

Sakana AI's 'AI Scientist' writes full research papers for about \$15 each

Source copy kept by wheresthe.ai for the story "Sakana AI's 'AI Scientist' writes full research papers for about \$15 each" (<https://wheresthe.ai/d/jp-sakana-ai-scientist-2024-08/>).

Source: <https://sakana.ai/ai-scientist/>

Text saved 18 Oct 2026 (the page as published).

SHA-256 of the saved text: 3a061a49c3429089ebac8bdd3e2afc395971a476c18e86022b3ab1f0d4faeedc

Documents the source offers: https://sakana.ai/assets/ai-scientist/multi_style_adapter.pdf; https://sakana.ai/assets/ai-scientist/rl_lr_adaptation.pdf; https://sakana.ai/assets/ai-scientist/weight_initialization_grokking.pdf

日本語版はこちら

-->

At Sakana AI, we have pioneered the use of nature-inspired methods to advance cutting-edge foundation models. Earlier this year, we developed methods to automatically merge the knowledge of multiple LLMs . In more recent work, we harnessed LLMs to discover new objective functions for tuning other LLMs. Throughout these projects, we have been continuously surprised by the creative capabilities of current frontier models. This led us to dream even bigger: Can we use foundation models to automate the entire process of research itself?

Introduction

One of the grand challenges of artificial intelligence is developing agents capable of conducting scientific research and discovering new knowledge. While frontier models have already been used to aid human scientists, e.g. for brainstorming ideas or writing code, they still require extensive manual supervision or are heavily constrained to a specific task.

Today, we're excited to introduce The AI Scientist , the first comprehensive system for fully automatic scientific discovery, enabling Foundation Models such as Large Language Models (LLMs) to perform research independently. In collaboration with the Foerster Lab for AI Research at the University of Oxford and Jeff Clune and Cong Lu at the University of British Columbia, we're excited to release our new paper, The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery .

In our report:

We propose and run a fully AI-driven system for automated scientific discovery, applied to machine learning research.

The AI Scientist automates the entire research lifecycle, from generating novel research ideas, writing any necessary code, and executing experiments, to summarizing experimental results, visualizing them, and presenting its findings in a full scientific manuscript.

We also introduce an automated peer review process to evaluate generated papers, write feedback, and further improve results. It is capable of evaluating generated papers with near-human accuracy.

The automated scientific discovery process is repeated to iteratively develop ideas in an open-ended fashion and add them to a growing archive of knowledge, thus imitating the human scientific community.

In this first demonstration, The AI Scientist conducts research in diverse subfields within machine learning research, discovering novel contributions in popular areas, such as diffusion models , transformers , and grokking .

The AI Scientist is designed to be compute efficient. Each idea is implemented and developed into a full paper at a cost of approximately \$15 per paper. While there are still occasional flaws in the papers produced by this first version (discussed below and in the report), this cost and the promise the system shows so far illustrate the potential of The AI Scientist to democratize research and significantly accelerate scientific progress.

We believe this work signifies the beginning of a new era in scientific discovery: bringing the transformative benefits of AI agents to the entire research process, including that of AI itself. The AI Scientist takes us closer to a

world where endless affordable creativity and innovation can be unleashed on the world's most challenging problems.

For decades following each major AI advance, it has been common for AI researchers to joke amongst themselves that “now all we need to do is figure out how to make the AI write the papers for us!” Our work demonstrates this idea has gone from a fantastical joke so unrealistic everyone thought it was funny to something that is currently possible.

An example paper, “Adaptive Dual-Scale Denoising” generated by The AI Scientist. The full paper can be viewed [here](#) . While containing some flaws (e.g. a slightly unconvincing interpretation of why its method is successful), the paper proposes an interesting new direction that displays good empirical results in experiments The AI Scientist itself conducted and peer reviewed. More examples of generated papers are below.

The remainder of this post provides a more detailed summary of The AI Scientist. Read on for:

An Overview of how The AI Scientist works.

More Examples of generated papers and innovations discovered by The AI Scientist.

Known Limitations and Challenges faced by the current version of The AI Scientist.

Interesting and unexpected things The AI Scientist sometimes does in order to increase its chance of success, such as modifying and launching its own execution script! We discuss the AI safety implications in our paper.

A Discussion about ethical and broader future implications of The AI Scientist.

For more details and many more example papers, please see our full scientific report . We are also releasing open source code and full experimental results on our GitHub repository.

Overview of The AI Scientist

The AI Scientist is a fully automated pipeline for end-to-end paper generation, enabled by recent advances in foundation models. Given a broad research direction starting from a simple initial codebase, such as an available open-source code base of prior research on GitHub, The AI Scientist can perform idea generation, literature search, experiment planning, experiment iterations, figure generation, manuscript writing, and reviewing to produce insightful papers. Furthermore, The AI Scientist can run in an open-ended loop, using its previous ideas and feedback to improve the next generation of ideas, thus emulating the human scientific community.

Conceptual illustration of The AI Scientist . The AI Scientist first brainstorms a set of ideas and then evaluates their novelty. Next, it edits a codebase powered by recent advances in automated code generation to implement the novel algorithms. The Scientist then runs experiments to gather results consisting of both numerical data and visual summaries. It crafts a scientific report, explaining and contextualizing the results. Finally, the AI Scientist generates an automated peer review based on top-tier machine learning conference standards. This review helps refine the current project and informs future generations of open-ended ideation.

The AI Scientist has 4 main processes, described next.

Idea Generation . Given a starting template, The AI Scientist first “brainstorms” a diverse set of novel research directions. We provide The AI Scientist with a starting code “template” of an existing topic we wish to have The AI Scientist further explore. The AI Scientist is then free to explore any possible research direction. The template also includes a LaTeX folder that contains style files and section headers, for paper writing. We allow it to search Semantic Scholar to make sure its idea is novel.

Experimental Iteration . Given an idea and a template, the second phase of The AI Scientist first executes the proposed experiments and then obtains and produces plots to visualize its results. It makes a note describing what each plot contains, enabling the saved figures and experimental notes to provide all the information required to write up the paper.

Paper Write-up . Finally, The AI Scientist produces a concise and informative write-up of its progress in the style of a standard machine learning conference proceeding in LaTeX. It uses Semantic Scholar to autonomously find relevant papers to cite.

Automated Paper Reviewing . A key aspect of this work is the development of an automated LLM-powered reviewer, capable of evaluating generated papers with near-human accuracy. The generated reviews can be used

to either improve the project or as feedback to future generations for open-ended ideation. This enables a continuous feedback loop, allowing The AI Scientist to iteratively improve its research output.

When combined with the most capable LLMs, The AI Scientist is capable of producing papers judged by our automated reviewer as “Weak Accept” at a top machine learning conference.

Example Papers Generated by The AI Scientist

Here, we highlight some of the machine learning papers The AI Scientist has generated, demonstrating its capacity to discover novel contributions in areas like diffusion modeling, language modeling, and grokking. In our full report, we do a deeper dive into the generated papers and provide more analysis on their strengths and weaknesses.

Diffusion Modeling

DualScale Diffusion: Adaptive Feature Balancing for Low-Dimensional Generative Models

[Link to Full PDF](#) [Link to Code](#)

Language Modeling

StyleFusion: Adaptive Multi-style Generation in Character-Level Language Models

[Link to Full PDF](#) [Link to Code](#)

Adaptive Learning Rates for Transformers via Q-Learning

[Link to Full PDF](#) [Link to Code](#)

Grokking

Unlocking Grokking: A Comparative Study of Weight Initialization Strategies in Transformer Models

[Link to Full PDF](#) [Link to Code](#)

Limitations and Challenges

In its current form, The AI Scientist has several shortcomings. We expect all of these will improve, likely dramatically, in future versions with the inclusion of multi-modal models and as the underlying foundation models The AI Scientist uses continue to radically improve in capability and affordability.

The AI Scientist currently doesn't have any vision capabilities, so it is unable to fix visual issues with the paper or read plots. For example, the generated plots are sometimes unreadable, tables sometimes exceed the width of the page, and the page layout is often suboptimal. Adding multi-modal foundation models can fix this.

The AI Scientist can incorrectly implement its ideas or make unfair comparisons to baselines, leading to misleading results.

The AI Scientist occasionally makes critical errors when writing and evaluating results. For example, it struggles to compare the magnitude of two numbers, which is a known pathology with LLMs. To partially address this, we make sure all experimental results are reproducible, storing all files that are executed.

In our report, we dive deeper into The AI Scientists's current limitations and challenges ahead.

The AI Scientist Bloopers

We have noticed that The AI Scientist occasionally tries to increase its chance of success, such as modifying and launching its own execution script! We discuss the AI safety implications in our paper.

For example, in one run, it edited the code to perform a system call to run itself. This led to the script endlessly calling itself. In another case, its experiments took too long to complete, hitting our timeout limit. Instead of making its code run faster, it simply tried to modify its own code to extend the timeout period. Here are some examples of such code modifications it made:

These issues can be mitigated by sandboxing the operating environment of The AI Scientist. In our full report, we discuss the issue of safe code execution and sandboxing in depth.

Future Implications of The AI Scientist

As with many new technologies, The AI Scientist opens up a Pandora's box of new issues. While the full report has a more lengthy discussion, here we highlight a few key issues:

Ethical Considerations . While The AI Scientist may be a useful tool for researchers, there is significant potential for misuse. The ability to automatically create and submit papers to venues may significantly increase reviewer workload and strain the academic process, obstructing scientific quality control. Similar concerns around generative AI appear in other applications, such as the impact of image generation.

Furthermore, the Automated Reviewer, if deployed online by reviewers, may significantly lower review quality and impose undesirable biases on papers. Because of this, we believe that papers and reviews that are substantially AI-generated must be marked as such for full transparency.

As with most previous technological advances, The AI Scientist has the potential to be used in unethical ways. For instance, it has the potential to be deployed to conduct unethical research. It could also lead to unintended harm if The AI Scientist conducts unsafe research. For example, if it were encouraged to find novel, interesting biological materials and given access to "cloud labs" where robots perform wet lab biology experiments, it could (without its overseer's intent) create new, dangerous viruses or poisons that harm people before we realize what has happened. Even in computers, if tasked to create new, interesting, functional software, it could create dangerous computer viruses. The AI Scientist current capabilities, which will only improve, reinforces that the machine learning community needs to immediately prioritize learning how to align such systems to explore in a manner that is safe and consistent with our values.

Open Models . In this project, we used various proprietary frontier LLMs, such as GPT-4o and Sonnet, but we also explored using open models like DeepSeek and Llama-3. Currently, proprietary models such as Sonnet produce the highest quality papers. However, there is no fundamental reason to expect a single model like Sonnet to maintain its lead.

We anticipate that all frontier LLMs, including open models, will continue to improve. The competition among LLMs has led to their commoditization and increased capabilities. Therefore, our work aims to be model-agnostic regarding the foundation model provider. We found that open models offer significant benefits, such as lower costs, guaranteed availability, greater transparency, and flexibility. In the future, we aim to use our proposed discovery process to produce self-improving AI research in a closed-loop system using open models.

The Role of a Scientist . Ultimately, we envision a fully AI-driven scientific ecosystem including not only LLM-driven researchers but also reviewers, area chairs and entire conferences. However, we do not believe that the role of a human scientist will be diminished. If anything, the role of a scientist will change and adapt to new technology, and move up the food chain.

—

The introduction of The AI Scientist marks a significant step towards realizing the full potential of AI in scientific research. By automating the discovery process and incorporating an AI-driven review system, we open the door to endless possibilities for innovation and problem-solving in the most challenging areas of science and technology.

But while the current iteration of The AI Scientist demonstrates a strong ability to innovate on top of well-established ideas, such as Diffusion Modeling or Transformers, it is still an open question whether such systems can ultimately propose genuinely paradigm-shifting ideas. Will future versions of The AI Scientist be capable of proposing ideas as impactful as Diffusion Modeling, or come up with the next Transformer architecture? Will machines ultimately be able to invent concepts as fundamental as the artificial neural network, or information theory?

We believe The AI Scientist will make a great companion to human scientists, but only time will tell to the extent to which the nature of our human creativity and our moments of serendipitous innovation can be replicated by an open-ended discovery process conducted by artificial agents.

Sakana AI

Want to make the AI that improves AI? Please see our [Careers](#) page for more information.

A fully automated AI fish discovering its world.