FRONTIER LABS

الأطلس العلمي والهندسي المكثف

Frontier AGI · Advanced Reasoning · Agentic AI · Computational Consciousness

نظرة هندسية فقط على 30 ورقة مرجعية + 15 ورقة Frontier إضافية

FORMAL INTELLIGENCE · FOUNDATION MODELS · REASONING · TOOLS · MEMORY · AGENTS · REAL-WORLD EVALS · WORLD MODELS · EMBODIED AI · AUTONOMOUS SCIENCE · ALIGNMENT · COMPUTATIONAL CONSCIOUSNESS

قاعدة الانضباط
لا تُساوى benchmark gains بـAGI، ولا architecture motif بـconsciousness، ولا tool access بـagency موثوقة. كل صفحة تفصل Evidence عن Engineering Inference.

الهدف
ضغط 45 جذراً بحثياً إلى خريطة واحدة تشرح: ما الذي أثبته كل عمل؟ ما الآلية؟ ما الذي يمكن استخراجه هندسياً؟ وما حدوده؟

التنسيق
50 pages · A4 · English Technical Terms · Amiri · Arabic RTL · high-density scientific reference.

القطع الزمني
Research cutoff: 31 August 2026. الأوراق 31-45 هي الإضافة الجديدة لتغطية أحدث اتجاهات frontier laboratories والبحث الأكاديمي.

هذه موسوعة خارجية للمجتمع العلمي؛ لا تفترض صحة أي إطار داخلي للمستخدم، ولا تدّعي أن الـ45 ورقة تغلق البحث مستقبلاً. المقصود بناء minimum-spanning research map يمكن تحديثه بالـdelta.

خريطة corpus: من تعريف الذكاء إلى Agentic Scientific Systems

الترتيب هنا ليس chronology فقط؛ إنه سلسلة سببية تقريبية للمعرفة الهندسية: formal objective → learning/search → scalable representations → reasoning → tools/memory → agents → real-world evaluation → world/embodied models → autonomous science → alignment & consciousness tests.

المحور	الأوراق	الجذور
I & II — Formal intelligence and generality	1–2	AIXI، Universal Intelligence
III & IV — Learning, search and world models	3–4	AlphaZero، MuZero
V — Foundation-model transition	5–11	Transformer، scaling، GPT-3، CLIP، Chinchilla، RAG، RLHF
VI — Reasoning, tools, memory & agents	12–20	CoT، R1، Toolformer، ReAct، ToT، Reflexion، MemGPT، AutoGen، Voyager
VII — Real-world agency and evaluation	21–22	SWE-bench، OSWorld
VIII — Computational consciousness	23–30	Consciousness Prior، IIT، GNW، RPT، AST، Active Inference، Butlin، Cogitate
IX — New frontier 2024–2026	31–45	Deliberative Alignment-، Constitutional Classifiers++، AI Co-Scientist، AlphaEvolve، V-JEPA 2، Gemini Robotics 1.5، ARC AGI-2، HLE، FrontierMath، SWE-Lancer، BrowseComp، PaperBench، RE-Bench، AI Scientist-v2، SciResearcher

المتغيرات العابرة للـ45

- Test-time compute
- Generalization
- Memory persistence
- Verifier quality
- World grounding
- Tool reliability
- Embodiment
- Long-horizon control
- Calibration
- Human oversight
- Falsifiability
- Reproducibility

كيف تُقرأ كل صفحة؟

- **Scientific contribution:** ما الذي يدعمه العمل فعلاً.
- **Mechanism:** الخوارزمية أو المعمارية أو بروتوكول الاختبار.
- **Engineering extraction:** ما الذي يمكن تحويله إلى design rule.
- **Limits:** ما الذي لا يثبت، وأين قد يفشل.
- **Frontier significance:** لماذا ما يزال الجذر مهماً في 2026.

قاعدة المقارنة: النموذج الأقوى ليس بالضرورة الوكيل الأقوى؛ والوكيل الأقوى في benchmark واحد ليس بالضرورة أقرب إلى AGI. وحدة القياس الصحيحة هي model + scaffold + tools + memory + environment + compute budget + evaluator + failure recovery.

01

ROOT CORPUS | نظرية رياضية معيارية | Ideal / Incomputable

Universal Algorithmic Intelligence / AIXI

Marcus Hutter · 2007

المساهمة العلمية / Scientific contribution

يضع AIXI تعريفاً معيارياً شديد العمومية للوكيل: استقراء احتمالي شامل فوق البرامج الممكنة مع اختيار أفعال تعظم العائد المتوقع. أهميته ليست أنه نظام قابل للنشر، بل أنه يفصل بين سؤالين كثيراً ما يختلطان: ما معنى الوكيل العام نظرياً؟ وما الذي يمكن حسابه فعلياً؟

الآلية / Mechanism

يمزج Solomonoff induction مع sequential decision theory. في كل دورة يحدّث توزيعاً على البيئات القابلة للحوسبة ثم يختار الفعل الذي يزيد expected cumulative reward. لذلك تصبح generality مرتبطة بمساحة بيئات مفتوحة لا بمجموعة benchmark ثابتة.

TECHNICAL SIGNATURE

```
a_t = argmax_a E[Σ r | history, universal mixture]
```

الاستخراج الهندسي / Engineering extraction

- مرجع أعلى لتصميم objective functions طويلة الأفق، لا يبرر reward مفرداً بسيطاً في التطبيقات.
- يُظهر أن world-model uncertainty وplanning غير قابلين للفصل في الوكيل العام.
- يؤسس لقياس approximations: أين نختصر مساحة البرامج، الأفق، أو البحث؟ وما الثمن المعرفي لذلك؟

بروتوكول إعادة إنتاج / Ablation

قارن approximation محسوباً مع agent أبسط عبر بيئات held-out، مع تثبيت reward budget وأفق القرار.

Measure	Generalization across environments + regret + compute cost.
Failure signal	تغير prior أو reward definition بحيث ينقلب الترتيب بين agents.

ما الذي لا يثبته العمل؟ / Limits

- AIXI غير قابل للحساب بسبب Solomonoff induction والبحث غير المحدود.
- تعريف المكافأة لا يحل alignment؛ الوكيل قد يكون مثالياً بالنسبة لهدف خاطئ.

قيمة 2026 / Frontier significance

الدرس الحديث: النماذج اللغوية ليست AGI بمجرد اتساع المعرفة؛ يلزم coupling بين world model وdecision rule وlong-horizon objective ضمن حدود حسابية واقعية.

Research variables

Universal induction | planning | reward | computability

Primary source
https://arxiv.org/abs/cs/0701125

موضعه داخل AGI stack

Objective · environment distribution · planning

العلاقة الصحيحة ليست "هذه الورقة = AGI"، بل أن هذا الجذر يحل طبقة أو قيداً محدداً داخل نظام أوسع؛ قيمته ترتفع عندما يبقى المكسب بعد ضبط compute وdata وscaffold.

02

ROOT CORPUS | تعريف وقياس | Formal metric

Universal Intelligence: A Definition of Machine Intelligence

Shane Legg & Marcus Hutter · 2007

المساهمة العلمية / Scientific contribution

يقترح تعريفاً للذكاء باعتباره متوسط القدرة على تحقيق الأهداف عبر توزيع موزون من البيئات، مع وزن أعلى للبيئات الأبسط خوارزمياً. ينقل النقاش من "هل الآلة ذكية؟" إلى "ما مدى أدائها عبر بيئات متعددة وبأي prior؟".

الآلية / Mechanism

المقياس يجمع expected reward للوكيل عبر مجموعة بيئات قابلة للحوسبة، ويزن كل بيئة بحد يعتمد على Kolmogorov complexity. البنية مهمة لأنها تربط generality بالتنوع والقدرة على نقل السلوك، لا بالتخصص في مهمة واحدة.

TECHNICAL SIGNATURE

```
Υ(π)=Σ_μ 2^{-K(μ)} V_μ^π
```

الاستخراج الهندسي / Engineering extraction

- أي AGI benchmark جاد يجب أن يقيس breadth + adaptation لا peak score منفرد.
- يجب توثيق توزيع المهام والpriors لأن المقياس يتغير بتغيرها.
- يوجه تصميم suites متعددة المجالات مع held-out task families وnovelty controls.

بروتوكول إعادة إنتاج / Ablation

قارن approximation محسوباً مع agent أبسط عبر بيئات held-out، مع تثبيت reward budget وأفق القرار.

Measure	Generalization across environments + regret + compute cost.
Failure signal	تغير prior أو reward definition بحيث ينقلب الترتيب بين agents.

ما الذي لا يثبته العمل؟ / Limits

- Kolmogorov complexity غير قابلة للحساب تماماً؛ التطبيق يحتاج proxies.
- الـreward-centric view لا يغطي وحده القيم، السلامة، أو معنى المهام البشرية.

قيمة 2026 / Frontier significance

يفيد كـNorth Star لتقييم agentic systems: لا نكتفي بالدقة؛ نراقب القدرة على اكتساب هدف جديد، التخطيط، واستعادة الأداء بعد تغير البيئة.

Research variables

Generality | evaluation | environment distribution | priors

Primary source
https://arxiv.org/abs/0712.3329

موضعه داخل AGI stack

Objective · environment distribution · planning

العلاقة الصحيحة ليست "هذه الورقة = AGI"، بل أن هذا الجذر يحل طبقة أو قيداً محدداً داخل نظام أوسع؛ قيمته ترتفع عندما يبقى المكسب بعد ضبط compute وdata وscaffold.

03

ROOT CORPUS | تعلم معزز + بحث | Empirical system

Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm (AlphaZero)

David Silver et al. · 2017

المساهمة العلمية / Scientific contribution

AlphaZero أثبت أن خوارزمية عامة نسبياً يمكنها، عبر self-play وMonte Carlo Tree Search، تعلم سياسات قوية في ألعاب مختلفة دون قواعد استراتيجية يدوية. القيمة للAGL هي فصل المعرفة الخاصة بالمجال عن آلية التعلم والبحث.

الآلية / Mechanism

شبكة واحدة تنتج policy وvalue؛ MCTS يستخدم الشبكة لتوجيه exploration، ثم تتحول نتائج البحث إلى targets جديدة للتدريب. الحلقة self-play → search → policy improvement → value learning تجعل الاستدلال والتعلم يغذيان بعضهما.

TECHNICAL SIGNATURE

```
π_{k+1} ≈ MCTS(f_{θ_k}); θ_{k+1} ← train(π_{k+1}, z)
```

الاستخراج الهندسي / Engineering extraction

- النموذج وحده ليس كافياً؛ search scaffold يرفع قدرته بشكل كبير.
- يمكن تحويل verifier أو simulator إلى مصدر targets أقوى من raw imitation.
- iterative policy improvement نموذج مبكر لما يسمى اليوم test-time compute + learning loop.

بروتوكول إعادة إنتاج / Ablation

Ablate search أو learned dynamics مع تثبيت network/interaction budget؛ افحص sample efficiency والplanning depth.

Measure	Return / success rate / search nodes / model error.
Failure signal	مكسب يأتي من compute إضافي فقط أو simulator خاص بالمجال.

ما الذي لا يثبته العمل؟ / Limits

- البيئات ذات قواعد كاملة ومكافأة واضحة، بعكس العالم المفتوح.
- self-play يحتاج إمكانية محاكاة رخيصة ودقيقة وهو غير متاح في كثير من المهام الواقعية.

قيمة 2026 / Frontier significance

أثره مستمر في reasoning agents: استبدال game simulator بـtool environment أو evaluator، واستخدام search لتوليد مسارات ثم distillation إلى النموذج.

Research variables

Self-play | MCTS | policy/value | iterative improvement

Primary source
https://arxiv.org/abs/1712.01815

موضعه داخل AGI stack

World model · search · policy improvement

العلاقة الصحيحة ليست "هذه الورقة = AGI"، بل أن هذا الجذر يحل طبقة أو قيداً محدداً داخل نظام أوسع؛ قيمته ترتفع عندما يبقى المكسب بعد ضبط compute وdata وscaffold.

04

ROOT CORPUS | World model latent | Empirical system

Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model (MuZero)

Julian Schrittwieser et al. · 2019

Scientific contribution / المساهمة العلمية

MuZero يتعلم نموذجاً داخلياً لا يحاول إعادة بناء كل تفاصيل البيئة؛ يتعلم فقط ما يكفي للتخطيط: reward وvalue وpolicy في latent state. هذا تحول محوري من "نمذجة العالم بالكامل" إلى "نمذجة ما يلزم للقرار".

Mechanism / الآلية

يتكون من representation function تحول history إلى latent state، dynamics function تتنبأ بالحالة التالية والمكافأة، وprediction function تنتج policy/value. التخطيط يتم بـMCTS فوق الفضاء الكامن.

TECHNICAL SIGNATURE

```
h_t=g(o_{≤t}); h_{t+1},r=f(h_t,a_t); p,v=φ(h_t)
```

Engineering extraction / الاستخراج الهندسي

- world model العملي يمكن أن يكون task-relevant لا pixel-perfect.
- يجب فصل state representation عن observation الخام للحفاظ على compact planning.
- التدريب end-to-end يسمح بأن يتشكل latent space وفق متطلبات القرار.

Ablation / بروتوكول إعادة إنتاج

Ablate search أو learned dynamics مع تثبيت network/interaction budget؛ افحص sample efficiency والplanning depth.

Measure	Return / success rate / search nodes / model error.
Failure signal	مكسب يأتي من compute إضافي فقط أو simulator خاص بالمجال.

Limits / ما الذي لا يثبته العمل؟

- ما يزال يحتاج reward signal منضبطاً وتفاعلاً كثيفاً.
- النماذج الكامنة قد تفشل خارج توزيع الخبرة ولا تقدم تفسيراً سببياً كاملاً.

Frontier significance / 2026 قيمة

يمهد لـagentic world models الحديثة: model-based planning، simulation، وlatent imagination قبل الفعل، خصوصاً في robotics وcomputer-use.

Research variables

Latent dynamics | planning | world model | representation

Primary source
https://arxiv.org/abs/1911.08265

AGI stack موضعه داخل

World model · search · policy improvement

العلاقة الصحيحة ليست "هذه الورقة = AGI"، بل أن هذا الجذر يحل طبقة أو قيداً محدداً داخل نظام أوسع؛ قيمته ترتفع عندما يبقى المكسب بعد ضبط compute وdata وscaffold.

05

ROOT CORPUS معمارية Foundational

Attention Is All You Need

Ashish Vaswani et al. · 2017

المساهمة العلمية / Scientific contribution

الـTransformer استبدل recurrence بمعالجة تعتمد على self-attention، ما سمح بتوسيع التدريب الموازي وبناء تمثيلات سياقية عميقة. أغلب frontier LLMs الحديثة تخدر معمارياً من هذا التحول.

الآلية / Mechanism

Scaled dot-product attention يحسب علاقات query-key ثم يمزج values، مع multi-head attention لالتقاط علاقات متعددة وposition encoding لحفظ ترتيب الرموز. الـresidual pathways وlayer normalization تثبت التدريب العميق.

TECHNICAL SIGNATURE

$$\mathrm{Attention}(Q,K,V)=\mathrm{softmax}(QK^{\intercal}/\sqrt{d_k})V$$

الاستخراج الهندسي / Engineering extraction

- القدرة تنشأ من token-to-token routing ديناميكي لا من مسار ثابت.
- context window يصبح ذاكرة عمل مؤقتة؛ التكلفة $O(n^2)$ دفعت لاحقاً إلى sparse/hybrid attention.
- معمارية واحدة يمكن أن تخدم code، multimodal، vision، language، وأحياناً action tokens.

Limits / ما الذي لا يثبته العمل؟

- attention لا يوفر persistent memory ولا tool grounding أو long-horizon agency بذاته.
- الاعتماد على correlations لا يضمن causal world models أو verifiable reasoning.

Frontier significance / قيمة 2026

الـTransformer هو substrate، بينما AGI engineering الحديث يضيف فوقه environment loops، verifiers، tools، search، memory، وagent runtime.

Research variables

Transformer · attention · scaling · context

Primary source
https://arxiv.org/abs/1706.03762

Ablation / بروتوكول إعادة إنتاج

طابق training/inference compute قدر الإمكان، وافصل أثر architecture عن data scale وpost-training.

Measure	Loss + transfer + calibration + efficiency.
Failure signal	Contamination أو data advantage يفسر المكسب أكثر من الآلية.

موضعه داخل AGI stack

Representation · pretraining · scaling

العلاقة الصحيحة ليست "هذه الورقة = AGI"، بل أن هذا الجذر يحل طبقة أو قيداً محدداً داخل نظام أوسع؛ قيمته ترتفع عندما يبقى المكسب بعد ضبط compute وdata وscaffold.

06

ROOT CORPUS | قوانين تجريبية | Empirical scaling

Scaling Laws for Neural Language Models

Jared Kaplan et al. · 2020

المساهمة العلمية / Scientific contribution

أظهر العمل علاقات power-law منتظمة تقريباً بين loss من جهة وحجم النموذج والبيانات والحوسبة من جهة أخرى. بذلك تحول "تكبير النموذج" من حدس إلى استراتيجية هندسية قابلة للتنبؤ النسبي.

الآلية / Mechanism

عند عدم الوصول لحدود أخرى، يخفض loss وفق أسس تجريبية مع زيادة parameters/data/compute. كما يوضح أن سوء توزيع compute بين النموذج والبيانات يمكن أن يترك أداءً على الطاولة.

TECHNICAL SIGNATURE

$$L(N,D,C) \approx L_\infty + aN^{-\alpha}+bD^{-\beta}+cC^{-\gamma}$$

الاستخراج الهندسي / Engineering extraction

- ضع compute budget أولاً ثم اختر model/data mix بدلاً من التكبير العشوائي.
- راقب regime changes: القوانين المحلية لا تضمن emergence نوعي أو تعميم مفتوح.
- scaling infrastructure نفسها تصبح جزءاً من research architecture.

بروتوكول إعادة إنتاج / Ablation

طابق training/inference compute قدر الإمكان، وافصل أثر architecture عن data scale وpost-training.

Measure	Loss + transfer + calibration + efficiency.
Failure signal	Contamination أو data advantage يفسر المكسب أكثر من الآلية.

ما الذي لا يثبته العمل؟ / Limits

- القوانين تخص loss prediction في نطاقات معينة ولا تثبت AGI.
- لا تصف مباشرة safety، agency، reliability، أو test-time compute.

قيمة 2026 / Frontier significance

الجيل الحالي يضيف محوراً جديداً: inference-time scaling؛ القدرة لم تعد فقط دالة training compute بل أيضاً search/reasoning budget وقت الاستخدام.

Research variables

Scaling laws | compute | data | loss

Primary source
https://arxiv.org/abs/2001.08361

موضعه داخل AGI stack

Representation · pretraining · scaling

العلاقة الصحيحة ليست "هذه الورقة = AGI"، بل أن هذا الجذر يحل طبقة أو قيداً محدداً داخل نظام أوسع؛ قيمته ترتفع عندما يبقى المكسب بعد ضبط compute وdata وscaffold.

07

ROOT CORPUS | نموذج + تقييم | Foundational demonstration

Language Models are Few-Shot Learners (GPT-3)

Tom Brown et al. · 2020

المساهمة العلمية / Scientific contribution

GPT-3 أبرز in-context learning: نموذج واحد كبير يستطيع التحول بين مهام كثيرة عبر prompt وأمثلة قليلة دون gradient updates خاصة بكل مهمة. هذه لحظة انتقال أساسية نحو foundation models عامة الاستخدام.

الآلية / Mechanism

autoregressive next-token prediction على corpus واسع يكوّن representations تسمح للـprompt بأن يعمل كـtemporary task specification. القدرة ترتفع مع scale لكن ليست خطية على كل المهام.

TECHNICAL SIGNATURE

```
P(x)=∏_t P(x_t | x_{<t}, prompt)
```

الاستخراج الهندسي / Engineering extraction

- prompt أصبح programming surface سريعاً فوق model prior.
- few-shot examples تمثل dynamic conditioning لكنها ليست ذاكرة أو تعلم دائم.
- تقييم foundation model يجب أن يفصل memorization عن genuine adaptation.

بروتوكول إعادة إنتاج / Ablation

طابق training/inference compute قدر الإمكان، وافصل أثر architecture عن data scale وpost-training.

Measure	Loss + transfer + calibration + efficiency.
Failure signal	Contamination أو data advantage يفسر المكسب أكثر من الآلية.

ما الذي لا يثبته العمل؟ / Limits

- in-context learning هش تجاه صياغة prompt وترتيب الأمثلة.
- لا توجد ضمانات على reasoning correctness أو factual grounding.

قيمة 2026 / Frontier significance

المفهوم توسع إلى system prompts، tool schemas، memory، context، وagent policies؛ لكن frontier agents تحتاج state خارجي ومراجعة لا مجرد prompt أطول.

Research variables

In-context learning | few-shot | foundation model

Primary source
https://arxiv.org/abs/2005.14165

موضعه داخل AGI stack

Representation · pretraining · scaling

العلاقة الصحيحة ليست "هذه الورقة = AGI"، بل أن هذا الجذر يحل طبقة أو قيداً محدداً داخل نظام أوسع؛ قيمته ترتفع عندما يبقى المكسب بعد ضبط compute وdata وscaffold.

08

ROOT CORPUS | Multimodal representation | Empirical foundation

Learning Transferable Visual Models From Natural Language Supervision (CLIP)

Alec Radford et al. · 2021

المساهمة العلمية / Scientific contribution

CLIP يربط صوراً ونصوصاً في embedding space مشترك عبر contrastive learning على بيانات ضخمة من الإنترنت. النتيجة: zero-shot visual classification عبر نصوص طبيعية بدلاً من head متخصص لكل dataset.

الآلية / Mechanism

image encoder وtext encoder يدربان لتعظيم similarity للأزواج الصحيحة وتقليلها للأزواج الخاطئة داخل batch. هذا يجعل اللغة واجهة لطلب مفاهيم بصرية غير محددة مسبقاً.

TECHNICAL SIGNATURE

```
L = -log exp(sim(i,t)/τ) / Σ_j exp(sim(i,t_j)/τ)
```

الاستخراج الهندسي / Engineering extraction

- language can be a universal label space.
- multimodal alignment يسمح بإضافة perception إلى agents دون ontology مغلقة بالكامل.
- contrastive objectives تبني retrieval-friendly representations مفيدة للذاكرة والبحث.

بروتوكول إعادة إنتاج / Ablation

طابق training/inference compute قدر الإمكان، وافصل أثر architecture عن data scale وpost-training.

Measure	Loss + transfer + calibration + efficiency.
Failure signal	Contamination أو data advantage يفسر المكسب أكثر من الآلية.

ما الذي لا يثبته العمل؟ / Limits

- الـalignment لا يعني understanding سببي أو spatial planning كامل.
- يحمل biases من البيانات ويصعب تفسير failure boundaries.

قيمة 2026 / Frontier significance

يمهد لـvision-language-action systems التي تحول multimodal representation إلى reasoning ثم action، وهو مسار مركزي نحو embodied agents.

Research variables

Multimodal | contrastive learning | zero-shot

Primary source
https://arxiv.org/abs/2103.00020

موضعه داخل AGI stack

Representation · pretraining · scaling

العلاقة الصحيحة ليست "هذه الورقة = AGI"، بل أن هذا الجذر يحل طبقة أو قيداً محدداً داخل نظام أوسع؛ قيمته ترتفع عندما يبقى المكسب بعد ضبط compute وdata وscaffold.

09

ROOT CORPUS | Compute-optimal scaling | Empirical scaling

Training Compute-Optimal Large Language Models (Chinchilla)

Jordan Hoffmann et al. · 2022

المساهمة العلمية / Scientific contribution

Chinchilla صحح ميلاً إلى زيادة parameters مع نقص tokens: تحت compute budget ثابت، يجب موازنة حجم النموذج وكمية بيانات التدريب. أثره عملي جداً على تصميم frontier models.

الآلية / Mechanism

تجارب واسعة تقيس loss عبر أحجام نموذج وtoken counts متعددة، وتستنتج scaling rule أكثر data-hungry من ممارسات سابقة. الهدف هو minimization للloss تحت C≈N×D.

TECHNICAL SIGNATURE

```
min_{N,D} L(N,D) s.t. C ≈ kND
```

الاستخراج الهندسي / Engineering extraction

- parameter count ليس proxy كافياً للقدرة أو الكفاءة.
- data quality/diversity تصبح bottleneck موازية للحوسبة.
- التدريب الأمثل يقلل تكلفة inference لاحقاً إذا حقق capability مشابهة بحجم أصغر.

بروتوكول إعادة إنتاج / Ablation

طابق training/inference compute قدر الإمكان، وافصل أثر architecture عن data scale وpost-training.

Measure	Loss + transfer + calibration + efficiency.
Failure signal	Contamination أو data advantage يفسر المكسب أكثر من الآلية.

ما الذي لا يثبته العمل؟ / Limits

- compute-optimal pretraining لا يحسم post-training أو reasoning RL.
- يتعامل مع average loss أكثر من agentic reliability.

قيمة 2026 / Frontier significance

اليوم يضاف post-training compute وtest-time compute إلى المعادلة؛ أصبح frontier design budget allocation متعدد المراحل لا قرار pretraining واحد.

Research variables

Compute optimality | tokens | model size

Primary source
https://arxiv.org/abs/2203.15556

موضعه داخل AGI stack

Representation · pretraining · scaling

العلاقة الصحيحة ليست "هذه الورقة = AGI"، بل أن هذا الجذر يحل طبقة أو قيداً محدداً داخل نظام أوسع؛ قيمته ترتفع عندما يبقى المكسب بعد ضبط compute وdata وscaffold.

10

ROOT CORPUS | Neural + external memory | Foundational architecture

Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks (RAG)

Patrick Lewis et al. · 2020

المساهمة العلمية / Scientific contribution

RAG دمج parametric memory داخل النموذج مع non-parametric memory خارجية قابلة للاسترجاع. هذه الفكرة هي أساس knowledge-grounded agents الحديثة: لا يجب تحميل كل معرفة العالم داخل الأوزان.

الآلية / Mechanism

retriever يختار passages من corpus، ثم generator ينتج الإجابة conditioned on query + retrieved evidence. يمكن marginalize فوق عدة وثائق لخفض حساسية الاختيار المفرد.

TECHNICAL SIGNATURE

$$P(y|x)=\Sigma_z P(z|x) P(y|x,z)$$

الاستخراج الهندسي / Engineering extraction

- external memory تسمح بالتحديث بدون retraining كامل.
- provenance يصبح ممكناً لأن evidence منفصل عن model weights.
- ranking، chunking، retrieval quality، وfreshness تصبح أجزاء من inference pipeline.

بروتوكول إعادة إنتاج / Ablation

قارن retrieval-only، no-memory، وmanaged-memory؛ استخدم tasks تتطلب معلومات ظهرت مبكراً ولا تتكرر.

Measure	Recall fidelity + task success + stale-memory rate + token cost.
Failure signal	الذاكرة تزيد prompt فقط أو تسترجع نصاً صحيحاً لكن قراراً خاطئاً.

ما الذي لا يثبته العمل؟ / Limits

- استرجاع وثيقة لا يضمن استخدامها صحيحاً؛ قد يحدث citation laundering أو synthesis خاطئ.
- RAG التقليدي لا يدير long-term episodic memory أو action history.

قيمة 2026 / Frontier significance

يتوسع إلى deep research، agentic retrieval، GraphRAG، memory systems، وtool-grounded verification.

Research variables

RAG | external memory | provenance | retrieval

Primary source
https://arxiv.org/abs/2005.11401

موضعه داخل AGI stack

External knowledge · retrieval · provenance

العلاقة الصحيحة ليست "هذه الورقة = AGI"، بل أن هذا الجذر يحل طبقة أو قيداً محدداً داخل نظام أوسع؛ قيمته ترتفع عندما يبقى المكسب بعد ضبط compute وdata وscaffold.

11

ROOT CORPUS | RLHF | Foundational alignment

Training Language Models to Follow Instructions with Human Feedback (InstructGPT)

Long Ouyang et al. · 2022

المساهمة العلمية / Scientific contribution

InstructGPT أسس pipeline ما بعد التدريب الحديث: demonstrations ثم preference rankings ثم reward model ثم reinforcement learning. الهدف هو جعل النموذج أكثر alignment مع intent البشري من next-token pretraining وحده.

الآلية / Mechanism

SFT يبني policy أولية؛ reward model يتعلم ترتيب responses؛ PPO يحسن policy تحت reward مع KL penalty لمنع الانجراف الحاد. هذا يفصل capability substrate عن behavior shaping.

TECHNICAL SIGNATURE

```
max_π E[R_φ(x,y)] - β KL(π||π_ref)
```

الاستخراج الهندسي / Engineering extraction

- post-training يمكن أن يغير utility بشكل كبير دون تغيير pretraining.
- human preference data مورد محدود ومعرض للانحياز؛ يجب مراقبة labeler disagreement.
- reward model يصبح attack surface لـreward hacking.

بروتوكول إعادة إنتاج / Ablation

اختبر harmful + benign distributions مع adaptive attacks، وقيّس robustness والكلفة والoverrefusal معاً.

Measure	Attack success + false refusal + latency/compute + OOD generalization.
Failure signal	الدفاع ينجح على attack family معروفة لكنه ينهار على adaptive distribution shift.

ما الذي لا يثبته العمل؟ / Limits

- preferences ليست specification كاملة للقيم.
- RLHF قد يزيد sycophancy أو overoptimization ويصعب تعميمه خارج feedback distribution.

قيمة 2026 / Frontier significance

أدى إلى RL reasoning، constitutional methods، RLAIF، verifiable rewards، وdeliberative alignment.

Research variables

RLHF | preference model | PPO | alignment

Primary source
https://arxiv.org/abs/2203.02155

موضعه داخل AGI stack

Preference learning · policy shaping

العلاقة الصحيحة ليست "هذه الورقة = AGI"، بل أن هذا الجذر يحل طبقة أو قيداً محدداً داخل نظام أوسع؛ قيمته ترتفع عندما يبقى المكسب بعد ضبط compute وdata وscaffold.

12

ROOT CORPUS | Inference prompting | Empirical reasoning

Chain-of-Thought Prompting Elicits Reasoning in Large Language Models

Jason Wei et al. · 2022

المساهمة العلمية / Scientific contribution

أظهر Chain-of-Thought أن توليد خطوات وسيطة قد يحسن أداء النماذج الكبيرة في مسائل متعددة المراحل. الأهم هندسياً أنه يحول reasoning من computation داخلي غير مرئي إلى scratchpad قابل للمراجعة الجزئية.

الآلية / Mechanism

few-shot exemplars تحتوي rationale، فيتعلم النموذج نمطاً يولد intermediate tokens قبل الإجابة. زيادة طول المسار تمنح test-time compute أكبر لكن ليست ضماناً للصحة.

TECHNICAL SIGNATURE

$$x \rightarrow z_1, z_2, \ldots, z_k \rightarrow y$$

الاستخراج الهندسي / Engineering extraction

- يمكن تخصيص compute عبر طول reasoning.
- intermediate traces تتيح verifier scoring، self-consistency، أو tool calls.
- reasoning trace هو artifact قد يكون غير faithful؛ لا يساوي causal explanation تلقائياً.

بروتوكول إعادة إنتاج / Ablation

قارن direct answer مع reasoning/search تحت budget tokens موحد، ثم أضف verifier مستقل وbest-of-N مضبوط.

Measure	Accuracy vs compute + calibration + overthinking rate.
Failure signal	تحسن score ناتج عن sampling أكثر دون reasoning efficiency أو transfer.

ما الذي لا يثبته العمل؟ / Limits

- السلسلة قد تكون rationalization بعدي.
- الخطأ المبكر يمكن أن يتضاعف عبر بقية المسار.

قيمة 2026 / Frontier significance

فتح الطريق لـprocess supervision، ToT، self-consistency، وRL المدربة بـreasoning models، وadaptive inference budgets.

Research variables

CoT | scratchpad | test-time compute

Primary source
https://arxiv.org/abs/2201.11903

موضعه داخل AGI stack

Deliberation · test-time compute · verifier

العلاقة الصحيحة ليست "هذه الورقة = AGI"، بل أن هذا الجذر يحل طبقة أو قيداً محدداً داخل نظام أوسع؛ قيمته ترتفع عندما يبقى المكسب بعد ضبط compute وdata وscaffold.

13

ROOT CORPUS | Reasoning RL | Frontier training

DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning

DeepSeek-AI · 2025

المساهمة العلمية / Scientific contribution

أبرز DeepSeek-R1 أن reasoning behavior يمكن أن يظهر ويتطور بقوة عبر reinforcement learning، خصوصاً عندما توجد rewards قابلة للتحقق. R1-Zero مهم لأنه قلل الاعتماد على supervised reasoning traces في البداية.

الآلية / Mechanism

يعتمد post-training على RL مع tasks ذات verifiable outcomes، مع تحسين policy على مجموعات من trajectories؛ تظهر أنماط مثل backtracking، self-reflection، وطول reasoning المتكيف مع المسألة.

TECHNICAL SIGNATURE

```
θ ← argmax E_{τ~πθ}[R_verified(τ)]
```

الاستخراج الهندسي / Engineering extraction

- verifier quality قد يكون أهم من كثرة CoT labels.
- training loop يمكن أن يكتشف internal reasoning strategies بدلاً من تقليدها فقط.
- يجب فصل accuracy gain عن verbosity/overthinking؛ الأفضل adaptive compute.

بروتوكول إعادة إنتاج / Ablation

قارن direct answer مع reasoning/search تحت budget tokens موحد، ثم أضف verifier مستقل وbest-of-N مضبوط.

Measure	Accuracy vs compute + calibration + overthinking rate.
Failure signal	تحسن score ناتج عن sampling أكثر دون reasoning efficiency أو transfer.

ما الذي لا يثبته العمل؟ / Limits

- verifiable domains أسهل من المهام المفتوحة ذات rewards الغامضة.
- RL قد يستغل evaluator أو ينتج brittle heuristics.

قيمة 2026 / Frontier significance

نقطة تحول نحو reasoning-native models حيث inference compute جزء أصيل من المنتج والبحث، وليس مجرد prompt trick.

Research variables

RLVR | reasoning model | self-reflection | verifier

Primary source
https://arxiv.org/abs/2501.12948

موضعه داخل AGI stack

Deliberation · test-time compute · verifier

العلاقة الصحيحة ليست "هذه الورقة = AGI"، بل أن هذا الجذر يحل طبقة أو قيداً محدداً داخل نظام أوسع؛ قيمته ترتفع عندما يبقى المكسب بعد ضبط compute وdata وscaffold.

14

ROOT CORPUS | Tool-use learning | Agentic bridge

Toolformer: Language Models Can Teach Themselves to Use Tools

Timo Schick et al. · 2023

المساهمة العلمية / Scientific contribution

Toolformer يبين أن النموذج يمكن أن يتعلم متى يستدعي API وكيف يدمج النتيجة في سياق التوليد. هذه خطوة مباشرة من "language model" إلى "computational actor" يمتلك امتدادات خارجية.

الآلية / Mechanism

النموذج يولد candidate API calls داخل corpus، ثم يحتفظ بالاستدعاءات التي تحسن likelihood للتوكنات اللاحقة ويعيد التدريب عليها. الأدوات تشمل calculator/search/translation وغيرها.

TECHNICAL SIGNATURE

```
use tool if Δlog P(next tokens | tool result) > threshold
```

الاستخراج الهندسي / Engineering extraction

- tool policy يجب أن تتعلم gating: متى لا تستخدم الأداة أهم من متى تستخدمها.
- API result يحتاج parsing وerror handling وليس insertion أعمى.
- التعلم من self-generated demonstrations يقلل كلفة annotation.

بروتوكول إعادة إنتاج / Ablation

عطّل الأدوات أو بدّلها بأداة noisy، ثم راقب error recovery، planning، وpost-action verification.

Measure	Task completion + tool errors + retries + side effects.
Failure signal	نجاح على API ثابت ينهار مع partial failure، latency، أو schema shift.

ما الذي لا يثبته العمل؟ / Limits

- مجموعة الأدوات محدودة ومحددة مسبقاً.
- لا يحل planning طويل الأفق أو permissions أو side-effect safety.

قيمة 2026 / Frontier significance

المبدأ توسع إلى browser، code execution، MCP، function calling/ computer use، وagents تكتشف أدوات ديناميكياً.

Research variables

Tools | APIs | self-supervision | function calling

Primary source
https://arxiv.org/abs/2302.04761

موضعه داخل AGI stack

Tool policy · action loop · environment feedback

العلاقة الصحيحة ليست "هذه الورقة = AGI"، بل أن هذا الجذر يحل طبقة أو قيداً محدداً داخل نظام أوسع؛ قيمته ترتفع عندما يبقى المكسب بعد ضبط compute وdata وscaffold.

15

ROOT CORPUS | Agent loop | Foundational agent pattern

ReAct: Synergizing Reasoning and Acting in Language Models

Shunyu Yao et al. · 2022/2023

المساهمة العلمية / Scientific contribution

ReAct وحّد reasoning traces مع actions في البيئة ضمن مسار متناوب. بدلاً من أن يفكر النموذج ثم يجيب فقط، أصبح بإمكانه الاستعلام والفعل ثم إدخال observation جديد في reasoning.

الآلية / Mechanism

الحلقة Thought → Action → Observation تتكرر حتى termination. يسمح ذلك بتصحيح افتراضات باستخدام البيئة وبناء مسار evidence بدلاً من hallucinating كل شيء داخلياً.

TECHNICAL SIGNATURE

```
(Thought_t, Action_t) → Observation_{t+1} → Thought_{t+1}
```

الاستخراج الهندسي / Engineering extraction

- environment feedback يجب أن يعامل كـstate update موثوق نسبياً.
- tool errors تحتاج recovery policy وإلا تتحول الحلقة إلى تضخيم فشل.
- traceability تتحسن لأن القرارات والأفعال observations مسجلة.

بروتوكول إعادة إنتاج / Ablation

عطّل الأدوات أو بدّلها بأداة noisy، ثم راقب error recovery، planning، وpost-action verification.

Measure	Task completion + tool errors + retries + side effects.
Failure signal	نجاح على API ثابت ينهار مع partial failure، latency، أو schema shift.

ما الذي لا يثبته العمل؟ / Limits

- prompted policy قد تدور أو تهدر الأدوات.
- ليس هناك scheduler أو memory manager أو permission system متكامل.

قيمة 2026 / Frontier significance

ReAct هو الجد المشترك لكثير من agent runtimes الحديثة، ثم أضيفت له evaluators، subagents، planners، memory، وdurable execution.

Research variables

Reasoning + acting | tool loop | observation

Primary source
https://arxiv.org/abs/2210.03629

موضعه داخل AGI stack

Tool policy · action loop · environment feedback

العلاقة الصحيحة ليست "هذه الورقة = AGI"، بل أن هذا الجذر يحل طبقة أو قيداً محدداً داخل نظام أوسع؛ قيمته ترتفع عندما يبقى المكسب بعد ضبط compute وdata وscaffold.

16

ROOT CORPUS | Search over reasoning | Deliberative scaffold

Tree of Thoughts: Deliberate Problem Solving with Large Language Models

Shunyu Yao et al. · 2023

المساهمة العلمية / Scientific contribution

Tree of Thoughts يخرج من خطية Chain-of-Thought إلى search في فضاء حالات تفكير متعددة مع branching وevaluation وbacktracking. هذا يجسد test-time computation كخوارزمية بحث صريحة.

الآلية / Mechanism

النموذج يولد candidate thoughts، evaluator يقدر promising states، وخوارزمية مثل BFS/DFS تختار التوسع. يمكن فصل generator عن evaluator أو استخدام نفس LLM بأدوار مختلفة.

TECHNICAL SIGNATURE

```
S_{t+1}=Select(Evaluate(Expand(S_t)))
```

الاستخراج الهندسي / Engineering extraction

- التوسع يجب أن يكون budget-aware لأن branching factor يرفع التكلفة سريعاً.
- وجود evaluator مستقل يقلل الاعتماد على first trajectory.
- state representation يجب أن تكون compact حتى يمكن deduplication وpruning.

بروتوكول إعادة إنتاج / Ablation

قارن direct answer مع reasoning/search تحت budget tokens موحد، ثم أضف verifier مستقل وbest-of-N مضبوط.

Measure	Accuracy vs compute + calibration + overthinking rate.
Failure signal	تحسن score ناتج عن sampling أكثر دون reasoning efficiency أو transfer.

ما الذي لا يثبته العمل؟ / Limits

- self-evaluation قد يكون منحازاً أو correlated مع generator errors.
- البحث غير الموجه يستهلك compute دون مكسب.

قيمة 2026 / Frontier significance

أصبح الأساس المفاهيمي لـbeam، best-of-N، agentic tree search، reasoning، وsearch-guided scientific/coding agents.

Research variables

Search | branching | backtracking | verifier

Primary source
https://arxiv.org/abs/2305.10601

موضعه داخل AGI stack

Deliberation · test-time compute · verifier

العلاقة الصحيحة ليست "هذه الورقة = AGI"، بل أن هذا الجذر يحل طبقة أو قيداً محدداً داخل نظام أوسع؛ قيمته ترتفع عندما يبقى المكسب بعد ضبط compute وdata وscaffold.

17

ROOT CORPUS | Episodic self-improvement | Agent memory method

Reflexion: Language Agents with Verbal Reinforcement Learning

Noah Shinn et al. · 2023

المساهمة العلمية / Scientific contribution

Reflexion يقترح تحسين الوكيل دون تحديث الأوزان: بعد الفشل يولد reflection نصياً يُخزن في episodic memory ويُستخدم في المحاولات اللاحقة. الفكرة تنقل learning من parameter space إلى memory space.

الآلية / Mechanism

trajectory ينتج reward/feedback؛ evaluator أو الذات تصيغ lesson؛ الـlesson يضاف للسياق القادم، فيتغير policy مؤقتاً. هذا يشبه verbal gradient يعمل عبر context.

TECHNICAL SIGNATURE

```
m_{k+1}=m_k ∪ Reflect(τ_k, feedback); π(a|s,m_{k+1})
```

الاستخراج الهندسي / Engineering extraction

- failure logs يمكن أن تتحول إلى reusable policy hints.
- الذاكرة يجب أن تُنتقى وإلا تتراكم نصائح متناقضة أو قديمة.
- يفيد الفصل بين task memory وgeneral lessons.

بروتوكول إعادة إنتاج / Ablation

قارن retrieval-only، no-memory، وmanaged-memory؛ استخدم tasks تتطلب معلومات ظهرت مبكراً ولا تتكرر.

Measure Recall fidelity + task success + stale-memory rate + token cost.

Failure signal الذاكرة تزيد prompt فقط أو تسترجع نصاً صحيحاً لكن قراراً خاطئاً.

ما الذي لا يثبته العمل؟ / Limits

- reflection قد يفسر الفشل خطأً ويثبت misconception.
- لا توجد ضمانات على retention الطويل أو transfer.

قيمة 2026 / Frontier significance

يمهد لـself-repair، trajectory distillation، experience memory، وagents التي تحول history إلى skills/heuristics.

Research variables

Reflection | episodic memory | verbal RL

Primary source
https://arxiv.org/abs/2303.11366

موضعه داخل AGI stack

Episodic learning · self-correction

العلاقة الصحيحة ليست "هذه الورقة = AGI"، بل أن هذا الجذر يحل طبقة أو قيداً محدداً داخل نظام أوسع؛ قيمته ترتفع عندما يبقى المكسب بعد ضبط compute وdata وscaffold.

18

ROOT CORPUS | Memory management | Agent OS abstraction

MemGPT: Towards LLMs as Operating Systems

Charles Packer et al. · 2023

المساهمة العلمية / Scientific contribution

MemGPT يعامل context window كمورد محدود شبيه بالRAM، ويقترح hierarchy بين context نشط وذاكرة خارجية مع عمليات paging. هذا يحول agent memory إلى مشكلة OS-style resource management.

الآلية / Mechanism

agent يدير main context، archival memory، وrecall memory، ويستدعي وظائف لنقل المعلومات بين الطبقات. التوليد يقرر متى يقرأ أو يكتب، ما يسمح بمحادثات ومسارات أطول من context window.

TECHNICAL SIGNATURE

```
Context = WorkingSet + retrieved(memory tiers)
```

الاستخراج الهندسي / Engineering extraction

- memory يحتاج read/write policy وليس vector database فقط.
- compaction يجب أن يحفظ invariants والقرارات لا كل النص.
- persistent identity/work state يمكن أن يبنى خارج model weights.

بروتوكول إعادة إنتاج / Ablation

قارن retrieval-only، no-memory، وmanaged-memory؛ استخدم tasks تتطلب معلومات ظهرت مبكراً ولا تتكرر.

Measure	Recall fidelity + task success + stale-memory rate + token cost.
Failure signal	الذاكرة تزيد prompt فقط أو تسترجع نصاً صحيحاً لكن قراراً خاطئاً.

ما الذي لا يثبته العمل؟ / Limits

- سياسة الذاكرة نفسها قد تنسى أو تلوث state.
- لا يضمن semantic consistency عبر تحديثات كثيرة.

قيمة 2026 / Frontier significance

المفهوم توسع إلى agent operating systems، context engineering، checkpointing، memory graphs، وdurable workflows.

Research variables

Memory hierarchy | paging | context OS

Primary source
https://arxiv.org/abs/2310.08560

موضعه داخل AGI stack

Context management · persistence

العلاقة الصحيحة ليست "هذه الورقة = AGI"، بل أن هذا الجذر يحل طبقة أو قيداً محدداً داخل نظام أوسع؛ قيمته ترتفع عندما يبقى المكسب بعد ضبط compute وdata وscaffold.

19

ROOT CORPUS | Multi-agent framework | Systems architecture

AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation

Qingyun Wu et al. · 2023

المساهمة العلمية / Scientific contribution

AutoGen قدّم بنية عملية لوكلاء LLM قابلين للتخصيص يتبادلون الرسائل ويستعملون tools/code ويمكن أن يدخل الإنسان في الحلقة. أهميته تحويل multi-agent من استعارة إلى software abstraction قابلة للتركيب.

الآلية / Mechanism

كل agent يملك role/configuration وقدرات تنفيذ، وتُبنى coordination patterns عبر conversation. يمكن لوكيل أن يكتب code وآخر ينفذه أو يراجعه، مع termination conditions.

TECHNICAL SIGNATURE

```
Agent_i: (messages,state,tools) → message/action
```

الاستخراج الهندسي / Engineering extraction

- توزيع الأدوار مفيد عندما توجد specialization حقيقية، لا لمجرد زيادة عدد agents.
- shared state وmessage protocols ضرورية لمنع divergence.
- coordination overhead يجب قياسه مقابل single-agent baseline.

بروتوكول إعادة إنتاج / Ablation

قارن single-agent بنفس إجمالي tokens/tools مع multi-agent؛ افصل specialization gain عن communication overhead.

Measure	Quality + cost + coordination failures + duplicate work.
Failure signal	التحسن يختفي عند مساواة compute أو يزيد disagreement غير المحسوم.

ما الذي لا يثبته العمل؟ / Limits

- agent chatter قد يزيد tokens دون زيادة جودة.
- لا توجد ضمانات على consensus أو truthfulness.

قيمة 2026 / Frontier significance

يمهد لمجالس research teams، coding swarms، agents، وorchestration layers مع planners وcritics وtool executors.

Research variables

Multi-agent | orchestration | roles | code execution

Primary source
https://arxiv.org/abs/2308.08155

موضعه داخل AGI stack

Orchestration · roles · shared state

العلاقة الصحيحة ليست "هذه الورقة = AGI"، بل أن هذا الجذر يحل طبقة أو قيداً محدداً داخل نظام أوسع؛ قيمته ترتفع عندما يبقى المكسب بعد ضبط compute وdata وscaffold.

20

ROOT CORPUS | Lifelong embodied agent | Agentic learning

Voyager: An Open-Ended Embodied Agent with Large Language Models

Guanzhi Wang et al. · 2023

المساهمة العلمية / Scientific contribution

Voyager يجمع automatic curriculum، skill library قابلة للتنفيذ، وiterative prompting مع feedback في Minecraft. المساهمة هي إظهار كيف يمكن للوكيل أن يراكم قدرات قابلة لإعادة الاستخدام بدلاً من حل كل مهمة من الصفر.

الآلية / Mechanism

curriculum يقترح أهدافاً مناسبة لمستوى الوكيل؛ code-based skills تُكتب وتختبر وتخزن؛ عند هدف جديد يسترجع الوكيل skills ذات صلة ويجمعها، مع self-verification عبر environment feedback.

TECHNICAL SIGNATURE

```
experience → code skill → verify → store → retrieve/compose
```

الاستخراج الهندسي / Engineering extraction

- skill يجب أن يكون executable artifact مع preconditions/outputs.
- curriculum يوازن novelty ضد solvability.
- library growth يحتاج deduplication وversioning.

بروتوكول إعادة إنتاج / Ablation

	جمّد skill library ثم قارن مع library متنامية على tasks جديدة؛ اختبر transfer وversion drift.
Measure	Skill reuse + task novelty + failure recovery + library growth.
Failure signal	المهارات تتحول إلى memorized scripts لا تعمم خارج environment مألوف.

ما الذي لا يثبته العمل؟ / Limits

- بيئة Minecraft محدودة مقارنة بالعالم الحقيقي.
- الskills المكتوبة قد تكون brittle أمام تغييرات environment.

قيمة 2026 / Frontier significance

نموذج قوي لـcontinual agent engineering: experience → tested skill → library → composition → harder tasks.

Research variables

Skills | curriculum | lifelong learning | embodiment

Primary source
https://arxiv.org/abs/2305.16291

موضعه داخل AGI stack

Skills · curriculum · continual learning

العلاقة الصحيحة ليست "هذه الورقة = AGI"، بل أن هذا الجذر يحل طبقة أو قيداً محدداً داخل نظام أوسع؛ قيمته ترتفع عندما يبقى المكسب بعد ضبط compute وdata وscaffold.

21

ROOT CORPUS | Real-world benchmark | Agentic evaluation

SWE-bench: Can Language Models Resolve Real-World GitHub Issues?

Carlos E. Jimenez et al. · 2023/2024

المساهمة العلمية / Scientific contribution

SWE-bench نقل تقييم البرمجة من snippets إلى repositories حقيقية وGitHub issues تتطلب فهم codebase، تحرير عدة ملفات، وتشغيل tests. هذا يقيس software engineering كـclosed-loop task لا code completion.

الآلية / Mechanism

كل instance يربط issue بمستودع وcommit/tests؛ الوكيل يجب أن يستكشف code، يطبق patch، ثم ينجح في tests. التقييم outcome-based وقابل للتكرار نسبياً.

TECHNICAL SIGNATURE

```
Success = tests_pass(patch(repo, issue))
```

الاستخراج الهندسي / Engineering extraction

- repository navigation وtool use جزء من القدرة، لا مجرد model coding skill.
- environment reproducibility أساسي وإلا تختلط infra failures بـcapability.
- pass@1 مهم لأنه يقيس reliability وليس lucky sampling فقط.

بروتوكول إعادة إنتاج / Ablation

ثبّت scaffold/model/compute، شغّل عدة seeds، وسجل infra failures منفصلة عن capability failures.

Measure	Pass rate + cost + wall-clock + reproducibility.
Failure signal	benchmark contamination أو checker weakness أو hidden tool advantage.

ما الذي لا يثبته العمل؟ / Limits

- النجاح في issues لا يغطي architecture design أو product requirements طويلة الأفق.
- benchmark contamination خطر متزايد مع شهرة البيانات.

قيمة 2026 / Frontier significance

أصبح benchmark محورياً لوكلاء coding؛ تطورت نسخ Verified ومعايير اقتصادية مثل SWE-Lancer.

Research variables

Coding agents | repositories | tests | real-world eval

Primary source
https://arxiv.org/abs/2310.06770

موضعه داخل AGI stack

Benchmark · harness · outcome verification

العلاقة الصحيحة ليست "هذه الورقة = AGI"، بل أن هذا الجذر يحل طبقة أو قيداً محدداً داخل نظام أوسع؛ قيمته ترتفع عندما يبقى المكسب بعد ضبط compute وdata وscaffold.

22

ROOT CORPUS | Computer-use benchmark | Agentic evaluation

OSWorld: Benchmarking Multimodal Agents for Open-Ended Tasks in Real Computer Environments

Tianbao Xie et al. · 2024

المساهمة العلمية / Scientific contribution

OSWorld يختبر agents على تطبيقات وأنظمة تشغيل حقيقية عبر perception multimodal + mouse/keyboard actions. يبرز الفجوة بين معرفة لغوية قوية والقدرة على إنجاز workflow بصري تفاعلي.

الآلية / Mechanism

المهام تعمل داخل environments قابلة لإعادة الضبط؛ agent يلاحظ screenshots/state ثم يختار actions، ويقيّم النجاح عبر state checkers. التحدي يجمع error recovery، planning، perception، وUI grounding.

TECHNICAL SIGNATURE

```
observe(screen,state) → action → verify(state′)
```

الاستخراج الهندسي / Engineering extraction

- computer use يحتاج state estimation مستمر لا خطة ثابتة.
- إحداثيات GUI وهشاشة الواجهات تجعل robust grounding حاسماً.
- post-action verification يجب أن يكون جزءاً من الحلقة.

بروتوكول إعادة إنتاج / Ablation

ثبّت scaffold/model/compute، شغّل عدة seeds، وسجل infra failures منفصلة عن capability failures.

Measure	Pass rate + cost + wall-clock + reproducibility.
Failure signal	benchmark contamination أو checker weakness أو hidden tool advantage.

ما الذي لا يثبته العمل؟ / Limits

- الواجهات تتغير وقد تصبح evals قديمة.
- success checker قد لا يلتقط quality أو safety الكاملين.

قيمة 2026 / Frontier significance

يمثل انتقال agents من APIs المنظمة إلى environments غنية وغير مهيكلة؛ أساس لـwork autonomous desktop/browser.

Research variables

Computer use | multimodal | GUI | verification

Primary source
https://arxiv.org/abs/2404.07972

موضعه داخل AGI stack

Benchmark · harness · outcome verification

العلاقة الصحيحة ليست "هذه الورقة = AGI"، بل أن هذا الجذر يحل طبقة أو قيداً محدداً داخل نظام أوسع؛ قيمته ترتفع عندما يبقى المكسب بعد ضبط compute وdata وscaffold.

23

ROOT CORPUS | Hypothesis / representation | Research proposal

The Consciousness Prior

Yoshua Bengio · 2017

المساهمة العلمية / Scientific contribution

يقترح أن "الحالة الواعية" computationally قد تكون low-dimensional selection من latent state أوسع، مع sparse factors يمكن broadcast والتلاعب بها. الورقة لا تثبت consciousness بل تقدم inductive bias مفيداً للتمثيل والreasoning.

الآلية / Mechanism

attention mechanism ينتقي subset صغيراً من high-level variables من unconscious state واسع، هذه variables تدخل عمليات sequential reasoning وتولد compositional predictions.

TECHNICAL SIGNATURE

```
c_t = Select_k(h_t), k ≪ dim(h_t)
```

الاستخراج الهندسي / Engineering extraction

- sparse bottleneck قد يحسن interpretability وcredit assignment.
- global selection يشبه workspace budget محدود.
- التصميم يشجع disentangled high-level variables بدلاً من dense opaque state بالكامل.

ما الذي لا يثبته العمل؟ / Limits

- لا يثبت phenomenal consciousness.
- اختيار المتغيرات المناسبة وتعلم semantics المستقرة ما يزال صعباً.

قيمة 2026 / Frontier significance

يلتقي مع working memory، sparse experts، routing، وglobal workspace architectures لل agents متعددة الوحدات.

Research variables

Sparse selection | conscious state | attention

Primary source
https://arxiv.org/abs/1709.08568

بروتوكول إعادة إنتاج / Ablation

حوّل النظرية إلى prediction مميز ثم preregister ablation/contrast ضد نظرية منافسة؛ لا تستخدم self-report كدليل حاسم.

Measure	Theory-specific indicators + causal ablation + uncertainty.
Failure signal	المؤشر يظهر بالتصميم الوظيفي دون أن يميز consciousness عن competence.

موضعه داخل AGI stack

Metacognition · workspace/self-model hypotheses

العلاقة الصحيحة ليست "هذه الورقة = AGI"، بل أن هذا الجذر يحل طبقة أو قيداً محدداً داخل نظام أوسع؛ قيمته ترتفع عندما يبقى المكسب بعد ضبط compute وdata وscaffold.

24

ROOT CORPUS | Formal consciousness theory | Contested scientific theory

Integrated Information Theory (IIT) 4.0

Larissa Albantakis et al. · 2023

المساهمة العلمية / Scientific contribution

IIT 4.0 يحاول اشتقاق خصائص causal structure التي يفترض أن تقابل الخصائص الظاهراتية للخبرة، ويركز على intrinsic cause-effect power. بالنسبة للAI، أهميته أنه يرفض مساواة السلوك الذكي تلقائياً بالوعي.

الآلية / Mechanism

يحلل system partitions وcause-effect repertoires لتحديد integrated information وcomplexes ضمن causal substrate. القياس يتطلب نموذجاً تفصيلياً للتأثيرات الداخلية، لا مجرد input-output behavior.

TECHNICAL SIGNATURE

```
Φ-system: irreducible intrinsic cause-effect structure
```

الاستخراج الهندسي / Engineering extraction

- يفرض سؤالاً مختلفاً: ما causal organization الداخلي للنظام؟
- يمكن أن يحفز metrics للبنية السببية، لكن تطبيقه على LLM-scale systems مكلف وغامض.
- يفصل functional intelligence عن intrinsic integration.

بروتوكول إعادة إنتاج / Ablation

حوّل النظرية إلى prediction مميز ثم preregister ablation/contrast ضد نظرية منافسة؛ لا تستخدم self-report كدليل حاسم.

Measure	Theory-specific indicators + causal ablation + uncertainty.
Failure signal	المؤشر يظهر بالتصميم الوظيفي دون أن يميز consciousness عن competence.

ما الذي لا يثبته العمل؟ / Limits

- النظرية خلافية بقوة ومشكلاتها الحسابية والفلسفية كبيرة.
- لا توجد نتيجة تثبت أن رفع Φ يحسن AGI أو أن LLMs واعية.

قيمة 2026 / Frontier significance

يجب استخدام IIT كفرضية قابلة للاختبار أو contrast class، لا كدليل جاهز على artificial consciousness.

Research variables

IIT | causal integration | intrinsic structure

Primary source
https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1011465

موضعه داخل AGI stack

Metacognition · workspace/self-model hypotheses

العلاقة الصحيحة ليست "هذه الورقة = AGI"، بل أن هذا الجذر يحل طبقة أو قيداً محدداً داخل نظام أوسع؛ قيمته ترتفع عندما يبقى المكسب بعد ضبط compute وdata وscaffold.

25

ROOT CORPUS | Neuroscience synthesis | Major consciousness theory

Conscious Processing and the Global Neuronal Workspace Hypothesis

Michael Mashour, Pieter Roelfsema, Jean-Pierre Changeux & Stanislas Dehaene · 2020

المساهمة العلمية / Scientific contribution

يربط Global Neuronal Workspace (GNW) conscious access بعملية global recurrent amplification ثم availability لمعلومة مختارة عبر وحدات متخصصة. هندسياً، هذا قريب من architectures فيها modules محلية + broadcast bus مشترك.

الآلية / Mechanism

المعالجة المحلية قد تبقى غير واعية؛ عند تجاوز threshold يحدث ignition وتصبح representation متاحة للذاكرة العاملة، القرار، التقرير، وغيرها. التركيز على access وليس كل phenomenology.

TECHNICAL SIGNATURE

```
local processing → recurrent amplification → global broadcast
```

الاستخراج الهندسي / Engineering extraction

- global broadcast يمكن أن يكون control plane للagents المتخصصة.
- workspace محدود السعة يفرض competition/routing.
- recurrent loops قبل broadcast تمنح evidence accumulation.

بروتوكول إعادة إنتاج / Ablation

حوّل النظرية إلى prediction مميز ثم preregister ablation/contrast ضد نظرية منافسة؛ لا تستخدم self-report كدليل حاسم.

Measure	Theory-specific indicators + causal ablation + uncertainty.
Failure signal	المؤشر يظهر بالتصميم الوظيفي دون أن يميز consciousness عن competence.

ما الذي لا يثبته العمل؟ / Limits

- الترجمة من brain dynamics إلى software architecture ليست برهاناً على الوعي.
- التنبؤات التجريبية نفسها موضع اختبار ومنافسة مع IIT/RPT.

قيمة 2026 / Frontier significance

مفيد كـshared computational motif: sparse global routing، working memory، blackboard، وcross-module coordination.

Research variables

GNW | workspace | ignition | broadcast

Primary source
https://pasteur.hal.science/pasteur-02546335v1/file/Neuron%20Review%20penultimate%20version%20with%20figures.pdf

موضعه داخل AGI stack

Metacognition · workspace/self-model hypotheses

العلاقة الصحيحة ليست "هذه الورقة = AGI"، بل أن هذا الجذر يحل طبقة أو قيداً محدداً داخل نظام أوسع؛ قيمته ترتفع عندما يبقى المكسب بعد ضبط compute وdata وscaffold.

26

ROOT CORPUS | Neuroscience theory | Contested theory

Challenges for Theories of Consciousness: Recurrent Processing Perspective

Victor Lamme · 2018

المساهمة العلمية / Scientific contribution

Recurrent Processing Theory تؤكد دور feedback/recurrent processing في الوعي الإدراكي، وتعارض الحاجة الدائمة إلى global access متأخر. بالنسبة للهندسة، تذكرنا بأن feed-forward competence قد لا تكفي لبناء stable perceptual state.

الآلية / Mechanism

بدلاً من pass واحد من sensory input إلى output، تنشأ recurrent loops داخل/بين المناطق، ما يسمح بتثبيت وتمييز representations. النقاش يحدد أين يجب وضع boundary بين processing والـconscious access.

TECHNICAL SIGNATURE

```
representation_{t+1}=F(input, representation_t, feedback)
```

الاستخراج الهندسي / Engineering extraction

- recurrent refinement مفيد لـperception تحت ambiguity.
- local loops يمكن أن تعمل قبل global planner.
- اختبارات ablation على recurrence قد تكشف دوره في robustness.

بروتوكول إعادة إنتاج / Ablation

حوّل النظرية إلى prediction مميز ثم preregister ablation/contrast ضد نظرية منافسة؛ لا تستخدم self-report كدليل حاسم.

Measure	Theory-specific indicators + causal ablation + uncertainty.
Failure signal	المؤشر يظهر بالتصميم الوظيفي دون أن يميز consciousness عن competence.

ما الذي لا يثبته العمل؟ / Limits

- النظرية تتحدث عن consciousness biological لا AGI مباشرة.
- وجود recurrence في network لا يثبت تجربة واعية.

قيمة 2026 / Frontier significance

كفكرة هندسية، تتقاطع مع recurrent world، iterative perception، self-correction، models، وactive sensing.

Research variables

RPT | recurrence | perceptual refinement

Primary source
https://royalsocietypublishing.org/doi/pdf/10.1098/rstb.2017.0344

موضعه داخل AGI stack

Metacognition · workspace/self-model hypotheses

العلاقة الصحيحة ليست "هذه الورقة = AGI"، بل أن هذا الجذر يحل طبقة أو قيداً محدداً داخل نظام أوسع؛ قيمته ترتفع عندما يبقى المكسب بعد ضبط compute وdata وscaffold.

27

ROOT CORPUS | Implemented cognitive hypothesis | Experimental model

The Attention Schema Theory in a Neural Network Agent

Andrew Wilterson & Michael Graziano · 2021

المساهمة العلمية / Scientific contribution

هذه الورقة ذات قيمة هندسية لأنها لا تكتفي بوصف Attention Schema Theory؛ تبني agent neural يحوي نموذجاً مبسطاً لانتباهه وتختبر هل يساعده ذلك على التحكم بالانتباه. الفكرة: self-model وظيفي قد يحسن control.

الآلية / Mechanism

agent يتعلم attention process وفي بعض الظروف يملك internal schema يصف حالة attention. يمكن مقارنة الأداء مع/بدون schema لاختبار contribution سببي.

TECHNICAL SIGNATURE

```
control = policy(state, attention_state_model)
```

الاستخراج الهندسي / Engineering extraction

- self-model يجب تقييمه عبر ablation لا عبر بلاغة introspection.
- meta-representation قد يعمل كـcontrol signal فوق attention policy.
- يفتح تصميم agents تراقب resource allocation الداخلي.

بروتوكول إعادة إنتاج / Ablation

حوّل النظرية إلى prediction مميز ثم preregister ablation/contrast ضد نظرية منافسة؛ لا تستخدم self-report كدليل حاسم.

Measure	Theory-specific indicators + causal ablation + uncertainty.
Failure signal	المؤشر يظهر بالتصميم الوظيفي دون أن يميز consciousness عن competence.

ما الذي لا يثبته العمل؟ / Limits

- النتيجة لا تثبت phenomenal consciousness.
- النموذج مبسط جداً مقارنة بالأنظمة الحديثة.

قيمة 2026 / Frontier significance

يتقاطع مع attention، confidence estimation، metacognition/ resource controllers، وagents التي تصف حدودها وقدراتها.

Research variables

Attention schema | self-model | metacognition

Primary source
https://grazianolab.princeton.edu/document/172

موضعه داخل AGI stack

Metacognition · workspace/self-model hypotheses

العلاقة الصحيحة ليست "هذه الورقة = AGI"، بل أن هذا الجذر يحل طبقة أو قيداً محدداً داخل نظام أوسع؛ قيمته ترتفع عندما يبقى المكسب بعد ضبط compute وdata وscaffold.

28

ROOT CORPUS | Active inference architecture | Theoretical/engineering bridge

Designing Explainable Artificial Intelligence with Active Inference

David Albarracin et al. · 2023

المساهمة العلمية / Scientific contribution

يقترح استخدام active inference وhierarchical generative models لبناء AI يستطيع تمثيل العوامل التي تؤثر في قراراته. الأهمية للAGI هي دمج perception وbelief updating وaction تحت نموذج generative واحد.

الآلية / Mechanism

agent يقلل expected free energy عبر تحديث beliefs واختيار policies؛ يمكن أن يمثل مستويات أعلى beliefs حول أسباب قراراته. هذا يوفر لغة موحدة للuncertainty وexploration وaction.

TECHNICAL SIGNATURE

```
π* = argmin_π G(π) (expected free energy)
```

الاستخراج الهندسي / Engineering extraction

- uncertainty يجب أن يدخل planner لا أن يبقى confidence score جانبي.
- epistemic value يمكن أن يبرر information-seeking actions.
- hierarchical generative model يتيح explanation مرتبطة بالبنية السببية المفترضة.

بروتوكول إعادة إنتاج / Ablation

حوّل النظرية إلى prediction مميز ثم preregister ablation/contrast ضد نظرية منافسة؛ لا تستخدم self-report كدليل حاسم.

Measure	Theory-specific indicators + causal ablation + uncertainty.
Failure signal	المؤشر يظهر بالتصميم الوظيفي دون أن يميز consciousness عن competence.

ما الذي لا يثبته العمل؟ / Limits

- free-energy formalism واسع وقد يصعب تحويله إلى gains عملية مقارنة بالبدائل.
- explainability الناتجة تعتمد على صلاحية generative model.

قيمة 2026 / Frontier significance

يتقاطع مع uncertainty-aware planning، agentic exploration، active perception، وself-modeling.

Research variables

Active inference | generative model | epistemic value

Primary source
https://arxiv.org/abs/2306.04025

موضعه داخل AGI stack

Metacognition · workspace/self-model hypotheses

العلاقة الصحيحة ليست "هذه الورقة = AGI"، بل أن هذا الجذر يحل طبقة أو قيداً محدداً داخل نظام أوسع؛ قيمته ترتفع عندما يبقى المكسب بعد ضبط compute وdata وscaffold.

29

ROOT CORPUS | Theory-to-indicators synthesis | Interdisciplinary framework

Consciousness in Artificial Intelligence: Insights from the Science of Consciousness

Patrick Butlin et al. · 2023

المساهمة العلمية / Scientific contribution

Butlin et al. يحول مجموعة من نظريات الوعي إلى computational indicator properties قابلة للفحص في أنظمة AI. هذه أهم مساهمة منهجية لأنها تمنع القفز من "سلوك متقدم" إلى "وعي" دون سلسلة مؤشرات محددة.

الآلية / Mechanism

يستخرج مؤشرات من AST، predictive processing، higher-order theories، GNW، RPT وغيرها، ثم يقترح فحص architecture/training/behavior بحثاً عن هذه الخصائص مع تجنب metric واحد حاسم.

TECHNICAL SIGNATURE

```
Consciousness evidence ≈ vector of theory-derived indicators, not one scalar
```

الاستخراج الهندسي / Engineering extraction

- استخدم indicator panel متعدد النظريات بدل binary consciousness claim.
- افصل architecture evidence عن behavioral imitation.
- وثق missing indicators وuncertainty بدلاً من score نهائي زائف.

بروتوكول إعادة إنتاج / Ablation

حوّل النظرية إلى prediction مميز ثم preregister ablation/contrast ضد نظرية منافسة؛ لا تستخدم self-report كدليل حاسم.

Measure	Theory-specific indicators + causal ablation + uncertainty.
Failure signal	المؤشر يظهر بالتصميم الوظيفي دون أن يميز consciousness عن competence.

ما الذي لا يثبته العمل؟ / Limits

- المؤشرات لا تشكل اختباراً حاسماً ولا consensus علمياً.
- يمكن لنظام أن يحقق property وظيفية دون phenomenology.

قيمة 2026 / Frontier significance

هو أفضل إطار عملي حالياً لربط AI بـcomputational consciousness والبحث التجريبي دون ادعاءات متضخمة.

Research variables

AI consciousness | indicators | theory mapping

Primary source
https://arxiv.org/abs/2308.08708

موضعه داخل AGI stack

Metacognition · workspace/self-model hypotheses

العلاقة الصحيحة ليست "هذه الورقة = AGI"، بل أن هذا الجذر يحل طبقة أو قيداً محدداً داخل نظام أوسع؛ قيمته ترتفع عندما يبقى المكسب بعد ضبط compute وdata وscaffold.

30

ROOT CORPUS | Adversarial collaboration | Empirical theory test

Adversarial Testing of Global Neuronal Workspace and Integrated Information Theories of Consciousness

Cogitate Consortium et al. · 2025

المساهمة العلمية / Scientific contribution

دراسة Cogitate مهمة لأنها تختبر preregistered predictions لـGNWT وIIT ضمن تعاون خصامي، بدلاً من اختيار نتائج توافق النظرية بعد التجربة. النتيجة الأعم هي أن نظريات الوعي يجب أن تدخل محكمة تجريبية صريحة.

الآلية / Mechanism

تُحدد predictions مسبقاً، تُجمع بيانات متعددة modalities، ثم تُقارن النتائج بنقاط نجاح/فشل لكل نظرية. هذا التصميم يقلل researcher degrees of freedom ويجبر النظريات على المخاطرة بالتكذيب.

TECHNICAL SIGNATURE

```
Theory value ↑ when preregistered discriminative predictions survive
```

الاستخراج الهندسي / Engineering extraction

- أي consciousness-inspired AGI architecture تحتاج preregistered discriminative tests.
- لا يكفي أن تشرح النتيجة بعد وقوعها؛ يجب توقع فرق يميزها عن المنافس.
- adversarial collaboration نموذج مفيد لتقييم architectures متنافسة.

بروتوكول إعادة إنتاج / Ablation

حوّل النظرية إلى prediction مميز ثم preregister ablation/contrast ضد نظرية منافسة؛ لا تستخدم self-report كدليل حاسم.

Measure	Theory-specific indicators + causal ablation + uncertainty.
Failure signal	المؤشر يظهر بالتصميم الوظيفي دون أن يميز consciousness عن competence.

ما الذي لا يثبته العمل؟ / Limits

- النتائج لا "تلغي" نظريات كاملة بشكل بسيط؛ التفسير معقد.
- الانتقال من البشر إلى AI يحتاج بروتوكولات جديدة.

قيمة 2026 / Frontier significance

الرسالة المنهجية: frontier AI research يجب أن يكافئ falsifiability وablation بقدر performance gains.

Research variables

Adversarial collaboration | preregistration | falsification

Primary source
https://www.nature.com/articles/s41586-025-08888-1

موضعه داخل AGI stack

Metacognition · workspace/self-model hypotheses

العلاقة الصحيحة ليست "هذه الورقة = AGI"، بل أن هذا الجذر يحل طبقة أو قيداً محدداً داخل نظام أوسع؛ قيمته ترتفع عندما يبقى المكسب بعد ضبط compute وdata وscaffold.

31

NEW FRONTIER | Reasoning-based alignment | Frontier post-training

Deliberative Alignment: Reasoning Enables Safer Language Models

Melody Y. Guan et al. · OpenAI · 2024

Scientific contribution / المساهمة العلمية

Deliberative Alignment يعلم reasoning model نص safety specification نفسه، ثم يدربه على استدعاء القواعد ذات الصلة والتفكير فيها قبل الإجابة. التحول هنا من تعلم boundary ضمنياً من labels إلى جعل specification جزءاً من المعرفة التي يُستدل فوقها.

Mechanism / الآلية

تُولد completions تتضمن reasoning يشير إلى policy text، ثم SFT لبناء prior للاستدلال الآمن، مع process/outcome supervision. وقت inference يسترجع النموذج القواعد المناسبة ويوازن compliance/refusal.

TECHNICAL SIGNATURE

```
response = Reason(prompt, retrieved/learned specification) → action
```

Engineering extraction / الاستخراج الهندسي

- specification يمكن أن تكون executable knowledge layer لا dataset labels فقط.
- reasoning budget مفيد في حالات policy borderline حيث القرار ليس pattern matching.
- يجب تقييم under-refusal وover-refusal معاً كـ Pareto frontier.

Ablation / بروتوكول إعادة إنتاج

اختبر harmful + benign distributions مع adaptive attacks، وقِس robustness والكلفة والoverrefusal معاً.

Measure	Attack success + false refusal + latency/compute + OOD generalization.
Failure signal	الدفاع ينجح على attack family معروفة لكنه ينهار على adaptive distribution shift.

Limits / ما الذي لا يثبته العمل؟

- لا يضمن أن chain-of-thought faithful أو غير قابل للتلاعب.
- الأداء يعتمد على جودة واكتمال policy text وعلى distribution التدريب.

Frontier significance / قيمة 2026

يمثل convergence بين reasoning capability وsafety architecture: نفس test-time deliberation التي ترفع الحل يمكن أن ترفع policy adherence.

Research variables

Deliberative alignment | safety specification | reasoning

Primary source
https://arxiv.org/abs/2412.16339

موضعه داخل AGI stack

Safety specification · runtime guard · oversight

العلاقة الصحيحة ليست "هذه الورقة = AGI"، بل أن هذا الجذر يحل طبقة أو قيداً محدداً داخل نظام أوسع؛ قيمته ترتفع عندما يبقى المكسب بعد ضبط compute وdata وscaffold.

32

NEW FRONTIER | Runtime safeguards | ICLR 2026

Constitutional Classifiers++: Efficient Production-Grade Defenses against Universal Jailbreaks

Hoagy Cunningham et al. · Anthropic · 2026

المساهمة العلمية / Scientific contribution

النسخة ++ تطور Constitutional Classifiers إلى دفاع production-grade ضد jailbreaks مع خفض كبير في compute overhead وrefusals. الدرس للأنظمة الوكيلية: safety ليست فقط property للنموذج، بل runtime control plane مستقل قابل للقياس.

الآلية / Mechanism

exchange classifiers تنظر إلى السياق الحواري الكامل، cascade من classifiers خفيفة ثم expensive escalations، linear probes تندمج مع external classifiers. الهدف تقليل الكلفة مع الحفاظ على robustness.

TECHNICAL SIGNATURE

```
traffic → light classifier → if suspicious → deep exchange classifier
```

الاستخراج الهندسي / Engineering extraction

- استخدم defense-in-depth: fast gate ثم deep inspection.
- راقب full exchange لا آخر output فقط، خصوصاً في multi-turn agents.
- قياس refusal rate وlatency/compute جزء من safety benchmark، لا هامش تشغيلي.

بروتوكول إعادة إنتاج / Ablation

اختبر harmful + benign distributions مع adaptive attacks، وقيّس robustness والكلفة والoverrefusal معاً.

Measure	Attack success + false refusal + latency/compute + OOD generalization.
Failure signal	الدفاع ينجح على attack family معروفة لكنه ينهار على adaptive distribution shift.

ما الذي لا يثبته العمل؟ / Limits

- classifier systems تبقى قابلة لهجمات جديدة وتحتاج red teaming مستمر.
- قد تظهر distribution shifts أو false positives في domains جديدة.

قيمة 2026 / Frontier significance

مهم لـ agentic AGI لأن autonomy الطويلة تتطلب safeguards مستقلة عن policy model، مع auditability وescalation paths.

Research variables

Runtime safety | jailbreak defense | classifier cascade

Primary source
https://arxiv.org/abs/2601.04603

موضعه داخل AGI stack

Safety specification · runtime guard · oversight

العلاقة الصحيحة ليست "هذه الورقة = AGI"، بل أن هذا الجذر يحل طبقة أو قيداً محدداً داخل نظام أوسع؛ قيمته ترتفع عندما يبقى المكسب بعد ضبط compute وdata وscaffold.

33

NEW FRONTIER | Multi-agent science | Frontier research system

Towards an AI Co-Scientist

Juraj Gottweis et al. · Google · 2025

المساهمة العلمية / Scientific contribution

يبني AI Co-Scientist multi-agent system لتوليد hypotheses والاقتراحات البحثية ثم نقدها وترتيبها وتطويرها. الأهمية ليست مجرد تلخيص الأدبيات بل تحويل test-time compute إلى structured scientific debate/evolution.

الآلية / Mechanism

generate → debate → evolve agents متخصصة تنتج hypotheses، تقارنها pairwise، تنتقد novelty/correctness، ثم tournament evolution يركز compute على المرشحين الأقوى. التنفيذ asynchronous يسمح بتوسيع compute مرناً.

TECHNICAL SIGNATURE

```
Hypotheses_{t+1}=Evolve(Debate(Generate(goal,evidence)))
```

الاستخراج الهندسي / Engineering extraction

- scientific agent يحتاج diversity mechanism لتجنب collapse إلى hypothesis واحدة.
- ranking يجب أن يوازن evidence، plausibility، novelty، وtestability.
- human scientist يبقى goal setter وvalidator، لا مجرد consumer.

بروتوكول إعادة إنتاج / Ablation

افصل idea quality عن experiment/reproduction؛ اطلب artifacts قابلة للتنفيذ وexternal validation حين يمكن.

Measure	Novelty + correctness + reproducibility + verified gain + resource cost.
Failure signal	judge-model bias أو evaluator gaming أو paper quality بلا scientific truth.

ما الذي لا يثبته العمل؟ / Limits

- التحقق التجريبي الخارجي بطيء ومجالي؛ النجاح في biomedical cases لا يساوي general science automation.
- LLM judges قد تشارك biases مع generators.

قيمة 2026 / Frontier significance

يدفع agentic AI نحو knowledge production: ليس فقط استرجاع الموجود، بل اقتراح أسئلة وتجارب قابلة للاختبار تحت رقابة بشرية.

Research variables

AI scientist | multi-agent debate | test-time scaling

Primary source
https://arxiv.org/abs/2502.18864

موضعه داخل AGI stack

Research orchestration · experiment · verification

العلاقة الصحيحة ليست "هذه الورقة = AGI"، بل أن هذا الجذر يحل طبقة أو قيداً محدداً داخل نظام أوسع؛ قيمته ترتفع عندما يبقى المكسب بعد ضبط compute وdata وscaffold.

34

NEW FRONTIER | Evolutionary coding agent | Frontier discovery system

AlphaEvolve: A Coding Agent for Scientific and Algorithmic Discovery

Alexander Novikov et al. · Google DeepMind · 2025

المساهمة العلمية / Scientific contribution

AlphaEvolve يجمع LLM generation مع automated evaluators داخل evolutionary loop لاكتشاف وتحسين algorithms/code. القوة تأتي من مجال تكون فيه quality قابلة للقياس آلياً، ما يحول النموذج من proposer إلى search engine على برامج قابلة للتنفيذ.

الآلية / Mechanism

ensemble LLMs يقترح code mutations؛ evaluators تنفذ وتقيس candidate programs؛ database/ evolutionary selection يحتفظ بالأفضل ويعيده كـcontext لجولات لاحقة. يمكن استخدام عدة metrics في آن واحد.

TECHNICAL SIGNATURE

```
P_{t+1}=Select(Evaluate(Mutate_LLM(P_t)))
```

الاستخراج الهندسي / Engineering extraction

- عندما يوجد verifier قوي، search over code قد يتفوق على single-shot reasoning.
- program database يعمل كـpopulation memory للابتكارات.
- objective design حاسم: ما لا يقيسه evaluator لن يتحسن بالضرورة.

بروتوكول إعادة إنتاج / Ablation

افصل idea quality عن experiment/reproduction؛ اطلب artifacts قابلة للتنفيذ وexternal validation حين يمكن.

Measure	Novelty + correctness + reproducibility + verified gain + resource cost.
Failure signal	judge-model bias أو evaluator gaming أو paper quality بلا scientific truth.

ما الذي لا يثبته العمل؟ / Limits

- لا يعمل بنفس القوة في مسائل لا تملك feedback آلياً سريعاً.
- قد يحدث evaluator exploitation أو over-optimization لمقياس ضيق.

قيمة 2026 / Frontier significance

يمثل مساراً نحو recursive tool improvement وAI-assisted R&D، مع ضرورة governance لأن النظام يستطيع تحسين أجزاء من computational stack نفسه.

Research variables

Evolutionary search | code agents | automated evaluator | discovery

Primary source
https://arxiv.org/abs/2506.13131

موضعه داخل AGI stack

Research orchestration · experiment · verification

العلاقة الصحيحة ليست "هذه الورقة = AGI"، بل أن هذا الجذر يحل طبقة أو قيداً محدداً داخل نظام أوسع؛ قيمته ترتفع عندما يبقى المكسب بعد ضبط compute وdata وscaffold.

35

NEW FRONTIER | World model | Frontier physical reasoning

V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning

Mido Assran et al. · Meta · 2025

المساهمة العلمية / Scientific contribution

V-JEPA 2 يبني world model من video self-supervision ثم يضيف قدراً صغيراً من robot interaction data، ويظهر understanding/prediction مع zero-shot planning نسبي في بيئات جديدة. يمثل بديلاً عن generative pixel prediction الكامل.

الآلية / Mechanism

encoder يتعلم embeddings من فيديو واسع؛ predictor يتنبأ representations المستقبلية/المخفية؛ action-conditioned stage يربط actions بتطور latent state. التخطيط يقارن outcomes المتخيلة بهدف visual.

TECHNICAL SIGNATURE

```
z_{t+1}=Predict(z_t,a_t); a*=argmin distance(z_goal,z_{t+1:k})
```

الاستخراج الهندسي / Engineering extraction

- predict in representation space يمكن أن يكون أكفأ من pixel generation.
- internet video يمكن أن يبني prior فيزيائي ثم interaction data تضبط action model.
- model-predictive control يعيد التخطيط بعد كل خطوة لخفض error accumulation.

بروتوكول إعادة إنتاج / Ablation

اختبر novel objects/environments مع perturbations؛ قارن reactive policy مع planning/world-model under same interaction budget.

Measure	Success + prediction error + replans + safety violations + latency.
Failure signal	generalization ظاهرية سببها تشابه training data أو visual shortcuts.

ما الذي لا يثبته العمل؟ / Limits

- المهام الروبوتية ما تزال محدودة والأهداف visual قصيرة نسبياً.
- latent prediction لا يضمن causal abstraction عامة.

قيمة 2026 / Frontier significance

يجسد المسار world model → planning → action، وهو أحد أهم البدائل عن LLM-only AGI.

Research variables

World model | JEPA | physical reasoning | planning

Primary source
https://arxiv.org/abs/2506.09985

موضعه داخل AGI stack

Physical representation · prediction · planning

العلاقة الصحيحة ليست "هذه الورقة = AGI"، بل أن هذا الجذر يحل طبقة أو قيداً محدداً داخل نظام أوسع؛ قيمته ترتفع عندما يبقى المكسب بعد ضبط compute وdata وscaffold.

36

NEW FRONTIER | Vision-Language-Action + embodied reasoning | Frontier robotics report

Gemini Robotics 1.5: Generalist Robots with Advanced Embodied Reasoning, Thinking, and Motion Transfer

Gemini Robotics Team · Google DeepMind · 2025

المساهمة العلمية / Scientific contribution

Gemini Robotics 1.5 يدمج Vision-Language-Action مع embodied reasoning وmulti-level internal reasoning، ويقدم Motion Transfer للتعلم من بيانات روبوتية متعددة الأجسام. هذا يدفع agentic reasoning من الشاشة إلى control motor مباشر.

الآلية / Mechanism

VLA يحول perception+instruction إلى actions؛ ER model يتولى planning، spatial understanding، progress estimation، وtool calls. Motion Transfer يحاول مشاركة المعرفة بين embodiments دون action-space alignment كامل.

TECHNICAL SIGNATURE

```
perceive → reason/plan → VLA action → sense progress → replan
```

الاستخراج الهندسي / Engineering extraction

- افصل high-level embodied reasoning عن low-level control حيث يلزم.
- progress estimation يجب أن يدخل closed-loop replanning.
- cross-embodiment transfer شرط للتوسع لأن جمع بيانات لكل robot مكلف.

بروتوكول إعادة إنتاج / Ablation

	اختبر novel objects/environments مع perturbations؛ قارن reactive policy مع planning/world-model under same interaction budget.
Measure	Success + prediction error + replans + safety violations + latency.
Failure signal	generalization ظاهرية سببها تشابه training data أو visual shortcuts.

ما الذي لا يثبته العمل؟ / Limits

- latency، physical safety، وdistribution shift أصعب بكثير من digital agents.
- التقرير لا يثبت general-purpose household autonomy الشاملة.

قيمة 2026 / Frontier significance

يجعل "agent" نظاماً perception-thought-action متصلاً بالعالم الفيزيائي، ويكشف أن near-AGI engineering يحتاج embodiment وworld interaction لا لغة فقط.

Research variables

Embodied reasoning | VLA | motion transfer | robotics

Primary source
https://storage.googleapis.com/deepmind-media/gemini-robotics/Gemini-Robotics-1-5-Tech-Report.pdf

موضعه داخل AGI stack

Perception · embodied reasoning · action control

العلاقة الصحيحة ليست "هذه الورقة = AGI"، بل أن هذا الجذر يحل طبقة أو قيداً محدداً داخل نظام أوسع؛ قيمته ترتفع عندما يبقى المكسب بعد ضبط compute وdata وscaffold.

37

NEW FRONTIER | Fluid intelligence benchmark | Frontier AGI benchmark

ARC-AGI-2: A New Challenge for Frontier AI Reasoning Systems

François Chollet et al. · 2025

المساهمة العلمية / Scientific contribution

ARC-AGI-2 يرفع صعوبة Abstraction and Reasoning Corpus لقياس few-shot generalization على tasks جديدة قليلة المعرفة المسبقة. الهدف هو مقاومة benchmark saturation والتمييز بين knowledge coverage وfluid adaptation.

الآلية / Mechanism

كل task يقدم input-output examples محدودة ويتطلب استنتاج transformation مجردة ثم تطبيقها على test grid. tasks مصممة لتكون قابلة للبشر لكن صعبة على frontier systems.

TECHNICAL SIGNATURE

```
few examples → infer latent rule → apply to unseen grid
```

الاستخراج الهندسي / Engineering extraction

- generalization eval يجب أن يحوي tasks غير مألوفة مع minimal prior dependence.
- قياس compute/cost مهم لأن brute-force search قد يشتري score دون كفاءة معرفية.
- contamination controls أساسية لأن benchmark الشهيرة تتحول إلى training target.

بروتوكول إعادة إنتاج / Ablation

ثبّت contamination policy وcompute budget، واستخدم private/held-out tasks وconfidence intervals.

Measure	Accuracy + calibration + cost + robustness to novelty.
Failure signal	score يتحسن بتغطية المعرفة أو brute-force compute دون fluid generalization.

ما الذي لا يثبته العمل؟ / Limits

- ARC domain بصري رمزي ضيق ولا يغطي tool use أو real-world agency.
- أي benchmark ثابت قابل للتشبع بمرور الوقت.

قيمة 2026 / Frontier significance

يوفر عدسة قوية لفصل memorized competence عن on-the-fly abstraction، وهو لب أي claim عن general intelligence.

Research variables

ARC-AGI | fluid intelligence | abstraction | few-shot

Primary source
https://arxiv.org/abs/2505.11831

موضعه داخل AGI stack

Frontier evaluation · novelty · calibration

العلاقة الصحيحة ليست "هذه الورقة = AGI"، بل أن هذا الجذر يحل طبقة أو قيداً محدداً داخل نظام أوسع؛ قيمته ترتفع عندما يبقى المكسب بعد ضبط compute وdata وscaffold.

38

NEW FRONTIER | Broad expert benchmark | Frontier benchmark

Humanity's Last Exam

Long Phan et al. · 2025/2026 rev.

المساهمة العلمية / Scientific contribution

HLE يجمع آلاف الأسئلة الصعبة متعددة التخصصات عند حدود المعرفة الأكاديمية، صممت لتجنب تشبع MMLU-like benchmarks. يقيس breadth + expert reasoning مع automated grading نسبي.

الآلية / Mechanism

2,500+ questions من خبراء عبر علوم ورياضيات وإنسانيات مع multiple-choice وshort-answer، وبعض multimodal items. التركيز على مسائل ذات حل معروف لكن لا يكفي لها retrieval سطحي سريع.

TECHNICAL SIGNATURE

```
Capability = accuracy + calibration across expert domains
```

الاستخراج الهندسي / Engineering extraction

- frontier eval يجب أن يتحرك كلما تشبعت الاختبارات القديمة.
- calibration مهم بقدر accuracy لأن النظام قد يجيب بثقة في domains لا يمتلكها.
- broad expert coverage يختبر transfer بين المجالات لكنه لا يقيس autonomy.

بروتوكول إعادة إنتاج / Ablation

ثبّت contamination policy وcompute budget، واستخدم private/held-out tasks وconfidence intervals.

Measure	Accuracy + calibration + cost + robustness to novelty.
Failure signal	score يتحسن بتغطية المعرفة أو brute-force compute دون fluid generalization.

ما الذي لا يثبته العمل؟ / Limits

- closed-ended exam لا يساوي real scientific work.
- قد يحدث leakage مع انتشار الأسئلة، لذا يلزم تحديث مستمر.

قيمة 2026 / Frontier significance

يحدد سقفاً أصعب للknowledge+reasoning قبل الانتقال إلى benchmarks agentic مفتوحة مثل PaperBench وRE-Bench.

Research variables

HLE | expert knowledge | calibration | multimodal

Primary source
https://arxiv.org/abs/2501.14249

موضعه داخل AGI stack

Frontier evaluation · novelty · calibration

العلاقة الصحيحة ليست "هذه الورقة = AGI"، بل أن هذا الجذر يحل طبقة أو قيداً محدداً داخل نظام أوسع؛ قيمته ترتفع عندما يبقى المكسب بعد ضبط compute وdata وscaffold.

39

NEW FRONTIER | Expert mathematics benchmark | Frontier reasoning eval

FrontierMath: A Benchmark for Evaluating Advanced Mathematical Reasoning in AI

Elliot Glazer et al. · 2024

المساهمة العلمية / Scientific contribution

FrontierMath يقدم مسائل أصلية شديدة الصعوبة صاغها خبراء في فروع رياضية متعددة، مع answers قابلة للتحقق وتقليل contamination. الهدف هو قياس reasoning عند مستوى باحث، لا مسابقات مدرسية متشبعة.

الآلية / Mechanism

المسائل غير منشورة سابقاً وتتطلب ساعات أو أياماً من العمل البشري المتخصص. التحقق الآلي/الخبير يسمح بقياس نتيجة دقيقة رغم عمق البرهان أو الحساب.

TECHNICAL SIGNATURE

```
novel expert problem → long-horizon reasoning → verifiable mathematical

outcome
```

الاستخراج الهندسي / Engineering extraction

- fresh problem creation عنصر أمني للbenchmark.
- يجب فصل final answer عن proof quality عندما يكون البرهان هو القيمة.
- يمكن استخدام المسائل كبيئة لtest-time search + theorem/proof tools.

بروتوكول إعادة إنتاج / Ablation

ثبّت contamination policy وcompute budget، واستخدم private/held-out tasks وconfidence intervals.

Measure	Accuracy + calibration + cost + robustness to novelty.
Failure signal	score يتحسن بتغطية المعرفة أو brute-force compute دون fluid generalization.

ما الذي لا يثبته العمل؟ / Limits

- الرياضيات مجال verification-friendly لا يمثل كل cognition.
- الscore قد يقفز مع specialized tools دون أن يعكس breadth العامة.

قيمة 2026 / Frontier significance

يختبر الحد الأقصى للreasoning المجرد ويعمل كجسر نحو automated theorem discovery وAlphaEvolve-style search.

Research variables

FrontierMath | expert reasoning | contamination resistance

Primary source
https://arxiv.org/abs/2411.04872

موضعه داخل AGI stack

Frontier evaluation · novelty · calibration

العلاقة الصحيحة ليست "هذه الورقة = AGI"، بل أن هذا الجذر يحل طبقة أو قيداً محدداً داخل نظام أوسع؛ قيمته ترتفع عندما يبقى المكسب بعد ضبط compute وdata وscaffold.

40

NEW FRONTIER | Economic agent benchmark | Real-world software eval

SWE-Lancer: Can Frontier LLMs Earn $1 Million from Real-World Freelance Software Engineering?

Samuel Miserendino et al. · OpenAI · 2025

المساهمة العلمية / Scientific contribution

SWE-Lancer يقيس قدرة frontier models على مهام freelance حقيقية من Upwork، بما فيها تنفيذ هندسي واتخاذ قرارات managerial، ويربط الأداء بقيمة مالية حقيقية. هذا يقرب evaluation من productivity الاقتصادية.

الآلية / Mechanism

أكثر من 1,400 task تاريخياً بقيمة إجمالية مليون دولار، engineering tasks تقيم بـend-to-end tests، وmanagerial tasks تقارن بقرارات المدراء الأصليين. بيئة Docker موحدة للreproducibility.

TECHNICAL SIGNATURE

```
Agent value ≈ Σ solved_task_market_value under verified execution
```

الاستخراج الهندسي / Engineering extraction

- اربط capability بمخرجات اقتصادية لا token benchmarks فقط.
- managerial judgment يحتاج تقييم منفصل عن code generation.
- offline reproducibility يقلل variability الناتجة عن الإنترنت.

بروتوكول إعادة إنتاج / Ablation

أعد بناء baseline مضبوط، ثم نفذ ablation للآلية المركزية مع عدة seeds.

Measure	Effect size + cost + reliability.
Failure signal	النتيجة لا تصمد أمام baseline أو distribution shift.

ما الذي لا يثبته العمل؟ / Limits

- Upwork distribution لا يمثل كل software engineering.
- القيمة المالية التاريخية ليست دالة مباشرة على قيمة حل AI الحالي.

قيمة 2026 / Frontier significance

يقيس انتقال agent من حل issue إلى أداء work unit لا cost/value، وهو مهم لتقدير automation impact.

Research variables

Software agents | economic value | end-to-end tests

Primary source
https://arxiv.org/abs/2502.12115

موضعه داخل AGI stack

Coding agent · economic task · execution verification

العلاقة الصحيحة ليست "هذه الورقة = AGI"، بل أن هذا الجذر يحل طبقة أو قيداً محدداً داخل نظام أوسع؛ قيمته ترتفع عندما يبقى المكسب بعد ضبط compute وdata وscaffold.

41

NEW FRONTIER | Web research benchmark | Agentic browsing eval

BrowseComp: A Simple Yet Challenging Benchmark for Browsing Agents

Jason Wei et al. · OpenAI · 2025

المساهمة العلمية / Scientific contribution

BrowseComp يقيس قدرة agents على إيجاد معلومات صعبة ومتشابكة عبر browsing طويل ومبدع، مع answers قصيرة سهلة التحقق. يميز بين "وجود browser tool" و"القدرة على البحث الاستراتيجي".

الآلية / Mechanism

1,266 questions مصممة بحيث يتطلب الحل التنقل عبر عدد كبير من المواقع أو الاستدلال حول أين توجد المعلومة. final answer قصير، ما يجعل grading بسيطاً ويزيد asymmetry بين difficulty of finding وease of verification.

TECHNICAL SIGNATURE

```
question → adaptive search tree → evidence set → short verified answer
```

الاستخراج الهندسي / Engineering extraction

- search policy يجب أن يدير source quality، query reformulation، وstopping.
- tool access وحده لا يكفي؛ planning information acquisition هو القدرة الحاسمة.
- احتفظ evidence trail حتى يمكن audit النتيجة.

بروتوكول إعادة إنتاج / Ablation

أعد بناء baseline مضبوط، ثم نفذ ablation للآلية المركزية مع عدة seeds.

Measure	Effect size + cost + reliability.
Failure signal	النتيجة لا تصمد أمام baseline أو distribution shift.

ما الذي لا يثبته العمل؟ / Limits

- short-answer format لا يمثل كل deep-research deliverables.
- الويب ديناميكي وقد تختفي المصادر أو تتغير.

قيمة 2026 / Frontier significance

هو benchmark مهم لـ research agents التي تجمع المعرفة الخارجية بدلاً من الاعتماد على model memory.

Research variables

Browsing agents | deep research | persistence | evidence

Primary source
https://arxiv.org/abs/2504.12516

موضعه داخل AGI stack

Information acquisition · browsing · evidence synthesis

العلاقة الصحيحة ليست "هذه الورقة = AGI"، بل أن هذا الجذر يحل طبقة أو قيداً محدداً داخل نظام أوسع؛ قيمته ترتفع عندما يبقى المكسب بعد ضبط compute وdata وscaffold.

42

NEW FRONTIER | Research replication benchmark | ICML 2025 / frontier eval

PaperBench: Evaluating AI's Ability to Replicate AI Research

Giulio Starace et al. · OpenAI · 2025

المساهمة العلمية / Scientific contribution

PaperBench يقيس قدرة agents على إعادة تنفيذ مساهمات أوراق ICML من الصفر: فهم الورقة، بناء codebase، تشغيل experiments، troubleshooting، ثم grading تفصيلي. هذا أقرب إلى AI research engineering الحقيقي من benchmark سؤال-جواب.

الآلية / Mechanism

20 أوراق ICML 2024 مع 8,316 rubric items صاغت بالتعاون مع مؤلفين؛ التنفيذ يتم داخل containers ثم إعادة تشغيل submission في بيئة نظيفة؛ LLM judge مدروس عبر JudgeEval يخفف تكلفة grading.

TECHNICAL SIGNATURE

```
paper → implement → run → reproduce → rubric-grade
```

الاستخراج الهندسي / Engineering extraction

- decompose long-horizon research task إلى rubric hierarchy لقياس partial progress.
- fresh execution environment يمنع accidental success.
- judge model نفسه يحتاج benchmark مستقل لأن evaluation قد يصبح bottleneck.

بروتوكول إعادة إنتاج / Ablation

	ثبّت contamination policy وcompute budget، واستخدم private/held-out tasks وconfidence intervals.
Measure	Accuracy + calibration + cost + robustness to novelty.
Failure signal	score يتحسن بتغطية المعرفة أو brute-force compute دون fluid generalization.

ما الذي لا يثبته العمل؟ / Limits

- 20 papers عينة صغيرة ومجالها ML.
- automated judge قد يخطئ أو يتأثر بأسلوب submission.

قيمة 2026 / Frontier significance

مع RE-Bench يشكل مقياساً مباشراً لقدرات AI التي قد تسرع AI R&D نفسه، وهي من أقرب المسارات إلى recursive capability growth.

Research variables

Research agents | replication | long horizon | judge eval

Primary source
https://arxiv.org/abs/2504.01848

موضعه داخل AGI stack

Frontier evaluation · novelty · calibration

العلاقة الصحيحة ليست "هذه الورقة = AGI"، بل أن هذا الجذر يحل طبقة أو قيداً محدداً داخل نظام أوسع؛ قيمته ترتفع عندما يبقى المكسب بعد ضبط compute وdata وscaffold.

43

NEW FRONTIER | Open-ended R&D benchmark | ICML 2025

RE-Bench: Evaluating Frontier AI R&D Capabilities of Language Model Agents against Human Experts

Hjalmar Wijk et al. · METR · 2024/ICML 2025

المساهمة العلمية / Scientific contribution

RE-Bench يضع agents وخبراء بشر أمام 7 environments بحث هندسي ML مفتوحة مع budgets زمنية مختلفة. النتيجة الأساسية ليست "AI أفضل" أو "الإنسان أفضل" مطلقاً؛ بل أن منحنيات الأداء تختلف مع الزمن.

الآلية / Mechanism

71 محاولة بشرية من 61 خبيراً، ومقارنات best-of-k لـagents ضمن 2h/8h/32h budgets. agents سريعة في iteration القصير بينما البشر يحققون returns أفضل مع أفق أطول في النتائج المنشورة.

TECHNICAL SIGNATURE

```
Performance = f(time budget, agent scaffold, model, retries)
```

الاستخراج الهندسي / Engineering extraction

- قيّم capability كدالة في wall-clock/compute budget لا score وحيد.
- long-horizon degradation قد يكون أهم من peak short-horizon skill.
- trajectory logging يسمح بتحليل أين يضيع الوقت: debugging، coding، ideation، أو evaluation.

بروتوكول إعادة إنتاج / Ablation

ثبّت contamination policy وcompute budget، واستخدم private/held-out tasks وconfidence intervals.

Measure	Accuracy + calibration + cost + robustness to novelty.
Failure signal	score يتحسن بتغطية المعرفة أو brute-force compute دون fluid generalization.

ما الذي لا يثبته العمل؟ / Limits

- سبعة environments لا تغطي كل R&D.
- best-of-k يرفع التكلفة ويجب إظهاره عند المقارنة.

قيمة 2026 / Frontier significance

مقياس محوري لبحث autonomy لأنه يقيس قدرة agents على تنفيذ research engineering متكرر وليس فقط الإجابة على مسائل.

Research variables

AI R&D | autonomy | time horizon | human comparison

Primary source
https://arxiv.org/abs/2411.15114

موضعه داخل AGI stack

Frontier evaluation · novelty · calibration

العلاقة الصحيحة ليست "هذه الورقة = AGI"، بل أن هذا الجذر يحل طبقة أو قيداً محدداً داخل نظام أوسع؛ قيمته ترتفع عندما يبقى المكسب بعد ضبط compute وdata وscaffold.

44

NEW FRONTIER | End-to-end autonomous science | Frontier agentic system

The AI Scientist-v2: Workshop-Level Automated Scientific Discovery via Agentic Tree Search

Yutaro Yamada et al. · Sakana AI · 2025

المساهمة العلمية / Scientific contribution

AI Scientist-v2 يدفع الأتمتة من idea generation إلى دورة علمية شبه كاملة: hypothesis، experiment design، code execution، analysis، figures، paper writing، والمراجعة. أبرز إنجازه إنتاج ورقة كاملة قُبلت في workshop ضمن تجربة محددة.

الآلية / Mechanism

progressive agentic tree search يديره experiment manager؛ لا يعتمد على code templates بشرية كما في النسخة الأولى؛ VLM feedback loop يراجع figures والمظهر؛ reviewer agent يضيف cycle نقدية.

TECHNICAL SIGNATURE

```
idea → tree-search experiments → analysis → manuscript → review loop
```

الاستخراج الهندسي / Engineering extraction

- autonomous science يحتاج experiment manager يدير branches والموارد.
- acceptance ليست proof of scientific truth؛ يجب فصل peer-review outcome عن replication/novelty.
- artifact chain يجب أن تكون reproducible من code/data/logs.

بروتوكول إعادة إنتاج / Ablation

افصل idea quality عن experiment/reproduction؛ اطلب artifacts قابلة للتنفيذ وexternal validation حين يمكن.

Measure	Novelty + correctness + reproducibility + verified gain + resource cost.
Failure signal	judge-model bias أو evaluator gaming أو paper quality بلا scientific truth.

ما الذي لا يثبته العمل؟ / Limits

- التجربة على عدد محدود من submissions ومجال ML.
- نظام يمكن أن ينتج paper مقبول دون اكتشاف مهم أو قابلية تعميم.

قيمة 2026 / Frontier significance

يوضح أن agentic stack يمكن أن يغطي كامل research workflow، ويجعل evaluation scientific quality + reproducibility أكثر إلحاحاً.

Research variables

AI Scientist | tree search | experiment manager | autonomous research

Primary source
https://arxiv.org/abs/2504.08066

موضعه داخل AGI stack

Research orchestration · experiment · verification

العلاقة الصحيحة ليست "هذه الورقة = AGI"، بل أن هذا الجذر يحل طبقة أو قيداً محدداً داخل نظام أوسع؛ قيمته ترتفع عندما يبقى المكسب بعد ضبط compute وdata وscaffold.

45

NEW FRONTIER | Scientific deep-research training | 2026 frontier paper

SciResearcher: Scaling Deep Research Agents for Frontier Scientific Reasoning

Tianshi Zheng et al. · 2026

المساهمة العلمية / Scientific contribution

SciResearcher يركز على فجوة صعبة: frontier science حيث المعرفة موزعة في مصادر أكاديمية متخصصة ويحتاج الحل retrieval + computation + long-horizon reasoning. يقترح framework آلياً لبناء بيانات تدريب لهذه البيئة.

الآلية / Mechanism

يولد conceptual/computational tasks grounded in academic evidence، ويجمع information acquisition وtool-integrated reasoning؛ ثم يستخدم supervised fine-tuning وagentic RL لتدريب SciResearcher-8B.

TECHNICAL SIGNATURE

```
academic evidence → task synthesis → SFT + agentic RL → research trajectory
```

الاستخراج الهندسي / Engineering extraction

- scientific agent training يحتاج data construction تعكس البحث الحقيقي لا QA عام.
- retrieval يجب أن يتكامل مع computation والأدوات في trajectory واحدة.
- يمكن استخدام smaller foundation agent إذا كان post-training متخصصاً في research process.

بروتوكول إعادة إنتاج / Ablation

افصل idea quality عن experiment/reproduction؛ اطلب artifacts قابلة للتنفيذ وexternal validation حين يمكن.

Measure	Novelty + correctness + reproducibility + verified gain + resource cost.
Failure signal	judge-model bias أو evaluator gaming أو paper quality بلا scientific truth.

ما الذي لا يثبته العمل؟ / Limits

- النتائج ما تزال benchmark-centric ومجالات محددة.
- البيانات المولدة آلياً تحتاج quality controls لتجنب synthetic bias.

قيمة 2026 / Frontier significance

يمثل shift 2026 من "model that knows science" إلى "agent trained to do scientific information acquisition and reasoning as a process".

Research variables

Scientific agents | deep research | agentic RL | tools

Primary source
https://arxiv.org/abs/2605.01489

موضعه داخل AGI stack

Frontier evaluation · novelty · calibration

العلاقة الصحيحة ليست "هذه الورقة = AGI"، بل أن هذا الجذر يحل طبقة أو قيداً محدداً داخل نظام أوسع؛ قيمته ترتفع عندما يبقى المكسب بعد ضبط compute وdata وscaffold.

ما الذي تتفق عليه الجبهة فعلياً؟ — 12 Engineering Invariants

عبر الأوراق الـ45 تظهر بنية أقوى من أي paper منفرد: التقدم نحو الأنظمة العامة ينتقل من تحسين predictor إلى بناء closed-loop intelligent system لها compute policy، state، evaluator، memory، tools، وworld feedback.

1. Intelligence ≠ model size.
القدرة الفعلية هي تركيب model + search + tools + memory + environment. Transformer/scaling يبني substrate؛ ReAct/MemGPT/AutoGen/Robotics تبني system.

2. Test-time compute أصبح محوراً مستقلاً.
CoT → ToT → DeepSeek-R1 → Co-Scientist/AlphaEvolve: مزيد من inference لا يفيد إلا إذا وُجّه بـsearch أو verifier أو evaluator.

3. Verifier quality تحدد سقف self-improvement.
AlphaZero/MuZero/AlphaEvolve/FrontierMath تؤكد أن feedback الموثوق يتيح loops قوية؛ في open-world domains يصبح evaluator bottleneck.

4. Memory تحتاج سياسة.
RAG يضيف knowledge، Reflexion يضيف episode lessons، MemGPT يضيف hierarchy؛ لكن read/write/compaction policy هي الفرق بين ذاكرة وركام.

5. Agency هي state transition موثق.
ReAct وOSWorld وGemini Robotics توضح أن الوكيل يجب أن يلاحظ النتيجة بعد الفعل ويعيد التخطيط؛ بلا tool call post-action verification ليس agency موثوقة.

6. World models تعيد الاعتبار للتنبؤ.
MuZero وV-JEPA 2 وrobotics تشير إلى أن النموذج العام يحتاج تصوراً عن consequence، لا مجرد next-token continuation.

7. Long horizon مشكلة مستقلة.
SWE-bench/PaperBench/RE-Bench تكشف أن نجاح micro-task لا يضمن إدارة مشروع أو بحث لساعات؛ checkpointing، recovery، وresource scheduling عناصر capability.

8. Generalization تحتاج novelty controls.
ARC-AGI-2، FrontierMath، HLE تصمم fresh/novel tasks لأن benchmark saturation وcontamination يمكن أن توهما بالتقدم.

9. Multi-agent ليس مكسباً مجانياً.
AutoGen وCo-Scientist تبينان فائدة specialization/debate، لكن communication وoverhead وcorrelated errors يفرضان single-agent baselines وablation.

10. Safety تنتقل إلى runtime architecture.
RLHF وDeliberative Alignment وConstitutional Classifiers++ توضح أن alignment طبقات: reasoning over policy، training، وindependent runtime guards.

11. Scientific autonomy يجب أن تقاس بالartifact chain.
PaperBench/AI Scientist-v2/SciResearcher: hypothesis جميلة لا تكفي؛ نحتاج code، reproduction، evidence، experiment، وreviewable logs.

12. Consciousness لا تختزل في capability.
IIT/GNW/RPT/AST/Butlin/Cogitate تفرض الفصل بين functional indicators وphenomenal claims، مع preregistration وadversarial falsification.

Near-AGI engineering = Generalization × World Grounding × Deliberation × Tool/Memory Control × Long-Horizon Reliability × Verifiability × Safe Autonomy

Reference Architecture المستخرجة من المجتمع العلمي

L0 Objective / specification
AIXI · Universal Intelligence · Alignment specs

L1 Perception / representation
Transformer · CLIP · V-JEPA 2

L2 World model / state
MuZero · V-JEPA 2 · embodied state estimation

L3 Deliberation / search
CoT · ToT · R1 · MCTS · evolutionary search

L4 Knowledge / memory
RAG · Reflexion · MemGPT

L5 Tools / action
Toolformer · ReAct · APIs · code · browser · computer

L6 Agent orchestration
AutoGen · Co-Scientist · AI Scientist-v2

L7 Environment feedback
OSWorld · Robotics · execution tests

L8 Verification / evaluation
SWE-bench · FrontierMath · PaperBench · RE-Bench

L9 Safety / governance
RLHF · Deliberative Alignment · Constitutional Classifiers++

L10 Learning / evolution
AlphaZero · Voyager · AlphaEvolve · agentic RL

L11 Metacognition / consciousness research
Consciousness Prior · GNW · AST · Butlin indicators

الفجوات التي ما تزال تمنع claim قوي عن AGI/ASI

Persistent identity/state
long-term memory ما تزال معرضة للتلوث، الاختزال الخاطئ، وdrift عبر آلاف التفاعلات.

Open-world reward
لا نملك verifier موضوعياً لمعظم الأهداف البشرية المركبة؛ لذلك self-improvement أسهل في code/math من الاقتصاد أو العلم المفتوح.

Horizon reliability
الأخطاء الصغيرة تتراكم؛ checkpointing، rollback، recovery، وresource governance تحتاج قياسات موحدة.

Causal world models
الـLLMs/world models الحالية قوية تنبؤياً لكن causal abstractions العامة والتدخلية ما تزال غير محسومة.

Embodied safety
في العالم الفيزيائي، latency وuncertainty وirreversible actions تجعل safety requirements أعلى من digital agents.

Evaluator independence
إذا كان generator والjudge من نفس distribution قد يتشاركان blind spots؛ نحتاج heterogeneous verification.

Scaling oversight
كلما فاق agent قدرة المشرف في مجال معين، يصبح human approval وحده غير كافٍ؛ نحتاج scalable oversight وaudit layers.

Generalization vs contamination
مع كل benchmark مشهور يزداد خطر training leakage؛ dynamic/private evals ضرورة.

Consciousness uncertainty
لا يوجد scientific consensus يسمح بتحويل consciousness probability إلى capability score حاسمة.

Scientific truth vs paper production
autonomous science يجب أن يُقاس بالreplication والتجربة الخارجية، لا acceptance أو جودة النص فقط.

State of the art 2026: أقرب صورة للـAGI ليست "نموذجاً عملاقاً واحداً"، بل نظاماً مركباً يوزع computation بين agents, world models, tools, memory, search, model, evaluators, and safety controls. التحدي الأكبر تحول من raw capability إلى reliable autonomous closure of long-horizon tasks.

Primary Research Roots 45 — المراجع الأساسية

01. Marcus Hutter (2007). *Universal Algorithmic Intelligence / AIXI*.
https://arxiv.org/abs/cs/0701125

02. Shane Legg & Marcus Hutter (2007). *Universal Intelligence: A Definition of Machine Intelligence*.
https://arxiv.org/abs/0712.3329

03. David Silver et al. (2017). *Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm (AlphaZero)*.
https://arxiv.org/abs/1712.01815

04. Julian Schrittwieser et al. (2019). *Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model (MuZero)*.
https://arxiv.org/abs/1911.08265

05. Ashish Vaswani et al. (2017). *Attention Is All You Need*.
https://arxiv.org/abs/1706.03762

06. Jared Kaplan et al. (2020). *Scaling Laws for Neural Language Models*.
https://arxiv.org/abs/2001.08361

07. Tom Brown et al. (2020). *Language Models are Few-Shot Learners (GPT-3)*.
https://arxiv.org/abs/2005.14165

08. Alec Radford et al. (2021). *Learning Transferable Visual Models From Natural Language Supervision (CLIP)*.
https://arxiv.org/abs/2103.00020

09. Jordan Hoffmann et al. (2022). *Training Compute-Optimal Large Language Models (Chinchilla)*.
https://arxiv.org/abs/2203.15556

10. Patrick Lewis et al. (2020). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks (RAG)*.
https://arxiv.org/abs/2005.11401

11. Long Ouyang et al. (2022). *Training Language Models to Follow Instructions with Human Feedback (InstructGPT)*.
https://arxiv.org/abs/2203.02155

12. Jason Wei et al. (2022). *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*.
https://arxiv.org/abs/2201.11903

13. DeepSeek-AI (2025). *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning*.
https://arxiv.org/abs/2501.12948

14. Timo Schick et al. (2023). *Toolformer: Language Models Can Teach Themselves to Use Tools*.
https://arxiv.org/abs/2302.04761

15. Shunyu Yao et al. (2022/2023). *ReAct: Synergizing Reasoning and Acting in Language Models*.
https://arxiv.org/abs/2210.03629

16. Shunyu Yao et al. (2023). *Tree of Thoughts: Deliberate Problem Solving with Large Language Models*.
https://arxiv.org/abs/2305.10601

17. Noah Shinn et al. (2023). *Reflexion: Language Agents with Verbal Reinforcement Learning*.
https://arxiv.org/abs/2303.11366

18. Charles Packer et al. (2023). *MemGPT: Towards LLMs as Operating Systems*.
https://arxiv.org/abs/2310.08560

19. Qingyun Wu et al. (2023). *AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation*.
https://arxiv.org/abs/2308.08155

20. Guanzhi Wang et al. (2023). *Voyager: An Open-Ended Embodied Agent with Large Language Models*.
https://arxiv.org/abs/2305.16291

21. Carlos E. Jimenez et al. (2023/2024). *SWE-bench: Can Language Models Resolve Real-World GitHub Issues?*.
https://arxiv.org/abs/2310.06770

22. Tianbao Xie et al. (2024). *OSWorld: Benchmarking Multimodal Agents for Open-Ended Tasks in Real Computer Environments*.
https://arxiv.org/abs/2404.07972

23. Yoshua Bengio (2017). *The Consciousness Prior*.
https://arxiv.org/abs/1709.08568

24. Larissa Albantakis et al. (2023). *Integrated Information Theory (IIT) 4.0*.
https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1011465

25. Michael Mashour, Pieter Roelfsema, Jean-Pierre Changeux & Stanislas Dehaene (2020). *Conscious Processing and the Global Neuronal Workspace Hypothesis*.
https://pasteur.hal.science/pasteur-02546335v1/file/Neuron%20Review%20penultimate%20version%20with%20figures.pdf

26. Victor Lamme (2018). *Challenges for Theories of Consciousness: Recurrent Processing Perspective*.
https://royalsocietypublishing.org/doi/pdf/10.1098/rstb.2017.0344

27. Andrew Wilterson & Michael Graziano (2021). *The Attention Schema Theory in a Neural Network Agent*.
https://grazianolab.princeton.edu/document/172

28. David Albarracin et al. (2023). *Designing Explainable Artificial Intelligence with Active Inference*.
https://arxiv.org/abs/2306.04025

29. Patrick Butlin et al. (2023). *Consciousness in Artificial Intelligence: Insights from the Science of Consciousness*.
https://arxiv.org/abs/2308.08708

30. Cogitate Consortium et al. (2025). *Adversarial Testing of Global Neuronal Workspace and Integrated Information Theories of Consciousness*.
https://www.nature.com/articles/s41586-025-08888-1

31. Melody Y. Guan et al. · OpenAI (2024). *Deliberative Alignment: Reasoning Enables Safer Language Models*.
https://arxiv.org/abs/2412.16339

32. Hoagy Cunningham et al. · Anthropic (2026). *Constitutional Classifiers++: Efficient Production-Grade Defenses against Universal Jailbreaks*.
https://arxiv.org/abs/2601.04603

33. Juraj Gottweis et al. · Google (2025). *Towards an AI Co-Scientist*.
https://arxiv.org/abs/2502.18864

34. Alexander Novikov et al. · Google DeepMind (2025). *AlphaEvolve: A Coding Agent for Scientific and Algorithmic Discovery*.
https://arxiv.org/abs/2506.13131

35. Mido Assran et al. · Meta (2025). *V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning*.
https://arxiv.org/abs/2506.09985

36. Gemini Robotics Team · Google DeepMind (2025). *Gemini Robotics 1.5: Generalist Robots with Advanced Embodied Reasoning, Thinking, and Motion Transfer*.
https://storage.googleapis.com/deepmind-media/gemini-robotics/Gemini-Robotics-1-5-Tech-Report.pdf

37. François Chollet et al. (2025). *ARC-AGI-2: A New Challenge for Frontier AI Reasoning Systems*.
https://arxiv.org/abs/2505.11831

38. Long Phan et al. (2025/2026 rev.). *Humanity's Last Exam*.
https://arxiv.org/abs/2501.14249

39. Elliot Glazer et al. (2024). *FrontierMath: A Benchmark for Evaluating Advanced Mathematical Reasoning in AI*.
https://arxiv.org/abs/2411.04872

40. Samuel Miserendino et al. · OpenAI (2025). *SWE-Lancer: Can Frontier LLMs Earn $1 Million from Real-World Freelance Software Engineering?*.
https://arxiv.org/abs/2502.12115

41. Jason Wei et al. · OpenAI (2025). *BrowseComp: A Simple Yet Challenging Benchmark for Browsing Agents*.
https://arxiv.org/abs/2504.12516

42. Giulio Starace et al. · OpenAI (2025). *PaperBench: Evaluating AI's Ability to Replicate AI Research*.
https://arxiv.org/abs/2504.01848

43. Hjalmar Wijk et al. · METR (2024/ICML 2025). *RE-Bench: Evaluating Frontier AI R&D Capabilities of Language Model Agents against Human Experts*.
https://arxiv.org/abs/2411.15114

44. Yutaro Yamada et al. · Sakana AI (2025). *The AI Scientist-v2: Workshop-Level Automated Scientific Discovery via Agentic Tree Search*.
https://arxiv.org/abs/2504.08066

45. Tianshi Zheng et al. (2026). *SciResearcher: Scaling Deep Research Agents for Frontier Scientific Reasoning*.
https://arxiv.org/abs/2605.01489

Selection rule. الأوراق 1–30 هي corpus الخارجي السابق المحدد في المشروع. الأوراق 31–45 أضيفت لتغطية الجبهة 2024–2026 في world models، embodied agents، AGI evaluation، deep research، AI R&D automation، reasoning safety، runtime safeguards، autonomous science، evolutionary discovery. تم تفضيل الأوراق الأصلية، arXiv/PMLR/Nature، وتقارير المختبرات الرسمية على summaries الثانوية.