

Synthese FR — Towards End-to-End Automation of AI Research

Nature, Vol. 651, 26 mars 2026

C. Lu, C. Lu, R.T. Lange, Y. Yamada, S. Hu, J. Foerster, D. Ha, J. Clune

Sakana AI, University of Oxford, University of British Columbia, Vector Institute

Synthese exhaustive en francais. Termes techniques conserves en anglais.
Source originale : <https://doi.org/10.1038/s41586-026-10265-5> License :
Creative Commons Attribution 4.0 International

Introduction

L'automatisation de la science est une ambition ancienne de l'intelligence artificielle. Avant les LLM (large language models), l'IA se limitait a des taches etroites : decouverte de structures chimiques (DENDRAL, Buchanan & Feigenbaum, 1978), preuves mathematiques (Lenat, 1977), decouverte de materiaux (Merchant et al., 2023 ; Szymanski et al., 2023), prediction de la structure 3D des proteines (AlphaFold, Jumper et al., 2021). D'autres systemes analysaient des jeux de donnees pre-collectes pour en extraire des insights (Langley et al., 1987 ; Ifargan et al., 2025).

Avec l'emergence des foundation models (modeles de fondation -- grands modeles pre-entraines generalistes), le role de l'IA s'est elargi : generation d'hypotheses, revues de litterature automatiques, assistance au code experimental. Mais **aucun systeme n'avait encore parcouru le cycle complet de la recherche** -- de la conception a la publication -- de facon autonome.

Ce paper presente **The AI Scientist**, un pipeline qui automatise l'integralite du processus scientifique de bout en bout (end-to-end) : ideation, recherche bibliographique, planification et execution d'experiences, analyse des resultats, redaction du manuscrit, et peer review automatique. Le domaine cible est la recherche en machine learning, ou les experiences sont entierement computationnelles.

Contribution centrale : un manuscrit genere par The AI Scientist a passe le premier tour de peer review d'un workshop de l'ICLR 2025 (International Conference on Learning Representations), l'une des conferences de machine learning les plus prestigieuses. Le workshop avait un taux d'acceptation de 70%.

1. Architecture du pipeline — The AI Scientist

The AI Scientist execute sequentiellement quatre phases principales (Fig. 1a du paper). Le systeme a ete teste dans deux modes : template-based (a partir d'un code fourni par un humain) et template-free (exploration ouverte sans code initial).

1.1 Phase 1 : Generation d'idees (Idea Generation)

Le systeme construit iterativement une **archive** de directions de recherche et d'hypotheses, dans un sous-domaine du machine learning specifie par l'utilisateur. Ce mecanisme d'archive croissante est inspire des algorithmes d'open-endedness (MAP-Elites, Mouret & Clune, 2015) qui maintiennent une collection diversifiee d'artefacts.

Pour chaque direction, le systeme genere : - Un titre descriptif - Le raisonnement (pourquoi cette idee est interessante) - Un plan experimental detaille - Des auto-evaluations sur des echelles 1-10 : interestingness, novelty, feasibility

Filtrage par nouveaute : chaque idee est verifiee automatiquement contre la litterature scientifique existante via l'API Semantic Scholar et l'accès web. Les idées ayant une forte similarité sémantique avec des travaux existants sont écartées. Le systeme effectue jusqu'a **10 tours de recherche bibliographique** par idee, en raffinant ses requetes a chaque tour.

Le prompt positionne le systeme comme "an ambitious AI PhD student who is looking to publish a paper that will contribute significantly to the field".

Variante template-free : l'ideation est plus abstraite, non liee a une implementation de code initiale. Le systeme genere des propositions de haut niveau qui ressemblent a l'abstract d'un article scientifique : probleme de recherche, methode proposee, resultats attendus. Ce processus est integre a un module de revue de litterature qui identifie les lacunes du domaine (knowledge gaps).

1.2 Phase 2 : Execution experimentale

Deux modes d'execution distincts sont proposes.

Mode template-based

Le systeme recoit un **template de code** qui reproduit un run d'entrainement d'un algorithme populaire sur un benchmark standard (par exemple, entrainer un petit transformer sur les oeuvres de Shakespeare via nanoGPT).

Le systeme elabore un plan experimental multi-etapes (jusqu'a **5 experiences**), puis les execute sequentiellement en utilisant **Aider** (assistant de codage open-source, Gauthier, 2024) pour modifier la codebase.

Robustesse aux erreurs : le systeme detecte automatiquement les echecs d'execution, capture les logs d'erreur et invoque une instance d'Aider pour le debugging automatique. Jusqu'a **4 cycles de correction** par experience. Timeout : **7 200 secondes** (2 heures) par experience.

Tous les resultats (metriques, graphiques, observations) sont consignes dans un **journal experimental** qui sert de memoire pour les etapes suivantes.

Mode template-free (open-ended)

Ce mode, plus ambitieux, supprime la dependance a un code de depart fourni par un humain. Il introduit plusieurs ameliorations majeures :

Experiment progress manager : un gestionnaire de progression qui coordonne quatre etapes structurees d'experimentation scientifique, chacune avec des criteres d'arret explicites :

1. **Preliminary investigation** (investigation preliminaire) -- se termine quand un prototype fonctionnel a ete execute avec succes
2. **Hyperparameter tuning** (reglage des hyperparametres) -- se termine quand les experiences se stabilisent (convergence des courbes d'entrainement, execution reussie sur au moins 2 datasets)
3. **Research agenda execution** (execution de l'agenda de recherche) -- se termine quand le budget computationnel est epuise
4. **Ablation studies** (etudes d'ablation) -- se termine quand le budget est epuise

Chaque noeud a un runtime maximum de **1 heure**. A la fin de chaque etape, un evaluateur base sur un LLM examine tous les noeuds feuilles et selectionne le plus prometteur comme racine de l'etape suivante.

Parallelized agentic tree search (recherche arborescente agentique parallelisee) : le workflow sequentiel du mode template-based est remplace par un **arbre de recherche parallelise**. Chaque noeud experimental suit un cycle d'execution :

1. Claude Sonnet 4 genere un plan d'experimentation et le code Python correspondant
2. Le code est execute dans un interpreteur Python
3. En cas d'erreur : le noeud est marque "buggy" et le cycle s'arrete
4. En cas de succes : phase de plotting -- le systeme genere des visualisations
5. Un VLM (vision-language model, GPT-4o) critique les visualisations (labels, legendes, coherence)
6. Le noeud est marque "non-buggy" s'il passe la critique VLM

A chaque iteration, le systeme selectionne plusieurs noeuds a etendre en parallele. Avec une probabilite predeterminee, un noeud buggy est choisi (priorite au debugging) ; sinon, un noeud non-buggy est selectionne pour amelioration. La selection entre noeuds non-buggy suit une strategie **best-first search** guidee par GPT-4o (metriques de performance, dynamique d'entrainement, qualite des graphiques).

Types de noeuds specialises : - **Hyperparameter nodes** : exploration systematique de configurations alternatives (etape 2) - **Ablation nodes** : evaluation des composants critiques (etape 4) - **Replication nodes** : replications avec differentes random seeds pour le calcul de statistiques (moyenne, ecart-type) - **Aggregation nodes** : consolidation et visualisation des resultats de replication (pas de nouvelle experience, uniquement aggregation)

Acces generalise aux datasets : le systeme peut integrer dynamiquement des datasets depuis le HuggingFace Hub. Dix exemples de datasets sont listes dans le prompt, et le systeme genere automatiquement le code de chargement. Les datasets locaux peuvent aussi etre utilises via des instructions ajoutees au prompt.

Duree typique du processus complet en mode template-free : **de plusieurs heures a plus de 15 heures**, selon la complexite du probleme.

1.3 Phase 3 : Redaction du manuscrit (Manuscript Generation)

Le systeme synthetise les resultats en un article scientifique complet au format de conference machine learning.

Mode template-based : utilise Aider pour remplir un template LaTeX standard section par section (introduction, methodes, resultats, conclusion). La section resultats est ecrite en analysant le journal experimental. Pour la section related work, le systeme interroge l'API Semantic Scholar (jusqu'a **20 tours de recherche**) et genere une justification textuelle pour chaque citation. Le manuscrit subit plusieurs passes d'edition automatique. Compilation LaTeX avec correction automatique des erreurs (jusqu'a **5 tours de correction**).

Mode template-free : passe a la generation directe de LaTeX via un modele de raisonnement (OpenAI o1), suivi de cycles de reflexion (self-reflection, Shinn et al., 2024). Le systeme agrege les resultats experimentaux de plusieurs etapes en figures composites. Des prompts specifiques guident la redaction selon le format du workshop cible (par exemple, le format ICBINB sur les resultats negatifs). Chaque cycle de reflexion integre le feedback de linters LaTeX et de revues VLM pour ameliorer la qualite des figures et l'alignement texte-figures.

Construction des citations : pour chaque citation potentielle, le systeme genere une justification d'inclusion indiquant comment utiliser la reference de facon appropriee dans le manuscrit. Le processus compare les resultats de la litterature au manuscrit genere sur **20 tours**.

1.4 Phase 4 : Peer review automatique (Automated Reviewer)

Le systeme inclut un **Automated Reviewer** construit sur o4-mini (OpenAI). Ce composant emule le processus de peer review d'une conference de premier rang, en suivant les guidelines officielles de la conference NeurIPS.

Processus de review : le reviewer traite le PDF du manuscrit et produit : - Un resume structure - Listes de forces (strengths) et faiblesses (weaknesses) - Questions aux auteurs - Limites et preoccupations ethiques identifiees - Scores numeriques : soundness, presentation, contribution, score global (1-10), niveau de confiance - Decision binaire : accept ou reject

Le prompt positionne le systeme comme "an AI researcher who is reviewing a paper that was submitted to a prestigious ML venue".

Mecanisme d'ensemble : pour chaque paper, **5 reviews independantes** sont generees, puis agregees en une **meta-review**. Le LLM joue alors le role d'un area chair qui cherche un consensus entre les 5 reviews individuelles. Ce mecanisme d'ensemble ameliore la robustesse de l'evaluation.

2. Technologies fondamentales

The AI Scientist s'appuie sur un ensemble de technologies complementaires :

LLMs autoregressifs (GPT-4, Claude, Gemini, o-series) : generation de texte par modelisation de la probabilite conditionnelle du prochain token. Capacites emergentes de raisonnement et de generation de code obtenues par scaling des donnees et du modele.

Agentic patterns (patterns agentiques -- schemas de comportement autonome pour LLM) : few-shot prompting (Brown et al., 2020) et self-reflection (Reflexion, Shinn et al., 2024) pour ameliorer performance et fiabilite.

Aider (Gauthier, 2024) : assistant de codage open-source conçu pour implémenter des fonctionnalités, corriger des bugs et refactoriser du code dans des codebases existantes. Utilise dans le mode template-based.

Tree search agentique : recherche arborescente parallélisée pour l'exploration expérimentale dans le mode template-free. Remplace l'approche séquentielle d'Aider.

VLM (vision-language model) : GPT-4o utilise pour critiquer les visualisations générées, vérifier l'alignement figures-légendes, et fournir du feedback pendant l'expérimentation et la rédaction.

Modèles spécialisés par rôle (mode template-free) : - OpenAI o3 : ideation et critique de code (raisonnement fort) - Claude Sonnet 4 : génération de code - GPT-4o : tâches vision-langage (rapport coût/efficacité) - o4-mini : raisonnement au stade du review (rapport coût/efficacité) - OpenAI o1 : génération directe de LaTeX pour la rédaction

API Semantic Scholar : recherche bibliographique automatisée, vérification de nouveauté, construction de la section related work.

HuggingFace Hub : accès dynamique à des datasets publics pour élargir les capacités expérimentales.

3. Resultats

3.1 Qualite des manuscrits generes

L'analyse par l'Automated Reviewer révèle une **tendance claire : la qualité des papiers augmente avec l'amélioration des modèles sous-jacents** (Fig. 1b). La corrélation entre la date de sortie du modèle et le score du papier est statistiquement significative ($P < 0.00001$, $R^2 = 0.517$).

Modèles testés dans le mode template-based (3 replications par point) : GPT-4, Sonnet-3, Gemini-1.5, GPT-4o, Sonnet-3.5. Mode template-free (6 replications par point) : Sonnet-3.5, GPT-4o, Gemini-2.0, o1, Sonnet-3.7, o3, Sonnet-4, Gemini-2.5.

Avec les generations recentes de modeles, The AI Scientist produit en moyenne des papers qui **approchent le seuil d'acceptabilite limite** (borderline acceptability) pour les workshops de conferences de machine learning.

Une **forte correlation entre le budget de compute alloue par paper et la qualite resultante** est egalement observee (Fig. 3c). Avec un tree search plus profond (plus de noeuds experimentaux), les scores s'ameliorent de maniere consistante. Ceci indique que le model scale et l'inference-time compute jouent tous deux un role important.

3.2 Peer review et soumission ICLR

Pour tester le systeme dans les conditions reelles d'evaluation de la science humaine, les auteurs ont soumis **3 manuscrits generes** au processus formel de peer review du workshop **I Can't Believe It's Not Better (ICBINB)** de l'ICLR 2025, avec l'accord des organisateurs et l'approbation ethique de l'University of British Columbia (IRB H24-02652).

Protocole : - Le mode template-free a ete adapte en lui donnant simplement le theme du workshop (investigation des limites du deep learning, cas ou des idees d'amelioration n'ont pas fonctionne) - Les outputs les plus prometteurs ont ete filtres manuellement a chaque etape - Criteres de selection : alignement avec le sujet du workshop, implementation correcte du code, correction du formatage du manuscrit - Le workflow scientifique complet (ideation, codage, redaction) a ete realise **sans aucune modification humaine** - Les 3 soumissions etaient parmi les 43 papers evalues pour le workshop - Les reviewers etaient informes que certaines soumissions etaient generees par IA, mais pas lesquelles (processus en aveugle) - Protocole pre-etabli : tous les papers generes par IA seraient retires apres le review, quel que soit le resultat

Resultats : - **1 paper sur 3 a ete accepte** : score moyen de **6,33** (scores individuels : 6 -- weak accept, 7 -- accept, 6 -- weak accept), au-dessus du seuil moyen d'acceptation du workshop, dans le top 45% des soumissions - Les organisateurs ont declare que le paper aurait ete accepte selon toute vraisemblance s'il n'avait pas ete retire conformement au protocole - Ce paper accepte rapportait un **resultat negatif**, en accord avec la thematique du workshop - Les 2 autres papers n'ont pas atteint le seuil d'acceptation

Evaluation interne : l'equipe de chercheurs humains a conclu que si un des papers atteignait le niveau workshop, **aucun ne satisfaisait le niveau de la**

conference principale ICLR (taux d'acceptation 32% vs 70% pour le workshop).

3.3 Validation de l'Automated Reviewer

L'Automated Reviewer a été benchmark contre les décisions humaines en utilisant les données ICLR du dataset OpenReview.

Metriques de performance (Table 1) :

Metrique	Humain (NeurIPS)	Automated Reviewer (avant cutoff)	Automated Reviewer (apres cutoff)
Balanced accuracy	0.66	0.69 +/- 0.04	0.66 +/- 0.03
Accuracy	0.73	0.65 +/- 0.10	0.63 +/- 0.09
F1 score	0.49	0.62 +/- 0.09	0.67 +/- 0.09
AUC	0.65	0.69 +/- 0.09	0.65 +/- 0.10
FPR	0.17	0.45 +/- 0.10	0.52 +/- 0.10
FNR	0.52	0.17 +/- 0.08	0.17 +/- 0.07

Points clés : - L'Automated Reviewer atteint une **performance comparable ou supérieure** aux humains sur les métriques F1, AUC et balanced accuracy - Les **tests statistiques** confirment ces résultats : tests z à deux échantillons (pas de différence significative avant cutoff $P = 0.319$, après cutoff $P = 0.921$) ; tests bootstrap non-paramétriques sur le F1 (surperformance automatisée $P < 0.001$) - **Contamination des données** : une légère baisse de la balanced accuracy de 69% à 66% est observée après le knowledge cutoff, suggérant une possible contamination minimale. Mais les résultats post-cutoff restent comparables aux performances humaines - Benchmark humain de référence : l'expérience de cohérence NeurIPS 2021 (Beygelzimer et al., 2021), où environ 10% des soumissions ont été envoyées à deux comités de review indépendants

Les auteurs notent une **limitation** de cette comparaison : les pools de papiers ICLR et NeurIPS sont différents (shift de distribution). Cependant, ICLR est la seule conférence majeure ML publiant toutes les décisions accept/reject, et

l'experience NeurIPS 2021 est le seul benchmark moderne de coherence inter-reviewers.

3.4 Correlation qualite/modele

La Fig. 1b montre que les scores des papers augmentent de facon **monotone** avec la date de sortie du modele sous-jacent, pour les deux modes (template-based et template-free). Cette tendance est statistiquement significative ($P < 0.00001$) et implique fortement que les **futures versions du systeme seront substantiellement plus performantes** a mesure que les modeles de fondation continuent de progresser.

De plus, les couts des systemes IA continuent de decroitre de facon exponentielle tandis que les capacites augmentent de facon exponentielle (AI Index Report, Maslej et al., 2024), suggerant un potentiel d'amelioration substantielle.

4. Modes d'operation

4.1 Focused mode (template-based)

Dans ce mode, le systeme recoit un template de code humain (par exemple, un script d'entrainement nanoGPT sur Shakespeare, enwiki8 ou text8). Le workflow est lineaire : ideation, execution sequentielle via Aider, redaction, review.

Avantages : simplicity, controlabilite, resultats reproductibles dans un perimetre bien defini. **Limites** : la creativite du systeme est contrainte par le template initial. Les idees generees sont des extensions ou variations du code de depart.

4.2 Open-ended mode (template-free)

Ce mode ne recoit aucun code initial. Il genere ses propres scripts, utilise le tree search parallelise, integre des datasets depuis HuggingFace, et emploie une combinaison de modeles specialises (o3, Claude Sonnet 4, GPT-4o, o4-mini, o1).

Avantages : exploration plus large, capacite a aborder des sujets non anticipes, meilleure scalabilite computationnelle grace au parallelisme.
Limites : cout computationnel plus eleve (plusieurs heures a 15+ heures), complexite accrue du pipeline, risque plus eleve de bugs dans le code auto-genere.

4.3 Comparaison des deux modes

Le mode template-free a ete utilise pour la soumission ICLR, demontrant sa capacite a s'adapter a un cadre reel de soumission simplement en specif iant le theme du workshop dans le prompt. Le mode template-based a ete utilise pour des evaluations systematiques sur des benchmarks controles. Les deux modes montrent une amelioration de la qualite avec de meilleurs modeles (Fig. 1b), mais le mode template-free beneficie davantage du scaling du compute (Fig. 3c) grace a sa strategie de tree search.

5. Limites et modes d'echec

Les auteurs documentent honnemment les limites du systeme :

Taux d'acceptation : 1 sur 3 au workshop ICLR (taux d'acceptation du workshop : 70%). Le systeme ne peut pas encore satisfaire de facon consistante le seuil des workshops, et encore moins celui des conferences principales (taux d'acceptation ICLR 2025 : 32%).

Modes d'echec documentes (failure modes) : - Generation d'idees naives ou sous-developpees - Implementations incorrectes de l'idee principale - Manque de rigueur methodologique approfondie - Erreurs dans l'implementation experimentale - Duplication de figures entre le texte principal et l'appendice -
Hallucinations multiples : citations inexactes, faits inventes, descriptions erronees de resultats

Limites structurelles : - Le systeme ne conduit que des **experiences computationnelles**. Les sciences necessitant des manipulations physiques (chimie, biologie) sont hors de portee pour l'instant, bien que les auteurs evoquent les progres rapides des laboratoires chimiques automatises (Boiko et al., 2023) - La longueur des taches que l'IA peut accomplir de facon fiable **double tous les 7 mois** (METR, Kwa et al., 2025), ce qui suggere que

beaucoup de goulots d'etranglement actuels pourraient etre resolus a court terme - Certaines faiblesses de l'IA se sont revelees **etonnamment difficiles a resoudre** : vulnerability aux adversarial examples (Szegedy et al., 2014 ; Nguyen et al., 2015), hallucinations et confiance excessive (Maynez et al., 2020), bien que des progres aient ete realises (GPT-5 System Card, Huang et al., 2025) - Il n'est **pas clair dans quelle mesure les systemes IA peuvent produire de veritables sauts conceptuels creatifs** en science

6. Implications ethiques et risques

Les auteurs identifient plusieurs categories de risques :

Surcharge du systeme de peer review : la capacite de generer des papers a grande echelle pourrait submerger des systemes de review deja sous pression, degradant la qualite du filtrage scientifique.

Bruit dans la litterature scientifique : multiplication de publications de faible qualite, rendant plus difficile l'identification des contributions significatives.

Inflation des credentials de recherche : possibilite d'utiliser le systeme pour gonfler artificiellement un CV academique.

Appropriation d'idees : risque de reprendre les idees d'autrui sans credit adequat.

Elimination d'emplois : menace potentielle pour les postes de chercheurs, en particulier au niveau junior.

Experiences non ethiques ou dangereuses : un systeme autonome pourrait conduire des experiences dont les implications ethiques n'ont pas ete evaluees.

Dual use : les capacites du systeme pourraient etre detournees a des fins malveillantes.

Mesures prises par les auteurs : - Obtention de l'approbation ethique de l'IRB de l'University of British Columbia (H24-02652) - Cooperation complete avec la direction de l'ICLR et les organisateurs du workshop - Les reviewers humains ont ete informes que certaines soumissions pouvaient etre generees

par IA (avec option d'opt-out) - Decision pre-etablie de retirer tous les papers generes par IA apres le review, quel que soit le resultat, pour eviter de creer un precedent avant que la communaute n'ait etabli des standards clairs de divulgation et d'evaluation - Les auteurs soulignent qu'il est **critique de developper des normes** pour l'utilisation responsable de ces systemes

7. Discussion et perspectives

Trajectoire d'amelioration : les auteurs argumentent que, historiquement en machine learning, une fois qu'un systeme commence a fonctionner (meme avec des defauts evidents), en quelques annees avec le scaling (compute, donnees, meilleurs modeles, meilleures techniques), les capacites deviennent surprenantes et peuvent depasser la performance humaine. Ils invitent a evaluer cette technologie non sur ses capacites actuelles, mais sur sa **trajectoire probable**.

Correlation modele/qualite : la forte correlation entre la qualite du modele sous-jacent et la qualite des papers generes ($R^2 = 0.517$, $P < 0.00001$) est le signal le plus fort du paper. Elle implique que les ameliorations futures des foundation models se traduiront directement en ameliorations de The AI Scientist, sans modification du pipeline lui-meme.

Scaling du compute : la qualite s'ameliore avec le nombre de noeuds experimentaux dans le tree search (Fig. 3c), ce qui ouvre la voie a des ameliorations par simple investissement en inference-time compute.

Extension a d'autres domaines : le meme playbook pourrait s'appliquer a d'autres sciences ou les experiences sont automatisables (chimie avec laboratoires robotises, biologie computationnelle), ou ou des humains conduisent les experiences physiques et collectent les donnees.

Travaux connexes : les auteurs situent leur travail dans la continuite de DENDRAL (1978), AM et EURISKO (Lenat, 1984), les systemes de decouverte computationnelle (Langley et al., 1987), et les travaux recents sur l'automatisation partielle de la recherche (ResearchAgent, AutoSurvey, MAgentBench, AIDE, MLE-bench, Darwin Godel Machine).

Recherche future necessaire : - S'assurer que l'exploration ouverte par l'IA procede de facon **sure et alignee avec les valeurs humaines** (Bommasani

et al., 2021 ; Ecoffet et al., 2020) - Etudier et ameliorer la capacite des systemes IA a produire des **idees creatives representant de veritables sauts conceptuels** - Developper des normes communautaires pour la divulgation et l'evaluation des recherches generees par IA

8. Conclusion

The AI Scientist demontre qu'un systeme IA peut traverser l'ensemble du cycle de la recherche scientifique, de l'ideation a la peer review, et produire un manuscrit accepte dans un processus reel d'evaluation par les pairs d'un workshop d'une conference de premier rang en machine learning.

Les resultats quantitatifs cles sont : - 1 paper accepte sur 3 soumissions (score 6,33/10, seuil de workshop) - L'Automated Reviewer atteint une performance comparable aux reviewers humains (balanced accuracy 66-69%, F1 0.62-0.67 vs 0.49 pour les humains) - La qualite des papers croit de facon statistiquement significative avec la qualite du modele sous-jacent ($P < 0.00001$) - La qualite croit avec le budget de compute alloue

Ce resultat constitue un **milestone** -- le premier exemple documente d'un paper integralement genere par IA naviguant avec succes un processus de peer review -- tout en etant explicitement situe comme un point de depart. Le systeme ne peut pas encore produire de la science au niveau des publications principales, et de nombreux modes d'echec persistent. Les auteurs appellent a un developpement responsable de ces systemes, avec des normes communautaires claires, pour que cette capacite accelere la decouverte scientifique plutot qu'elle n'en degrade l'ecosysteme.

Sources

Les references ci-dessous sont celles citees dans le paper, dans l'ordre de citation.

1. Lenat, D. B. "Automated theory formation in mathematics." Proc. 5th IJCAI, 833-842 (1977).

2. Buchanan, B. G. & Feigenbaum, E. A. "Dendral and meta-dendral: their applications dimension." *Artif. Intell.* 11, 5-24 (1978).
3. OpenAI. "GPT-4 technical report." *arXiv:2303.08774* (2023).
4. Anthropic. "The Claude 3 Model Family: Opus, Sonnet, Haiku." (2024).
5. Grattafiori, A. et al. "The Llama 3 herd of models." *arXiv:2407.21783* (2024).
6. Jiang, Z. et al. "AIDE: AI-driven exploration in the space of code." *arXiv:2502.13138* (2025).
7. Chan, J. S. et al. "MLE-bench: evaluating ML agents on ML engineering." *Proc. ICLR* (2025).
8. Waltz, D. & Buchanan, B. G. "Automating science." *Science* 324, 43-44 (2009).
9. Langley, P. "Integrated systems for computational scientific discovery." *Proc. AAAI* 38, 22598-22606 (2024).
10. Langley, P., Simon, H. A., Bradshaw, G. L. & Zytkow, J. M. *Scientific Discovery: Computational Explorations of the Creative Process*. MIT (1987).
11. Lenat, D. B. & Brown, J. S. "Why AM and EURISKO appear to work." *Artif. Intell.* 23, 269-294 (1984).
12. Pyzer-Knapp, E. O. et al. "Accelerating materials discovery using AI, HPC and robotics." *npj Comput. Mater.* 8, 84 (2022).
13. Merchant, A. et al. "Scaling deep learning for materials discovery." *Nature* 624, 80-85 (2023).
14. Szymanski, N. J. et al. "An autonomous laboratory for accelerated synthesis of novel materials." *Nature* 624, 86-91 (2023).
15. Jumper, J. et al. "Highly accurate protein structure prediction with AlphaFold." *Nature* 596, 583-589 (2021).
16. Hayes, T. et al. "Simulating 500 million years of evolution with a language model." *Science* 387, 850-858 (2025).
17. Ifargan, T. et al. "Autonomous LLM-driven research -- from data to human-verifiable research papers." *NEJM AI* 2, Aloa2400555 (2025).
18. Falkenhainer, B. C. & Michalski, R. S. "Integrating quantitative and qualitative discovery: the ABACUS system." *Mach. Learn.* 1, 367-401 (1986).

19. Girotra, K. et al. "Ideas Are Dimes a Dozen: LLMs for Idea Generation in Innovation." Working paper, Cornell Univ. (2023).
20. Lehman, J. et al. "Open-endedness." In Handbook of Evolutionary Machine Learning, Springer Nature (2023).
21. Lu, C., Hu, S. & Clune, J. "Automated capability discovery via model self-exploration." arXiv:2502.07577 (2025).
22. Faldor, M. et al. "OMNI-EPIC: open-endedness via models of human notions of interestingness." Proc. ICLR (2025).
23. Hu, S., Lu, C. & Clune, J. "Automated design of agentic systems." Proc. ICLR (2025).
24. Baek, J. et al. "ResearchAgent: Iterative research idea generation over scientific literature with LLMs." Proc. NAACL (2025).
25. Wang, Y. et al. "AutoSurvey: LLMs can automatically write surveys." Proc. NeurIPS 37 (2024).
26. Huang, Q. et al. "MLAgentBench: evaluating language agents on ML experimentation." Proc. ICML 235 (2024).
27. Lu, C. et al. "Discovering preference optimization algorithms with and for LLMs." Proc. NeurIPS 37 (2024).
28. Ma, Y. J. et al. "Eureka: human-level reward design via coding LLMs." Proc. ICLR (2024).
29. Zhang, J. et al. "Darwin Godel Machine: open-ended evolution of self-improving agents." Proc. ICLR (2026).
30. Mouret, J.-B. & Clune, J. "Illuminating search spaces by mapping elites." arXiv:1504.04909 (2015).
31. Fricke, S. "Semantic Scholar." J. Med. Libr. Assoc. 106, 145 (2018).
32. Schick, T. et al. "Toolformer: language models can teach themselves to use tools." Proc. NeurIPS 36 (2023).
33. Gonzalez-Marquez, R. & Kobak, D. "Learning representations of learning representations." DMLR Workshop at ICLR (2024).
34. Beygelzimer, A. et al. "The NeurIPS 2021 consistency experiment." NeurIPS Blog (2021).
35. Efron, B. & Tibshirani, R. J. An Introduction to the Bootstrap. Chapman & Hall CRC (1993).
36. Lehmann, E. L. Testing Statistical Hypotheses. 1st ed., Wiley (1959).

37. Maslej, N. et al. "Artificial Intelligence Index Report 2024." arXiv: 2405.19522 (2024).
38. "I Can't Believe it's Not Better: Challenges in Applied Deep Learning." ICLR Workshop (2025). <http://openreview.net/group?id=ICLR.cc/2025/Workshop/ICBINB>
39. ICLR 2025 Fact Sheet. https://media.iclr.cc/Conferences/ICLR2025/ICLR2025_Fact_Sheet.pdf
40. Kwa, T. et al. "Measuring AI ability to complete long tasks." METR Blog (2025). <https://metr.org/blog/2025-03-19-measuring-ai-ability-to-complete-long-tasks/>
41. Nguyen, A., Yosinski, J. & Clune, J. "Deep neural networks are easily fooled." Proc. CVPR (2015).
42. Szegedy, C. et al. "Intriguing properties of neural networks." Proc. ICLR (2014).
43. Maynez, J. et al. "On faithfulness and factuality in abstractive summarization." Proc. ACL (2020).
44. OpenAI. "GPT-5 System Card." (2025). <https://openai.com/index/gpt-5-system-card/>
45. Huang, L. et al. "A survey on hallucination in large language models." ACM Trans. Inf. Syst. 43, 42 (2025).
46. Boiko, D. A. et al. "Autonomous chemical research with large language models." Nature 624, 570-578 (2023).
47. Bommasani, R. et al. "On the opportunities and risks of foundation models." arXiv:2108.07258 (2021).
48. Ecoffet, A., Clune, J. & Lehman, J. "Open questions in creating safe open-ended AI." Proc. Artificial Life Conference 32 (2020).
49. Wang, L. et al. "A survey on LLM-based autonomous agents." Front. Comput. Sci. 18, 186345 (2024).
50. Brown, T. et al. "Language models are few-shot learners." Proc. NeurIPS 33 (2020).
51. Shinn, N. et al. "Reflexion: language agents with verbal reinforcement learning." Proc. NeurIPS 36 (2024).
52. Gauthier, P. "Aider." GitHub (2024). <https://github.com/paul-gauthier/aider>
53. Vaswani, A. et al. "Attention is all you need." Proc. NeurIPS 30 (2017).

54. Stanley, K. O., Lehman, J. & Soros, L. "Open-endedness: the last grand challenge you've never heard of." O'Reilly Radar (2017).
55. Wolf, T. et al. "Transformers: state-of-the-art natural language processing." Proc. EMNLP (2020).
56. El-Kishky, A. "OpenAI o1 System Card." (2024).
57. OpenAI. "Introducing OpenAI o3 and o4-mini." (2025). <https://openai.com/index/introducing-o3-and-o4-mini/>
58. Karpathy, A. "NanoGPT." GitHub (2022). <https://github.com/karpathy/nanoGPT>
59. Karpathy, A. "The unreasonable effectiveness of recurrent neural networks." (2015).
60. Hutter, M. "The Hutter Prize." (2006). <http://prize.hutter1.net>
61. Mahoney, M. "About the Test Data." (2011). <http://mattmahoney.net/dc/textdata.html>
62. Mensah, P. K. et al. "CCMT: dataset for crop pest and disease detection." Data Brief 49, 109306 (2023).
63. Sagawa, S. et al. "Waterbirds." GitHub (2019). https://github.com/kohpangwei/group_DRO
64. Liu, Z. et al. "Deep learning face attributes in the wild." Proc. ICCV (2015).

Code source : - Template-based + Automated Reviewer : <https://github.com/SakanaAI/AI-Scientist> (Apache 2.0) - Template-free : <https://github.com/SakanaAI/AI-Scientist-v2> (Apache 2.0)