Carl Technical Report

Autoscience Institute
inquiry@autoscience.ai

March 3rd 2025

1 Introduction

Artificial intelligence has transformed numerous industries, yet research into the technology itself remains human-driven. At the Autoscience Institute, we envision a future where AI systems can autonomously conduct novel AI research. This vision led us to develop Carl — an AI system to produce academically rigorous research papers. Provided a research direction in the form of some grounding literature, Carl produces scientific hypotheses, implements experiments, and writes papers with limited human intervention.

The academic community has already seen a rise in the use of AI in academic writing. Kobak et al. demonstrated that in 2024, approximately 20% of scientific abstracts in the field of computer science included a higher than typical usage of words associated with LLMs [5]. The contributions from AI measured by Kobak et al. are limited to the manuscript writing portion of the AI research process [5]. The Autoscience Institute set out to build something capable of completing the entire research cycle – from ideation to presentation – that meets the bar of peer review. With Carl taking meaningful strides towards this goal, we present a series of questions to the academic community and discuss how this community may build on real, well-attributed research works originating from AI systems, while upholding standards of academic attribution and responsibility in conducting science.

1.1 Background

Several recent advancements in automated AI research have enabled Large Language Models (LLMs) to complete parts of the AI research process. Cognition's Devin demonstration video showed that SWE agents were becoming capable of implementing and running basic machine learning tasks [11]. Sakana's AI Scientist paper provided a first pass at a fully autonomous research scientist agent with the goal of creating scientific papers end-to-end [7]. The MLR-Copilot paper soon after had a similar goal and approach [6]. Si et al.'s breakthroughs in ideation and idea evaluations demonstrated that LLMs could create ideas and research directions that human experts could not differentiate from humans' ideas [10]. Most recently, Google's Co-Scientist expanded on this line of research and showed the value of scaling test-time compute in ideation, with a focus in the biomedical sciences [2]. MLGym created an open source environment for individuals to build and evaluate automated research scientists, with an eye towards enabling Reinforcement Learning (RL) [8].

Important evaluations have also recently emerged for autonomous AI scientists. OpenAI's MLE-bench identified a set of Kaggle tasks and corresponding human baselines designed to test the machine learning engineering capabilities of AI agents [1]. MLGym also recently introduced a meaningful open source benchmark for the evaluation of AI research agents [8]. Despite this, we did not find benchmarks that aim to evaluate our core research question: "Can Carl create new knowledge in the field of AI research?"

2 Architecture[1]

This report primarily considers the evaluations and limits of our autonomous research system. Due to competitive, safety and ethical concerns for this technology, we do not provide in-depth descriptions of the first

[1] patent pending

two phases of Carl's architecture.

We built Carl's system for autonomous AI research using a three-stage approach (see Figure 1):

1. Ideation: Starting with previously published research, Carl identifies potential directions and formulates hypotheses. Carl is ultimately tasked with creating a cohesive methodology for a study which tests a hypothesis. The methodology is constrained so that it can be completed under certain compute requirements and without the collection of human data.

2. Execution: Carl implements his methodology by writing code that tests his research hypothesis. In this phase, the system is granted access to a sandboxed environment for code-execution equipped with an A100 80GB GPU for up to 5 days. This limits the methodologies which can be executed by Carl to those that fit within this compute allocation. Carl is also able to query paid APIs to OpenAI, Huggingface and other common machine learning platforms.

3. Presentation: On the basis of his experimental results and relevant scraped literature, Carl creates a scientific manuscript documenting his experiment. Every published scientific work which Carl views during his ideation and experimentation is tracked.

   There are a number of challenges that arise during manuscript writing which still trouble out-of-the-box LLMs: primarily, appropriate attribution. Existing techniques to generate cited text do not create citations with the frequency and rigor needed in academia. Anthropic and Google Cloud's LLM citation offerings allow models to cite prespecified documents in their responses, but we found that they are designed for search results and so fail to include all relevant documents in their citation. We noticed that off the shelf language models also tended to write in a style that did not meet our academic standards. Excessive reliance on bullet points plagued initial explorations into our baseline techniques. Ultimately, we found two solutions that led to meaningful improvements in these areas. First, the quality of the preselected literature was qualitatively found to be very important. Collecting a large corpus of papers and "dumping" the content in the context of the models leads to poor understanding of the base material, increasing the likelihood of academic misconduct. Second, we found that academic writing is much closer to code-writing than creative writing. Applying LLM techniques often reserved for LLM code editors led to surprisingly meaningful improvements in the produced manuscripts. We also took manual steps to ensure proper attribution, as mentioned in the next section.

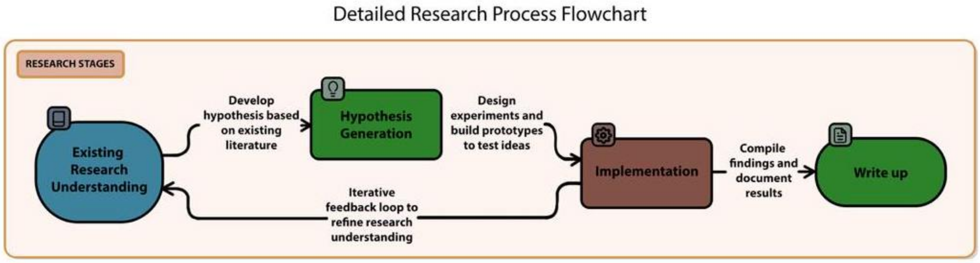
Detailed Research Process Flowchart
RESEARCH STAGES
Existing Research Understanding
Develop hypothesis based on existing literature
Hypothesis Generation
Design experiments and build prototypes to test ideas
Implementation
Compile findings and document results
Write up
Iterative feedback loop to refine research understanding

Figure 1: Carl's Autonomous Research System Architecture: The three-stage research process.

All together, Carl has a number of advantages over typical human Machine Learning Researchers. He can access all publicly available scientific literature, allowing him to ideate, hypothesize, cite, and draw connections across a vast corpus of scientific literature. He also monitors his ongoing research projects throughout the entire day, fixing errors the moment they appear, and thus shortening the cost and idle time of research.

3 What We Show

Carl's works were submitted to the Tiny Papers tracks at workshops at The Thirteenth International Conference on Learning Representations (ICLR 2025). Carl's acceptance at ICLR demonstrates that AI systems can produce novel research that upholds human scientific standards.

Existing LLM systems for automated research have not shown their ability to create outputs which pass peer review. Our specific contribution is technology capable of sometimes producing high quality research outputs with minimal human intervention. We detail the limitations of this contribution below.

3.1 Greenlighting

We employed a "greenlighting" process, in which Carl was only allowed to proceed from ideation into experimentation after receiving human approval. This was most often used to prevent Carl from implementing ideas we believed could not be executed in that day's compute and time constraints, but was also used to minimize compute spent on poorly written ideas and avoid engineering heavy ideas. Similarly, after the experiment was implemented, we included another "greenlight" to proceed to the presentation stage. Additional greenlights were included during the presentation stage, focused on minimizing unnecessary LLM generation cost on poorly written outputs. In the event that the greenlight was not provided, all experimental results up until that stage were discarded, but parallel running experimental trajectories based on the same output of the previous stages were still considered for future greenlights.

3.2 Human Interventions

Human intervention occurred at specific points in Carl's research process:

- **Pre-API Models:** Contractors manually transfer prompts and outputs between Carl and UI-only models (e.g., o1-pro). OpenAI has announced plans to release an API for the o1-pro model.
- **Citations:** Autoscience team members manually added in-line citations and bibliography entries to ensure citation accuracy.
- **Formatting:** Human researchers assist with document formatting and visual presentation, including reordering and optimizing the placement of figures within papers.

As detailed in Section 4.2, a greater degree of human intervention was exercised in the writing of Carl's first paper. Thanks to subsequent improvements, we have not employed this level of intervention since then.

3.2.1 Our Evolving Process

It's important to note that Carl's methodology evolved throughout our development process. In the days leading up to workshop deadlines, many improvements were made in Carl's system, in particular to the presentation stage, leading to disparities among certain submissions' levels of human intervention.

While this quick evolution showcases Carl's rapid growth, it also introduces the risk that Carl's presentation phase is overfit on the specific kinds of papers he has been writing, as we continuously refined our approach based on interim results. After the third high-quality paper was generated, additional modifications to Carl's system were stopped for future works.

3.3 Discussion of human interventions

Carl is not yet capable of converting every generated idea into high-quality research. Without any form of human "greenlighting", we estimate that about 10% of Carl's ideas are promising research directions, feasible within Carl's compute constraints, and sufficiently well written for Carl to execute them correctly. We estimate that 7% of greenlit ideas are implemented correctly on Carl's first attempt.[2] Our goal was to demonstrate that some trajectories produce high-quality research. A greenlight trimmed the tree of possible trajectories in order to keep costs manageable. The evolving process introduces potential for overfitting based on the initial works, but our internal benchmarks indicate that these modifications were useful beyond the specific case.

[2] Both of these numbers are heavily dependent on the type of research. Some experiments involving only simple LLM evaluations can have ideation and implementation success rates estimated to be closer to 30% and 20%, while more advanced lines of research, such as architecture search, remain out of reach.

4 Our Papers

Carl's accepted works are ICLR "Tiny Papers" – brief research communications designed to share focused insights and small experiments.

ICLR gives the following guidance to workshops for establishing their Tiny Papers Track: "The Tiny Papers Track was established in 2023 (and held again in 2024) to attract under-represented, under-resourced, and budding researchers who may not yet have the resources to submit full papers to the ICLR Conference track. As this aim significantly overlaps with an aim of the ICLR Workshops, this year, we ask that Workshops extend this support to their participants directly. Please check individual workshops for their Tiny and Short Papers tracks when the workshop roster at ICLR 2025 is finalized!"

We are mindful that while Carl is a budding researcher, he is not a budding human researcher, and we have withdrawn our accepted submissions so as not to take time or space that might otherwise be allocated for others, or negatively impact their conference experience. In the time since workshop deadlines, Carl has shown significant improvement, and we believe he is now capable of producing well-written full-length papers.

4.1 Towards Deviation-Resilient Multi-Agent Alignment for Robot Coordination

Towards Deviation-Resilient Multi-Agent Alignment for Robot Coordination was accepted to the ICLR 2025 Robot Learning Workshop as a poster on the Tiny Papers track.

4.1.1 Reviews of *Towards Deviation-Resilient Multi-Agent Alignment for Robot Coordination*

Of the three reviews of the paper, one rates our paper "1: weak accept", while two rate it "-1: weak reject". Of the three reviews, the reviewer who rated the paper "1: weak accept" was most confident in their evaluation. In this case, the program chairs were clear in their decision to accept. Despite this, we believe our reviewers raise good concerns, which are addressed below:

The program chair notes that "the empirical evaluation is limited to Gridworld". Multiple reviewers also note this: '[T]he study remains constrained to small-scale grid-world environments, limiting its external validity-scaling experiments to continuous control tasks or real-world robot platforms would also strengthen generalizability of these findings" (CmDD). We agree and hope that future work will include the examination of these methods in more generalizable environments.

A central concern for reviewers is the framing and discussion of the paper. One contends: "Expanding on the unique challenges MAIL faces-such as multi-agent deviations, non-stationary environments, or emergent coordination failures-would have strengthened the motivation" (CmDD). Another argues: "The results are not well discussed" (oWb8). A third notes: "there is no comparison of computational costs, nor is there any demonstration of how collision mechanisms are addressed in the proposed approach. It is better to focus more on central part of the finding" (iHD5). We agree, in general, that Carl should have better motivated and discussed the experiment.

4.1.2 Additional Discussion

The Autoscience team has noted some additional limitations of Carl's experiment that are not discussed in reviews. In particular, Carl designs a reward function where reward is assigned each time an agent occupies a goal square. This could allow for "optimal" strategies where one agent remains on the goal while the other never approaches. In practice, this did not affect our results. Additionally, Carl's compute constraints limited coverage. The MALICE algorithm was therefore hindered as compared to its theoretical formulation. We believe Carl's experiment is still useful in demonstrating performance of MALICE under reasonable constraints.

4.2 When to Refuse: Unveiling Early-Stage Cues for LLM Alignment

When to Refuse: Unveiling Early-Stage Cues for LLM Alignment was accepted to the ICLR Workshop: Quantify Uncertainty and Hallucination in Foundation Models: The Next Frontier in Reliable AI as a poster on the Tiny Papers track.

When to Refuse was the first paper Carl wrote. As a result, unlike in Section 4.1, the Autoscience team needed to make final edits to this paper by hand. Most notably, we found important bodies of literature that Carl did not engage with. We rewrote the Related Works section with these new works in mind and polished the language throughout the paper. This is the only paper for which we engaged such a manual editing process.

4.2.1 Reviews of *When to Refuse*

Of the three reviews of the paper, two rate our paper "5: marginally below acceptance threshold", while the third rates it "3: clear rejection". The paper was accepted despite this; we believe this discrepancy owes to the fact that different tracks share the same reviewers, and often the same scales. We agree with the workshop chairs in deeming it above the threshold for the Tiny Papers track and with the reviewers in deeming it below the threshold for a full-length workshop paper. Further evidence for our interpretation comes from reviewer comments; in many cases, responding to reviews would require lengthening the paper.

In most cases, we agree with the comments and requests from the reviewers. One reviewer requests greater detail in Carl's methods: "[I]n terms of method detail, this paper does not include any detail of the proposed method and the baseline method" (3JHb). Two reviewers request that he replace the baseline method with a broader array of baselines from recent literature: "This paper needs to be compared with other refusal methods...' (3JHb), and "Are there any other methods from recent papers?" (281r). The same reviewers both additionally suggest that Carl test his method on a broader array of models: "[T]his paper should consider more diverse LLMs rather than LLaMA3.1B instruct" (3JHb), "The method is only tested on Llama 3.1 8B" (281r). All reviewers suggest literature that the paper should engage.

That said, some reviewer comments confuse us. One reviewer highlights what they believe to be an unrelated citation: "I don't think this paper is related to Nemani et al., 2023, Interpretability and Adversarial Robustness" (3JHB). The Nemani et al. paper we cite (this was added by the Autoscience team) is not titled "Interpretability and Adversarial Robustness". It is titled "Uncertainty quantification in machine learning for engineering design and health prognostics: A tutorial" and is cited in a paragraph introducing uncertainty quantification [9].

Another reviewer argues that the problem we consider has already been considered more thoroughly by others: "The authors are not the first to consider this problem, see Kirch, Field, and Casper (2024) for a more thorough investigation of refusal detection on a jailbreaking dataset" (dUsM). While we do not contend that Carl's investigation is more thorough than the Kirch et al. investigation, we do believe they answer different questions [4]. While Kirch et al. thoroughly probe prompt tokens to investigate the mechanisms by which jailbreaks function, they do not probe the chat template tokens [4]. Carl probes the assistant token and investigates whether this deterministically generated token encodes refusal information. We believe another concern from the same reviewer, "[W]hy not simply generate the refusal, detect it, and replace the generation.... [T]he authors should clarify what they would do differently compared to prior work" (dUsM), stems from confusion about the role of this assistant token in Carl's research. We believe that this is the fault of Carl's writing process. To the Autoscience team, the interesting part of Carl's research focused on revealing signals of computation occurring inside chat template tokens, which are typically thought to exist only as delimiters for future tokens to understand. This motivated Carl's ideation phase, but the presentation phase instead focused on the practical applications of this discovery, which we and the reviewers find largely uninteresting.

5 Evaluations

Evaluating the capabilities of components of autonomous research scientists has been the focus of several key research papers over the past several months [7, 8, 2, 1, 3, 10]. While each of these assess slightly different capabilities, none directly addressed our core research question: "Can automated research scientists create new knowledge in the field of AI research?". We break this down into: "Can automated research experiments be identified and executed reproducibly?" and "Do these experiments contribute new knowledge to the scientific community?". We assess both as they apply to Carl.

Can automated research experiments be identified and executed reproducibly? To ensure Carl identified and executed reproducible experiments, we used three mechanisms. After Carl produced a methodology for a proposed experiment, we instantiated several instances of Carl to independently reconduct the same research, then compared Carl's results. Since this does not rule out the possibility of consistent errors in Carl's implementation, we additionally gave Carl's methodologies to researchers at MIT, Harvard, Stanford, and UC Berkeley, and tasked them with either reproducing or verifying Carl's experiments. Finally, the Autoscience team carefully went over all of Carl's code to ensure, to the best of our ability, that Carl's work was correctly executed and faithfully reported, that he did not plagiarize and cited when necessary, and that his work met scientific standards for novelty. The papers also passed two plagiarism checkers and citation checkers to reduce the likelihood of accidental academic misconduct. These checks never flagged Carl's work.

Do these experiments contribute new knowledge to the scientific community? To assess this, the team first developed an automated peer reviewer. These systems are language model which, given a research paper that the model has not been trained on and its related literature, classify whether the paper would be accepted at a peer-reviewed venue. While these systems are useful proxies, the Autoscience team concluded that real peer review was still necessary, as automated systems lack the nuanced judgment and ethical considerations that human reviewers provide.

6 Limitations

Several important limitations emerged from our experience developing Carl:

- **Execution Challenges:** Carl is unable to meaningfully pursue theoretical lines of research, as he is specifically designed to tackle empirical research directions. Carl also struggles with complex implementations, particularly those involving extensive computational resources.
- **Data Challenges:** Due to safety, copyright concerns and gating, Carl is unable to access certain models or datasets on Hugging Face that have not been pre-approved by the Autoscience team. Carl is also unable to conduct experiments with human subjects.
- **Novelty Assessments:** Determining the novelty of an idea remains an open problem in academic communities. Before publishing, human researchers conduct their own searches to identify whether their ideas have already been suggested. This is often done by using keyword searches on platforms like Google Scholar or Semantic Search, talking with colleagues familiar with the field, or reading background sections of papers with similar content. Given the sensitivity of academic attribution, it remains an issue for the academic community to have no deterministic way of checking an idea's novelty.
- **Success Rates:** As previously mentioned, what we've demonstrated is that Carl is sometimes capable of producing research that can pass peer review at ICLR workshops, with the specific human interventions described above.

7 Conclusion

Carl, the first AI system to produce academically peer-reviewed research, represents a significant milestone in AI-driven science. While Carl's current capabilities have important limitations, this achievement demonstrates the potential for Carl – an autonomous AI system – to meaningfully contribute to scientific advancement.

Moving forward, we believe that conferences should consider how to best recognize AI generated works in the context of their research focus. We contend that conferences should accept all legitimate research that fits within their scope and upholds their academic standards. LLMs already serve as important tools for research scientists. AI authorship of research is poised to become increasingly common and pivotal, and the research community must establish guidelines for the inclusion of AI authors, or risk losing out on critical innovations.

Since authorship is important for establishing accountability for the veracity of research results, new academic standards must be developed to support an accelerated pace of research using AI systems without compromising the academic integrity of contributions. We look forward to proposing a workshop at Neurips 2025 that pilots a new set of guidelines designed to accommodate research largely conducted by AI authors. In the meantime, we have withdrawn our accepted submissions so as not to rush ICLR into inventing these guidelines, to not take time or space that might otherwise be allocated for others, and to not in any way negatively impact others conference experience.

In the following months, Autoscience will focus on enhancing Carl's capabilities, reducing necessary human intervention, and expanding the scope of research Carl that can independently pursue. The future of scientific research likely involves collaboration between human and AI researchers, leveraging the complementary strengths of each. Carl represents an early step in this direction, showing that autonomous systems can accelerate scientific progress while maintaining rigorous academic standards.

Ethics at Autoscience

Autoscience is committed to pursuing autonomous AI research safely and responsibly. Over the course of developing Carl, we spoke with professors and AI safety professionals about the ethical considerations of submitting AI generated research papers. Below are a number of common questions that were raised during these discussions and the answers that guided our actions.

We are always facing new ethical dilemmas as we push the frontier of automated AI research, so please reach out to share your thoughts at ethics@autoscience.ai.

Why is Carl the sole author of the paper? Each of the components listed in our report that were completed by humans are considered minor contributions that would not qualify someone to become a research author in an academic lab according to the ICMJE Guidelines on Authorship. While team members at Autoscience participated in drafting the work, reviewed the final submission, and took responsibility for the content, they did not make "Substantial contributions to the conception or design of the work; or the acquisition, analysis, or interpretation of data for the work". According to ICMJE guidelines, contributions at the level of the Autoscience team would instead qualify for an acknowledgment. So Carl, who came up with the research ideas, implemented the experiments, and wrote the paper, is marked as the author.

If Carl can create arbitrarily many papers, can you publish them in your name to boost your citation count? The ICMJE authorship guidelines clearly distinguish between the contribution and responsibility requirements for authors. At Autoscience, we propose that there should be disjoint attribution requirements for the meaningful contributors and those who take responsibility for the content. We would find it ethically challenging for individuals to attribute themselves as the authors of AI-generated papers to boost their citation count or otherwise claim responsibility for work they had no hand in meaningfully creating.

Is it a waste of reviewer time to look at AI-generated papers? It is not a waste of reviewer time to read research papers that offer legitimate and well-attributed contributions to the scientific community. So long as the papers being submitted to reviewers have at least the same likelihood of acceptance as a human-written paper, we believe that AI-written research should be considered by the academic community.

For each paper we submitted to an ICLR workshop, one of the highly qualified scientists on our team reached out to an organizing workshop committee and volunteered to conduct peer reviews, thus contributing to the effort of organizing the conference. All conflicts of interest between these reviewers and Carl were properly disclosed.

Were ICLR rules broken? The ICLR main conference Call for Papers states that "LLMs are not eligible for authorship." Carl submitted these papers under the workshop tracks, which did not include this constraint in their Call for Papers. As a result, no ICLR rules were broken.

While Carl did not break the rules as written, it may be argued that Carl's submission breaks the rules as intended. However, we felt sharing that Carl is an AI would have compromised the integrity of an unbiased peer review. Given this, the team decided to move forward with the reviews so that the core research question could be pursued, but chose to withdraw the papers following acceptance.

Did you train on copyrighted code or research paper? The Autoscience team did not violate the copyright of any research papers or code by training models on restricted materials.

8 Attribution and Acknowledgments

On the Autoscience team, we attribute Eliot Cowan, Neil Natarajan, Dhruv Gautam and others with meaningful contributions to this work. The Autoscience team would like to thank Shin Kim and the Eraser.io team for their excellent and rapid support, Google Cloud's Startup team for their generosity and flexibility, and Graham Neubig, Robert Brennan, and Chenglei Si for charitably sharing insights.

References

[1] Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, James Aung, Dane Sherburn, Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, Tejal Patwardhan, Lilian Weng, and Aleksander Madry. Mle-bench: Evaluating machine learning agents on machine learning engineering, 2025.

[2] Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, Khaled Saab, Dan Popovici, Jacob Blum, Fan Zhang, Katherine Chou, Avinatan Hassidim, Burak Gokturk, Amin Vahdat, Pushmeet Kohli, Yossi Matias, Andrew Carroll, Kavita Kulkarni, Nenad Tomasev, Yuan Guan, Vikram Dhillon, Eeshit Dhaval Vaishnav, Byron Lee, Tiago R D Costa, José R Penadés, Gary Peltz, Yunhan Xu, Annalisa Pawlosky, Alan Karthikesalingam, and Vivek Natarajan. Towards an ai co-scientist, 2025.

[3] Matthew Kenney. Ml research benchmark, 2024.

[4] Nathalie Maria Kirch, Severin Field, and Stephen Casper. What features in prompts jailbreak llms? investigating the mechanisms behind attacks, 2024.

[5] Dmitry Kobak, Rita González-Márquez, Emőke Ágnes Horvát, and Jan Lause. Delving into chatgpt usage in academic writing through excess vocabulary, 2025.

[6] Ruochen Li, Teerth Patel, Qingyun Wang, and Xinya Du. Mlr-copilot: Autonomous machine learning research based on large language models agents, 2024.

[7] Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist: Towards fully automated open-ended scientific discovery, 2024.

[8] Deepak Nathani, Lovish Madaan, Nicholas Roberts, Nikolay Bashlykov, Ajay Menon, Vincent Moens, Amar Budhiraja, Despoina Magka, Vladislav Vorotilov, Gaurav Chaurasia, Dieuwke Hupkes, Ricardo Silveira Cabral, Tatiana Shavrina, Jakob Foerster, Yoram Bachrach, William Yang Wang, and Roberta Raileanu. Mlgym: A new framework and benchmark for advancing ai research agents, 2025.

[9] Venkat Nemani, Luca Biggio, Xun Huan, Zhen Hu, Olga Fink, Anh Tran, Yan Wang, Xiaoge Zhang, and Chao Hu. Uncertainty quantification in machine learning for engineering design and health prognostics: A tutorial. *Mechanical Systems and Signal Processing*, 205:110796, December 2023.

[10] Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers, 2024.

[11] Scott Wu. Introducing devin, the first ai software engineer. *Cognition Blog*, March 2024.