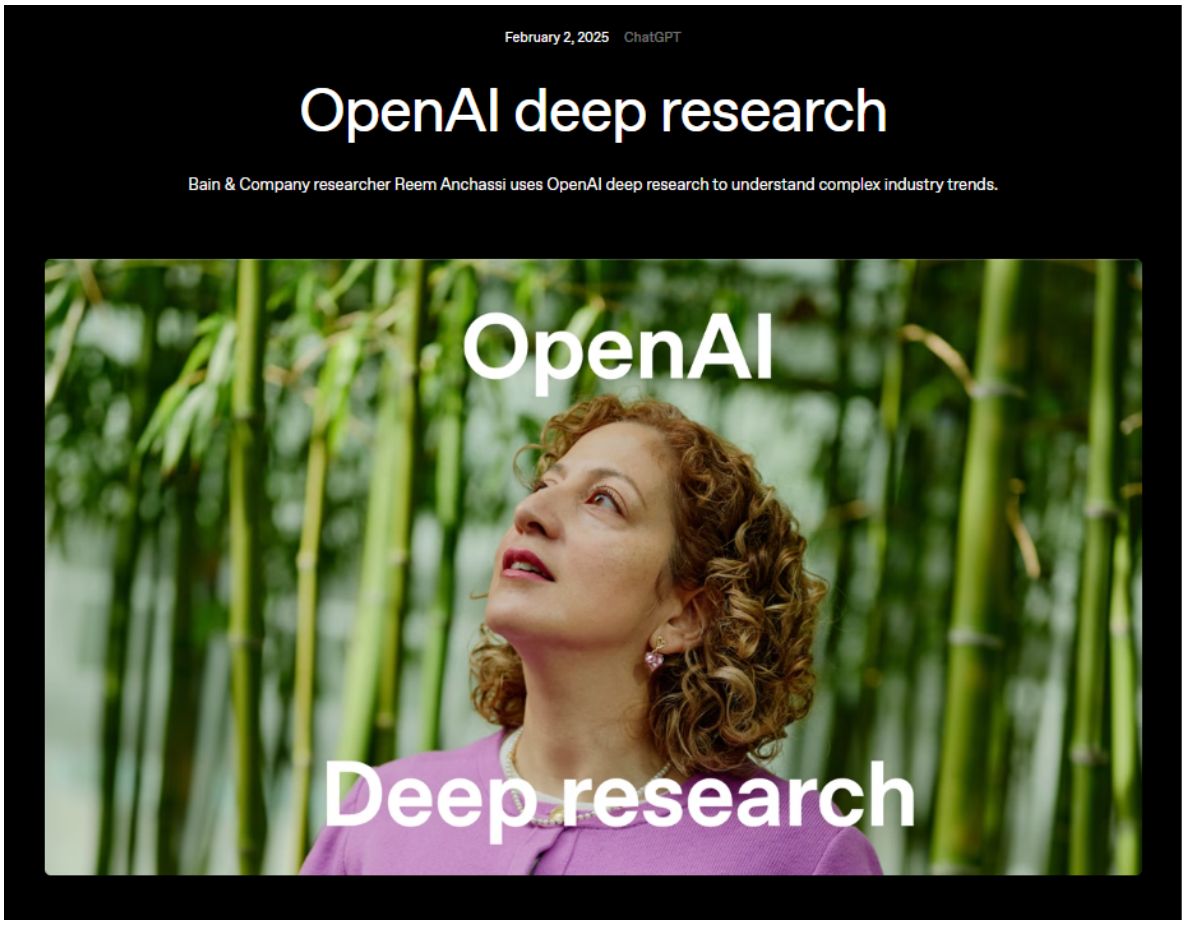
Masterclass in Social Research

AI for Social Research: Computational and Quantitative



LLMs and “interviews” (Geiecke and Jaravel 2024)

Conversations at Scale: Robust AI-led Interviews with a Simple Open-Source Platform*

Friedrich Geiecke
LSE
f.c.geiecke@lse.ac.uk

Xavier Jaravel
LSE & CEPR
x.jaravel@lse.ac.uk

November 20, 2024

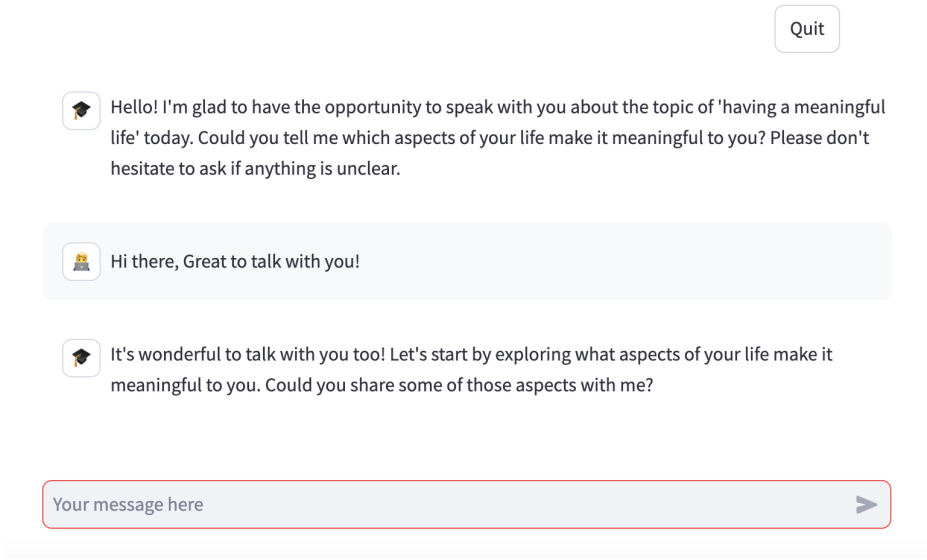
Abstract

The advent of large language models (LLMs) provides an opportunity to conduct qualitative interviews at a large scale, with thousands of respondents, creating a bridge between qualitative and quantitative methods. In this paper, we develop a simple, versatile open-source platform for researchers to run AI-led qualitative interviews. Our approach incorporates established best practices from the sociology literature, uses only a single LLM agent with low latency, and can be adapted to new interview topics almost instantaneously. We assess its robustness by drawing comparisons to human experts and using several respondents-based quality metrics. Its versatility is illustrated through four broad classes of applications: eliciting key factors in decision making, political views, views of the external world, and subjective mental states. High performance ratings are obtained in all of these domains. The platform is easy to use and deploy: we provide detailed explanations and code for researchers to swiftly set up and test their own AI-led interviews. In addition, we develop, validate, and share a simple LLM-based pipeline for textual analysis and coding of large volumes of interview transcripts.

Keywords: qualitative interviews; large language models; surveys.

- Use "single" LLM agent (OpenAI’s GPT-4o) to carry out interviews on different topics, such as
  - Meaning of life
  - Political preferences (over the 2024 French snap election)
  - Educational and occupational choices
- Evaluate LLM peformance
  - Compare AI-generated transcripts to human experts (in this case, sociologists)
  - Solicit assessment from respondents, e.g. how well the content of the interview captures their views (or "empathetic") and whether they would prefer to participate in an interview with AI or human researchers
- Why AI-led/assisted interviews might be a good idea
  - Scale and cost
  - Privacy and sensitivity bias

Figure 1 Chat Interface



Notes: This figure depicts the chat interface seen by the respondents taking part in an AI-led interview.

Table II Quality Metrics for the AI-led Interview on Meaning in Life

Panel A: Perceived quality of interview process, survey responses

	Fraction of Respondents
In the future, would you rather take the interview with	
... An AI	43%
... A human	19%
... I do not mind	38%
Would you have preferred to answer open-ended questions instead?	
... Yes	12%
... No	76%
... I do not mind	12%

Panel B: Perceived quality of interview content, survey responses

	AI-Led Interview (1)	Open Text Fields (2)
How well does it summarize what gives you a sense of meaning?	3.58	3.45
1 ("poorly") to 4 ("very well")	(s.e. 0.045)	(s.e. 0.039)
Are you able to clearly identify sources of meaning in your life?		
My thoughts are still evolving	34%	42%
I can clearly pinpoint sources of meaning in my life	52%	41%
I am somewhere in between	14%	17%
Number of words	460 (+142%)	190

Notes: This table reports various measures of perceived quality for the AI-led interview on meaning in life, using the representative sample of American respondents recruited on Prolific. Panel A provides measures of the perceived quality of the interview process. Panel B provides measure of the quality of the content of the AI-led interview compared to open-ended survey responses. Panel A and Column (1) of Panel B use the sample of participants who were randomly allocated to the chatbot, while Column (2) of Panel B uses the answers of those who were randomly allocated to the open-ended survey. The total number of respondents is 466.

LLMs and quantitative social research

- Survey and experiments: Using computational tools to design survey and simulate responses (see the NORC piece)
- Natural language processing (NLP): Using computational tools to analyze unstructured text, image or audio data (more later in Masterclass)
- Agent-based modeling (ABM): Using computational tools to model and similate the social world and human behaviors (e.g., segregation)

PA

Out of One, Many: Using Language Models to Simulate Human Samples

Lisa P. Argyle[1], Ethan C. Busby[1], Nancy Fulda[2], Joshua R. Gubler[1], Christopher Rytting[2] and David Wingate[2]

[1] *Department of Political Science, Brigham Young University, Provo, UT, USA. e-mail: lpargyle@byu.edu, ethan.busby@byu.edu, jgub@byu.edu*
[2] *Department of Computer Science, Brigham Young University, Provo, UT, USA. e-mail: nfulda@cs.byu.edu, christophermichaelrytting@gmail.com, wingated@cs.byu.edu*

Abstract

We propose and explore the possibility that language models can be studied as effective proxies for specific human subpopulations in social science research. Practical and research applications of artificial intelligence tools have sometimes been limited by problematic biases (such as racism or sexism), which are often treated as uniform properties of the models. We show that the "algorithmic bias" within one such tool—the GPT-3 language model—is instead both fine-grained and demographically correlated, meaning that proper conditioning will cause it to accurately emulate response distributions from a wide variety of human subgroups. We term this property *algorithmic fidelity* and explore its extent in GPT-3. We create "silicon samples" by conditioning the model on thousands of sociodemographic backstories from real human participants in multiple large surveys conducted in the United States. We then compare the silicon and human samples to demonstrate that the information contained in GPT-3 goes far beyond surface similarity. It is nuanced, multifaceted, and reflects the complex interplay between ideas, attitudes, and sociocultural context that characterize human attitudes. We suggest that language models with sufficient algorithmic fidelity thus constitute a novel and powerful tool to advance understanding of humans and society across a variety of disciplines.

Keywords: artificial intelligence, machine learning, computational social science, public opinion

Generative Agents: Interactive Simulacra of Human Behavior

Joon Sung Park
Stanford University
Stanford, USA
joonspk@stanford.edu

Joseph C. O'Brien
Stanford University
Stanford, USA
jobrien3@stanford.edu

Carrie J. Cai
Google Research
Mountain View, CA, USA
cjcai@google.com

Meredith Ringel Morris
Google DeepMind
Seattle, WA, USA
merrie@google.com

Percy Liang
Stanford University
Stanford, USA
pliang@cs.stanford.edu

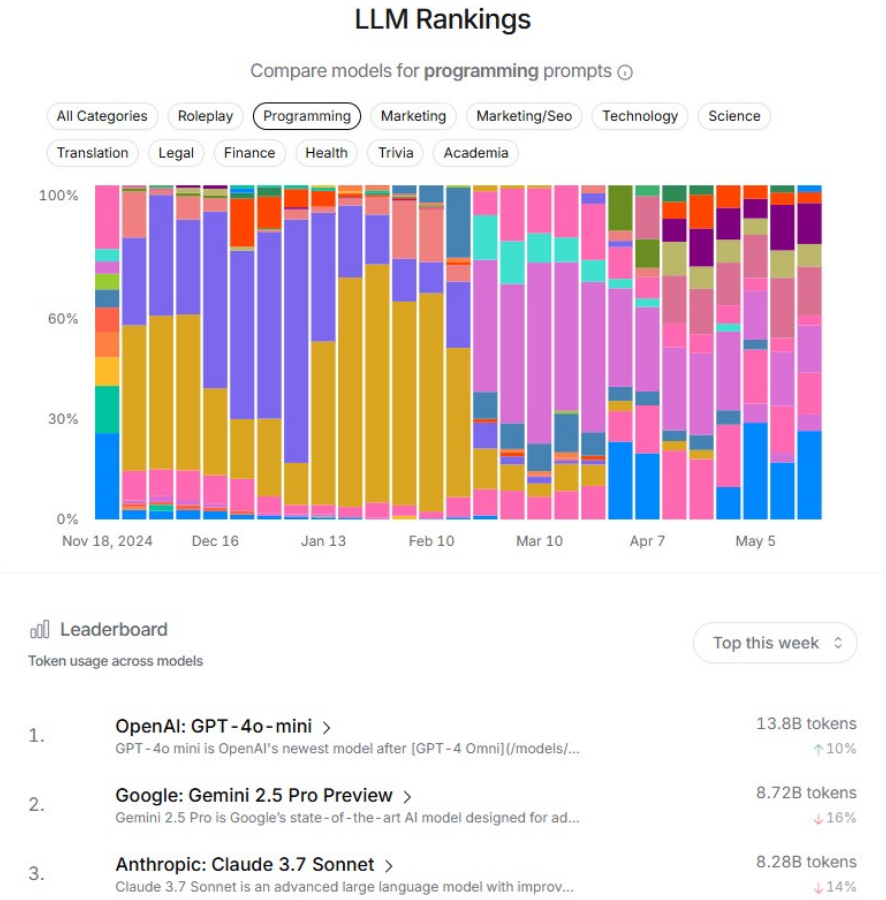
Michael S. Bernstein
Stanford University
Stanford, USA
msb@cs.stanford.edu

Demo: https://reverie.herokuapp.com/arXiv_Demo/

LLMs for coding and programming

- Claude Code (https://www.anthropic.com/claude-code)
- Google Gemini Code Assist (https://codeassist.google/)
- OpenAI Codex (https://chatgpt.com/codex)

LLM rankings (Openrouter)



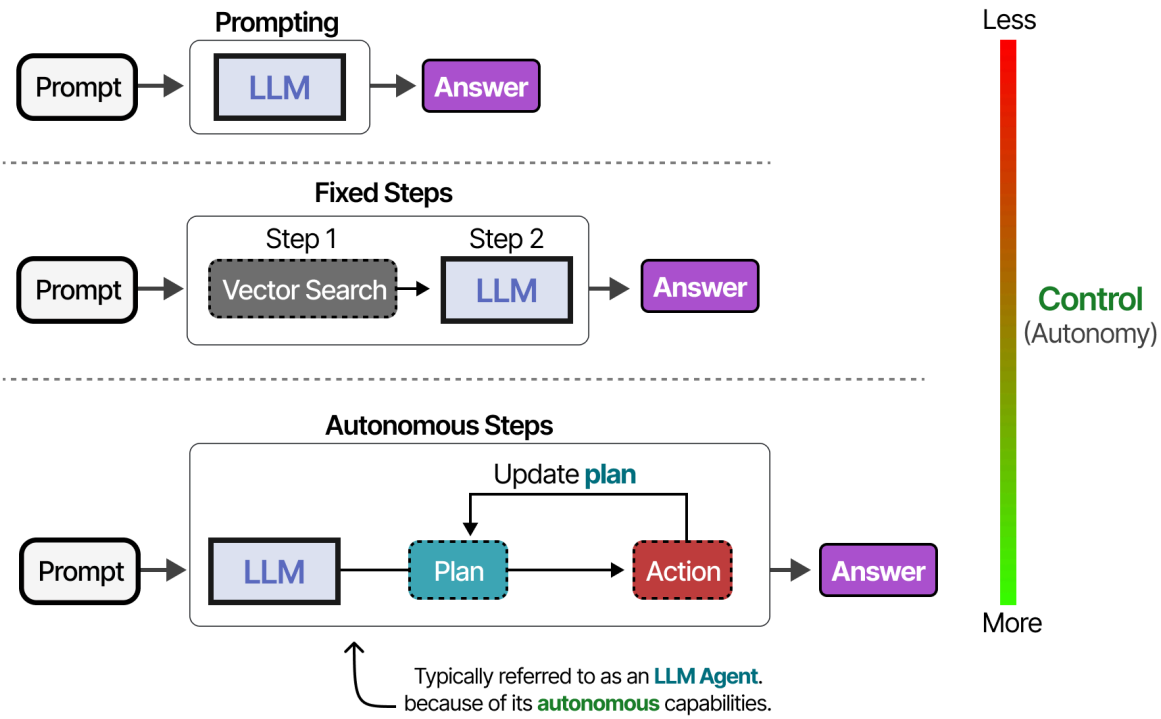
Leaderboard

Token usage across models

Top this week

1.	**OpenAI: GPT-4o-mini** > GPT-4o mini is OpenAI's newest model after [GPT-4 Omni](/models/...	13.8B tokens ↑10%
2.	**Google: Gemini 2.5 Pro Preview** > Gemini 2.5 Pro is Google's state-of-the-art AI model designed for ad...	8.72B tokens ↓16%
3.	**Anthropic: Claude 3.7 Sonnet** > Claude 3.7 Sonnet is an advanced large language model with improv...	8.28B tokens ↓14%

AI "agents"



https://code.visualstudio.com/api/extension-guides/language-model

Exercise: Simulating survey responses

- Form a group of three to four students so you can discuss findings among yourselves
- Use LLMs to carry out a "mock" survey
  - **Step 1**: Choose a question from YouGov (https://yougov.co.uk/topics/overview/trackers)
  - **Step 2**: Ask LLM the same question to see if they return similar responses (you can try to ask if they have any reference for their answers)
  - **Step 3**: Create different "respondents" using YouGov's demographic breakdowns
  - **Step 4**: Ask LLM the same question to see if they return similar responses based on the demograpic profile; repeat **Step 4** using different demograpic profiles
- Questions: How reliable are AI-generated responses, using survey data as the benchmark? How do you compare the performance of different LLMs? Does the topic you choose matter?

Alternative surveys for exercise

- British Social Attitudes (https://natcen.ac.uk/british-social-attitudes)
- Pew Research Center (https://www.pewresearch.org/datasets/)
- World Value Survey (https://www.worldvaluessurvey.org/wvs.jsp)

- ... and more

Political Analysis (2024), **32**, 401–416
doi:10.1017/pan.2024.5

PA

ARTICLE

Synthetic Replacements for Human Survey Data? The Perils of Large Language Models

James Bisbee, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel and Jennifer M. Larson

Political Science Department, Vanderbilt University, Nashville, TN, USA

Corresponding author: James Bisbee; Email: james.h.bisbee@vanderbilt.edu

(Received 2 May 2023; revised 18 January 2024; accepted 20 January 2024; published online 17 May 2024)

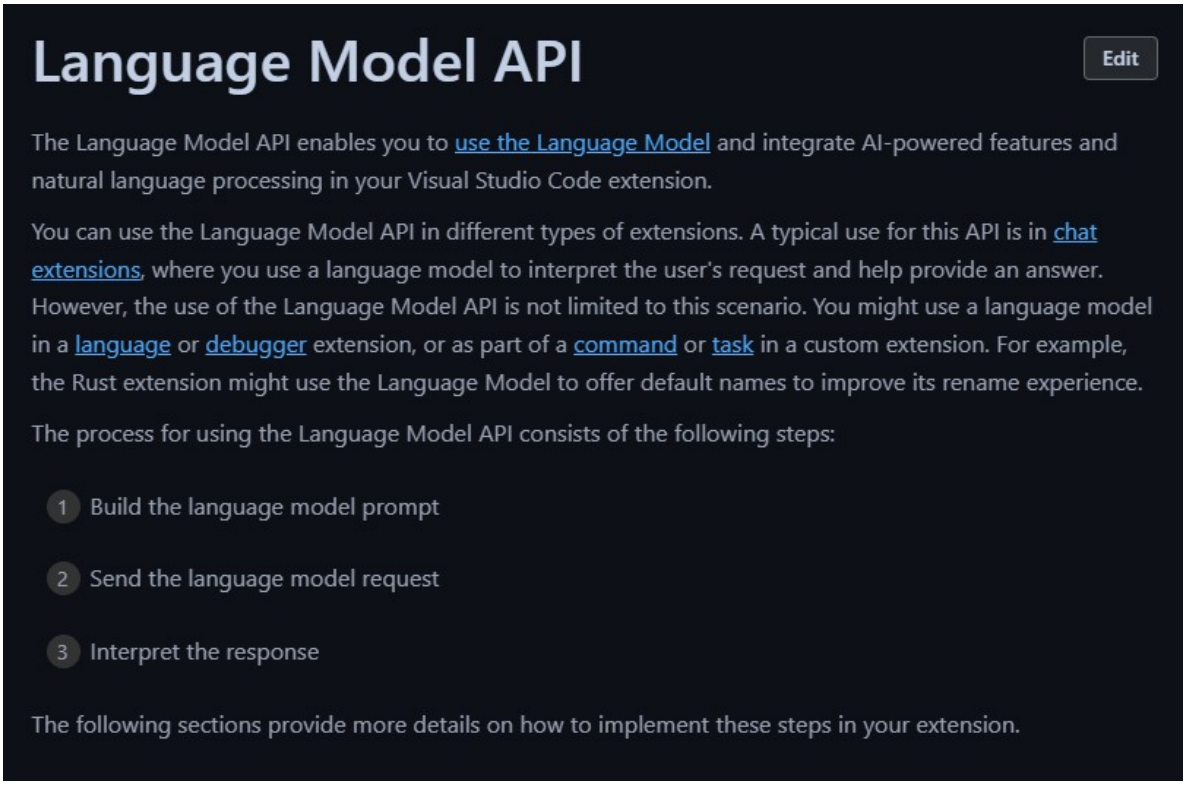
Abstract

Large language models (LLMs) offer new research possibilities for social scientists, but their potential as "synthetic data" is still largely unknown. In this paper, we investigate how accurately the popular LLM ChatGPT can recover public opinion, prompting the LLM to adopt different "personas" and then provide feeling thermometer scores for 11 sociopolitical groups. The average scores generated by ChatGPT correspond closely to the averages in our baseline survey, the 2016–2020 American National Election Study (ANES). Nevertheless, sampling by ChatGPT is not reliable for statistical inference: there is less variation in responses than in the real surveys, and regression coefficients often differ significantly from equivalent estimates obtained using ANES data. We also document how the distribution of synthetic responses varies with minor changes in prompt wording, and we show how the same prompt yields significantly different results over a 3-month period. Altogether, our findings raise serious concerns about the quality, reliability, and reproducibility of synthetic survey data generated by LLMs.

Keywords: ChatGPT; synthetic data; public opinion; research ethics

Edited by: Jeff Gill

Accessing LLMs using API

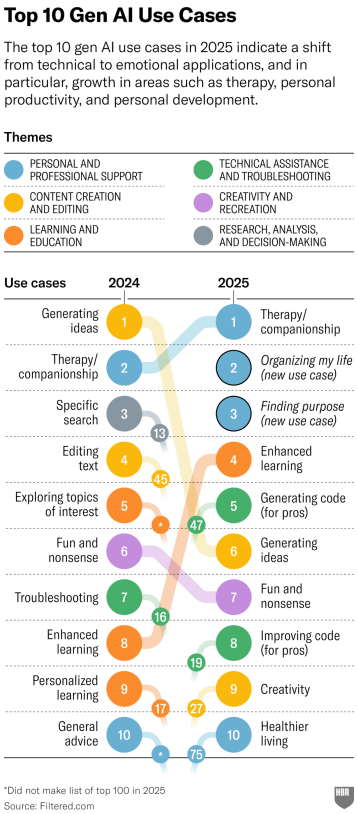


https://code.visualstudio.com/api/extension-guides/language-model

Concluding remarks

- The past three weeks only allow us to scratch the surface, and yet we have touched upon how GenAI can be used for three major research tasks: Analysis, coding and similation
- Learn how to use AI tools creatively to increase your research productivity; critical thinking remains crucial and should not undermine the agency of human researchers
- AI hype may as well be a bubble if we look at it again in a decade – use it with caution (and make sure you talk to your)
- Transparency is key: Explain how and why you use AI carefully in your research so your peers can evaluate the quality/rigour of your work

“How People Are Really Using Gen AI in 2025” (Havard Business Review)



https://hbr.org/2025/04/how-people-are-really-using-gen-ai-in-2025

TECHNOLOGY

‘The Worst Internet-Research Ethics Violation I Have Ever Seen’

The most persuasive “people” on a popular subreddit turned out to be a front for a secret AI experiment.

By Tom Bartlett

Illustration by The Atlantic

MAY 2, 2025 SHARE SAVE

MIT Economics

Home | News | Assuring An Accurate Research Record

Assuring an accurate research record

May 16th, 2025

Following the posting of the preprint paper "Artificial Intelligence, Scientific Discovery, and Product Innovation" on arXiv in November 2024, concerns were raised about the integrity of the research. MIT conducted an internal, confidential review and concluded that the paper should be withdrawn from public discourse.

In an effort to correct the research record, MIT has contacted arXiv to formally request that the paper be withdrawn and *The Quarterly Journal of Economics*, where it had been submitted. The letter on behalf of the Committee on Discipline to arXiv states:

"Earlier this year, the COD conducted a confidential internal review based upon allegations it received regarding certain aspects of this paper. While student privacy laws and MIT policy prohibit the disclosure of the outcome of this review, we are writing to inform you that MIT has no confidence in the provenance, reliability or validity of the data and has no confidence in the veracity of the research contained in the paper. Based upon this finding, we also believe that the inclusion of this paper in arXiv may violate arXiv's Code of Conduct.

Artificial Intelligence, Scientific Discovery, and Product Innovation*

Aidan Toner-Rodgers†
MIT

December 25, 2024

This paper studies the impact of artificial intelligence on innovation, exploiting the randomized introduction of a new materials discovery technology to 1,018 scientists in the R&D lab of a large U.S. firm. AI-assisted researchers discover 44% more materials, resulting in a 39% increase in patent filings and a 17% rise in downstream product innovation. These compounds possess more novel chemical structures and lead to more radical inventions. However, the technology has strikingly disparate effects across the productivity distribution: while the bottom third of scientists see little benefit, the output of top researchers nearly doubles. Investigating the mechanisms behind these results, I show that AI automates 57% of "idea-generation" tasks, reallocating researchers to the new task of evaluating model-produced candidate materials. Top scientists leverage their domain knowledge to prioritize promising AI suggestions, while others waste significant resources testing false positives. Together, these findings demonstrate the potential of AI-augmented research and highlight the complementarity between algorithms and expertise in the innovative process. Survey evidence reveals that these gains come at a cost, however, as 82% of scientists report reduced satisfaction with their work due to decreased creativity and skill underutilization.