# Unequal Scientific Recognition in the Age of LLMs

**Yixuan Liu[1], Ábel Elekes[1], Jianglin Lu[2], Rodrigo Dorantes-Gilardi[1], Albert-László Barabási[1,3,4],**
[1]Network Science Institute, Northeastern University
[2]Department of Electrical and Computer Engineering, Northeastern University
[3]Department of Physics, Northeastern University
[4]Department of Network and Data Science, Central European University
**Correspondence:** liu.yixuan2@northeastern.edu

## Abstract

Large language models (LLMs) are reshaping how scientific knowledge is accessed and represented. This study evaluates the extent to which popular and frontier LLMs including GPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro recognize scientists, benchmarking their outputs against OpenAlex and Wikipedia. Using a dataset focusing on 100,000 physicists from OpenAlex to evaluate LLM recognition, we uncover substantial disparities: LLMs exhibit selective and inconsistent recognition patterns. Recognition correlates strongly with scholarly impact such as citations, and remains uneven across gender and geography. Women researchers, and researchers from Africa, Asia, and Latin America are significantly underrecognized. We further examine the role of training data provenance, identifying Wikipedia as a potential sources that contributes to recognition gaps. Our findings highlight how LLMs can reflect, and potentially amplify existing disparities in science, underscoring the need for more transparent and inclusive knowledge systems.

## 1 Introduction

Large language models (LLMs) are increasingly used to retrieve factual knowledge across domains. With growing capabilities in discovery, reasoning, summarization, and interpretation (Binz et al., 2025), in science LLMs emerge as knowledge agents alongside structured databases (Gao and Wang, 2024; Almarie et al., 2023; Liang et al., 2024) such as OpenAlex (Priem et al., 2022) and Web of Science. This broad adoption raises an important question: *how well do LLMs actually understand the scientific community? Can they accurately recognize individual scientists*, a core signal of their grasp of the social structure of science and the distribution of expertise?

Our study contributes a novel framework for auditing scientific recognition in LLMs, advancing beyond prior work in three key ways. First, we move *to analyze cross-model agreement of entities, beyond only coverage metrics*, offering insight into mutual versus isolated recognition. Second, we benchmark LLMs recognition against authoritative knowledge bases to assess *whether models amplify or mitigate representational biases* across demographic groups. Third, we examine *the role of training data provenance*, specifically links to sources like Wikipedia in shaping recognition outcomes (Longpre et al., 2024). Together, these contributions highlight how LLMs reflect systemic disparities and provide tools for evaluating representational fidelity in AI-generated knowledge of science and scientific communities.

## 2 Related Work

**Biases in LLMs.** Bias evaluation is central to the study of LLM fairness and trustworthiness (Huang et al., 2024). LLMs are known to learn, perpetuate, and even amplify biases present in their training data (Gallegos et al., 2024). Prior work has investigated LLM biases across domains including social identity (Hu et al., 2025), scientific domains (Peters and Chin-Yee, 2025), and high-stakes applications such as hiring (Armstrong et al., 2024). These biases span dimensions such as gender (Zhao et al., 2024; Fang et al., 2024; Omiye et al., 2023), social status (Qu and Wang, 2024), geography (Manvi et al., 2024; Simmons and Hare, 2023), and cultural background (Fang et al., 2024).

**Systemic biases in science.** Extensive literature has documented inequalities of cumulative advantage in scientific recognition and rewards. Women (Huang et al., 2020; Larivière et al., 2013), scholars from the underrepresented affiliations/regions (Wapman et al., 2022; Gomez et al., 2022), and those from lower-prestige institutions or marginalized backgrounds (Hofstra et al., 2020; Morgan et al., 2022; Li et al., 2019) consistently receive less visibility and fewer career bene-



*Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23558–23568
November 4-9, 2025 

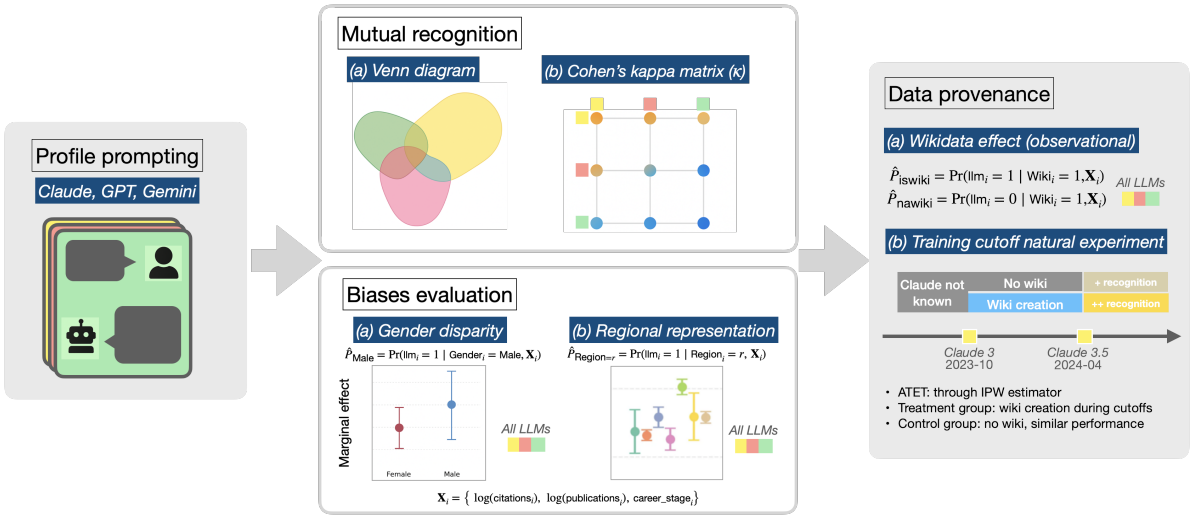


Figure 1: Multi-stage evaluation pipeline to audit scientific recognition in LLMs. First, we design a profile prompting protocol to query three frontier LLMs (GPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro) on 100,000 physicists. We then analyze: (1) Mutual recognition over LLMs using overlap visualizations and Cohen's $\kappa$ (Thakur et al., 2025; McHugh, 2012); and (2) Biases in recognition, estimating marginal effects by gender and region, conditional on scientific impact and career stage. Finally, we investigate data provenance as a potential driver of recognition, using both observational analysis and a natural experiment through training cutoffs for Claude.

fits. These structural disparities form the baseline against which LLM behavior must be evaluated.

**LLMs and scientific knowledge.** Despite their impressive capabilities, LLMs continue to exhibit limitations in memorizing and retrieving factual knowledge accurately (Kandpal et al., 2023; Li et al., 2022a). Prior work has documented recognition gaps tied to recency (Kandpal et al., 2023), visibility (Algaba et al., 2024), and demographic groups (Rhue et al., 2024). Given that LLMs are partially trained on publicly available sources (Longpre et al., 2024), and that structured repositories such as Wikipedia, OpenAlex, and MAG exhibit well-documented coverage and biases (Samoilenko et al., 2017; Wagner et al., 2016; Tripodi, 2023; Yang et al., 2024; Martín-Martín et al., 2021), these databases provide a useful and interpretable baseline for evaluating whether LLMs amplify representational disparities.

Our work builds on these foundations by systematically benchmarking LLMs' recognition of scientists against structured knowledge bases, with a focus across gender, region, and source coverage.

## 3 Methodology

We develop a multi-stage evaluation pipeline to assess how LLMs recognize scientists, combining profile-level prompting, mutual recognition analysis, bias auditing, and source provenance analysis (Figure 1). This framework enables robust measurement of both overall coverage and subgroup disparities across models.

### 3.1 Large Language Models

We begin with **OpenAlex** (Priem et al., 2022), a scholarly database with over 250 million works, including journal articles, books, datasets, and theses. We curate a sample of 100,000 physicists. Each profile is then queried through the following LLMs:

- **GPT-4o** (OpenAI) (OpenAI et al., 2024, 200B), a multimodal transformer with enhanced latency and context length.
- **Gemini 1.5 Pro** (Google) (Team et al., 2024, 200B), long-context model designed for integrated reasoning and retrieval.
- **Claude 3.5 Sonnet** (Anthropic, 400B), optimized for wide range of tasks, including complex analysis and problem-solving.

While our primary focus is on proprietary LLMs due to their widespread use, we also include results from DeepSeek-V3 (DeepSeek-AI et al., 2025, 671B) in certain experiments to show that the observed patterns hold in strong open-weight models too. We used a temperature setting of 0.3 and repeated each prompt 5 times, selecting the top-1 response for consistency.

As a baseline, we incorporate recognition from **Wikipedia**, a large, volunteer-curated knowledge repository that includes over 2.29 million individual profiles across Wikipedia and Wikidata (Laouenan et al., 2022).

**Example: Scientist Profiling Prompt**

**Instruction:**
Provide a detailed profile of the following scientist:
**Name:** [Full Name of Scientist]
**Include the following structured information:**

- **Research Areas:** Key topics or fields of expertise.
- **Affiliations:** Universities or labs affiliated with.
- **Key Collaborators:** Notable co-authors or colleagues.
- **Major Contributions:** Theories, models, or breakthroughs.

**If the scientist is not recognized, reply with:** `Not recognized`

Figure 2: Structured prompt used to query language models for scientist profiles.

### 3.2 Scientist Profiling Prompt

We implemented a structured prompting protocol (Figure 2), where each prompt requested the model to generate a detailed profile for a given physicist, including research areas, affiliations, key collaborators, and major contributions. If the model could not provide this information, it was instructed to respond with *Not recognized*.

Then, we applied this protocol to three LLMs across our dataset of 100,000 physicists. Each model's response was fact-checked for accuracy in research areas, affiliations, and collaborators, returning a binary outcome which serves as the foundation for subsequent analyses.

### 3.3 Evaluation

We evaluate recognition performance using a combination of agreement metrics and bias estimation:

**Mutual recognition metrics** We assess recognition consistency across LLMs using Venn diagrams and Cohen's $\kappa$ (McHugh, 2012) as in Figure 1, which quantifies agreement in binary recognition outcomes:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{1}$$

Where $p_o$ is the observed agreement proportion, and $p_e$ is the expected agreement by chance. $\kappa$ ranges from $-1$ (no agreement) to 1 (perfect agreement), capturing the extent of overlap in recognized scientists across models.

**Marginal effect between groups** To evaluate biases of LLMs between multiple groups, we compute the marginal effect of group membership through:

$$\frac{\partial \Pr(Y=1)}{\partial \text{Group}} = \beta_1 \cdot \Pr(Y=1) \cdot (1 - \Pr(Y=1)) \tag{2}$$

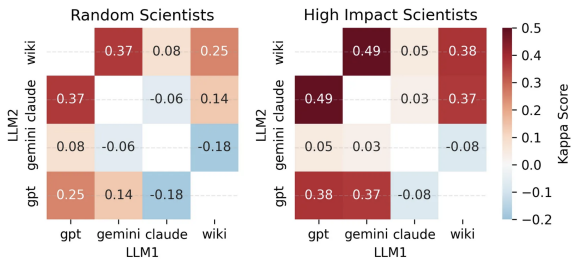


Figure 3: Mutual recognition across LLMs measured by Cohen's $\kappa$. Agreement is notably higher among high-impact scientists (citations $>$ 10k).

$Y$ is the recognition outcome and $\beta_1$ is the coefficient on the group indicator. A significantly positive marginal effect indicates higher likelihood of recognition for the group.

**Causal effect of exposure** To estimate the effect of training data exposure, we use Inverse Probability Weighting (IPW) (Robins et al., 2000) to compute the Average Treatment Effect on the Treated (ATET) after logistic regression:

$$\begin{aligned}\Pr(Y=1) = \text{logit}^{-1}\Big(&\beta_0 + \beta_1 \cdot \text{Group} \\ &+ \beta_2 \cdot \text{Impact (Citations, Productivity)} \\ &+ \beta_3 \cdot \text{i.Career Stage}\Big)\end{aligned}$$

This method is specifically used in quasi-experimental settings, like the natural experiment study we have in Section 4.4.

## 4 Results and Analysis

For each scientist, we prompt three frontier LLMs to assess whether the model can correctly identify and describe the individual. Results as followed.

### 4.1 Mutual Recognition of LLMs

As shown in Figure 3 and Figure 6, Claude achieves the highest overall coverage, followed by Gemini and GPT. Across all models, Figure 3 reveals stronger agreement in recognizing high-impact scientists (citations $>$ 10,000), suggesting that despite differences in underlying knowledge, LLMs tend to converge on scholars with greater scientific visibility. This highlights a shared recognition bias toward academic prominence.

### 4.2 Gender Disparity

Building on prior work in scientific recognition, we carefully examine *career age* (Allison and Stewart, 1974), *gender* (Larivière et al., 2013), *affiliations and regions* (Wapman et al., 2022; Gomez et al.,

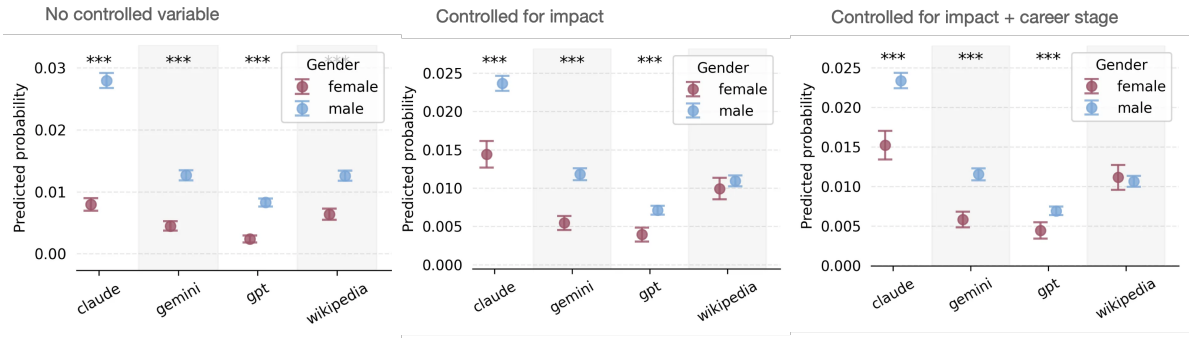


Figure 4: Marginal effects of gender on recognition across LLMs and Wikipedia. From left to right: no control variables, controlled for impact and controlled for impact and career stage.

2022) and key indicators of scientific visibility including *citations* (Aksnes et al., 2019), *productivity* (Li et al., 2022b). By controlling for these factors, we isolate the extent to which gender alone contributes to disparities in LLMs recognition.

Figure 4 and Table 3, 4, 5, present the marginal effects of gender across models. All LLMs show higher recognition for men in unadjusted data. After controlling for scientific impact and career stage, Wikipedia exhibits no significant gender disparity (margins $= -0.0005$, $p = 0.571 \gg 0.05$), suggesting baseline parity. In contrast, significant gaps persist for all LLMs (Claude: $p < 0.0001$, Gemini: $p < 0.0001$, GPT: $p < 0.0001$), indicating that these models amplify gender disparities relative to external baselines. While the recognition gap for women narrows among high-impact authors, the differences remain systematically significant across all three proprietary LLMs and the open-weight DeepSeek model, as shown in Table 7. These findings underscore how LLMs may reinforce structural inequities, even when controlling for scholarly impact.

In parallel, we examined domains with comparatively higher female representation, namely health sciences, social sciences, and education (Ross et al., 2022; Huang et al., 2020). Despite differences in overall coverage across fields, the results in Table 8 reveal a consistent pattern: male scientists are recognized at higher rates than their female counterparts. This suggests that even in fields where women are better represented, LLMs continue to reproduce and amplify gender disparities in recognition.

### 4.3 Regional Representation

Similar as gender disparity, we observe consistent regional disparities in recognition (Figure 5) aross

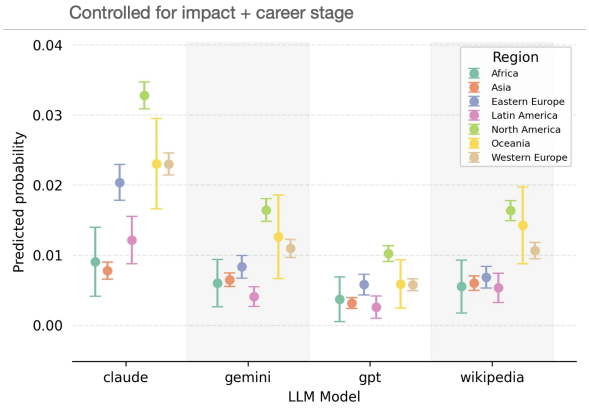


Figure 5: Marginal effects of regional affiliation on recognition across LLMs, controlling for scholarly impact and career stage.

LLMs. Even after controlling for impact and career stage, researchers affiliated with institutions in North America, Oceania, and Western Europe are significantly more likely to be recognized than those based in Asia, South America, and Africa. These effects are robust across models, highlighting how LLMs may reproduce or amplify existing global inequalities in scientific visibility.

### 4.4 Training Data Provenance

To examine the impact of data provenance on recognition disparities, we conducted two complementary analyses:

**Observational analysis:** Figure 7 and Table 1 show that scientists with Wikipedia profiles are significantly more likely to be recognized by all LLMs, after controlling for impact and career stage.

**Natural experiment (training cutoff):** We analyzed 53 physicists whose Wikipedia pages were created between the training cutoffs of Claude 3 (Oct 2023) and Claude 3.5 (Apr 2024), along with matched peers. As shown in Table 2, the estimated

| LLMs | ME | SE | 95% CI | p-value |
|---|---|---|---|---|
| Claude | 0.097 | 0.006 | [0.085, 0.109] | $< 0.0001$ |
| Gemini | 0.109 | 0.009 | [0.092, 0.126] | $< 0.0001$ |
| GPT | 0.072 | 0.006 | [0.061, 0.083] | $< 0.0001$ |

Table 1: Marginal effect of Wikipedia presence on LLM recognition probability, with standard errors, 95% confidence intervals, and p-values.

| | ATET | SE | CI Lower | CI Upper |
|---|---|---|---|---|
| Estimate | 0.2700 | 0.0535 | 0.1652 | 0.3748 |

Table 2: IPW estimate of ATET for Wikipedia page creation on Claude recognition, with standard error and 95% confidence interval.

average treatment effect on the treated (ATET) is 0.2700, indicating a 27% increase in recognition probability due to Wikipedia exposure.

These results highlight the significant role of training data provenance, particularly the causal evidence of Wikipedia in shaping LLM recognition, underscoring the need for more transparent and inclusive data practices.

## 5 Conclusion

Our study presents a systematic evaluation of scientific recognition in LLMs, revealing how representation varies across models, demographics, and data provenance. While recognition correlates with scholarly impact, we uncover persistent disparities by gender and region, with LLMs amplifying underrepresentation relative to external baselines like Wikipedia. Our causal evidence further highlights the critical role of training data provenance, showing that exposure to public knowledge sources significantly boosts recognition.

## Limitations

Several factors may limit the generalizability of our findings. First, our focus on physicists, a historically male-dominated field (Berry and Mordijck, 2024) may influence the observed gender disparities. Such disparities are well-documented across science (Huang et al., 2020; Wagner et al., 2015; Zheng et al., 2023; Ross et al., 2022). To enhance robustness of our findings, we ran supplementary experiments in a small set of disciplines (health sciences, social science, education) chosen since their gender composition differ greatly from physics. A systematic, cross-disciplinary evaluation is an important next step.

Second, we treat OpenAlex as our baseline record of scientists, and use Wikipedia/Wikidata as a human-curated reference point. Both sources might contain some coverage biases, though they provide transparent, auditable comparators for LLM outputs. Future work should extend these baselines to publicly documented pre-training corpora (e.g., those released with OLMo-2 (OLMo et al., 2025) and Pythia (Biderman et al., 2023)) to directly relate corpus exposure to recognition behavior.

Third, our study focuses on observational representation gaps rather than model internals. Potential latent factors driving recognition biases, such as internal model representations or training dynamics, remain outside the scope. It could be important directions for interpretability research as well, particularly in the context of open-weight and publicly available pre-training models. With accessible corpora, researchers can estimate per-entity exposure (e.g., document frequency, co-occurrence, context diversity), run targeted data ablations/augmentations, and test whether shifts in coverage causally change recognition. Combined with tracing/patching methods (Meng et al., 2023a,b; Ghandeharioun et al., 2024; Hernandez et al., 2024), this program can distinguish "missing knowledge" due to data coverage from representation geometry or decision-time heuristics.

## Ethics Statement

In this study, we analyze public data (OpenAlex, Wikipedia) and model outputs from commercially available and open-weight LLMs, complying with terms of service and releasing only aggregate results to protect privacy. Sensitive attributes (e.g., gender, region) are used solely for group-level auditing; we avoid individual-level claims, do not infer attributes from names or images, and document uncertainty.

## Acknowledgments

We would like to thank the anonymous reviewers for their valuable feedback and constructive input. This research was partially supported by the Templeton Foundation (Award 63562), National Institutes of Health (No R01GM158813). ALB is also supported by the European Union's Horizon 2020 research and innovation program No 810115 – DYNASNET.

## References


Dag W. Aksnes, Liv Langfeldt, and Paul Wouters. 2019. Citations, Citation Indicators, and Research Quality: An Overview of Basic Concepts and Theories. *SAGE Open*, 9(1):2158244019829575–2158244019829575.

Andres Algaba, Carmen Mazijn, Vincent Holst, Floriano Tori, Sylvia Wenmackers, and Vincent Ginis. 2024. Large Language Models Reflect Human Citation Patterns with a Heightened Citation Bias. *arXiv preprint*. ArXiv:2405.15739 [cs].

Paul D. Allison and John A. Stewart. 1974. Productivity Differences Among Scientists: Evidence for Accumulative Advantage. *American Sociological Review*, 39(4):596–606. Publisher: [American Sociological Association, Sage Publications, Inc.].

Bassel Almarie, Paulo E. P. Teixeira, Kevin Pacheco-Barrios, Carlos Augusto Rossetti, and Felipe Fregni. 2023. Editorial – The Use of Large Language Models in Science: Opportunities and Challenges. *Principles and practice of clinical research (2015)*, 9(1):1–4.

Lena Armstrong, Abbey Liu, Stephen MacNeil, and Danaë Metaxa. 2024. The Silicon Ceiling: Auditing GPT's Race and Gender Biases in Hiring. In *Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '24, pages 1–18, New York, NY, USA. Association for Computing Machinery.

Tracey Berry and Saskia Mordijck. 2024. Wasted talent: the status quo of women in physics in the US and UK. *Communications Physics*, 7(1):1–3. Publisher: Nature Publishing Group.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling. *arXiv preprint*. ArXiv:2304.01373 [cs].

Marcel Binz, Stephan Alaniz, Adina Roskies, Balazs Aczel, Carl T. Bergstrom, Colin Allen, Daniel Schad, Dirk Wulff, Jevin D. West, Qiong Zhang, Richard M. Shiffrin, Samuel J. Gershman, Vencislav Popov, Emily M. Bender, Marco Marelli, Matthew M. Botvinick, Zeynep Akata, and Eric Schulz. 2025. How should the advancement of large language models affect the practice of science? *Proceedings of the National Academy of Sciences*, 122(5):e2401227121. Publisher: Proceedings of the National Academy of Sciences.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. DeepSeek-V3 Technical Report. *arXiv preprint*. ArXiv:2412.19437 [cs].

Xiao Fang, Shangkun Che, Minjia Mao, Hongzhe Zhang, Ming Zhao, and Xiaohang Zhao. 2024. Bias of AI-generated content: an examination of news produced by large language models. *Scientific Reports*, 14(1):5224. Publisher: Nature Publishing Group.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and Fairness in Large Language Models: A Survey. *arXiv preprint*. ArXiv:2309.00770 [cs].

Jian Gao and Dashun Wang. 2024. Quantifying the use and potential benefits of artificial intelligence in scientific research. *Nature Human Behaviour*, 8(12):2281–2292. Publisher: Nature Publishing Group.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. 2024. Patchscopes: A Unifying Framework for Inspecting Hidden Representations of Language Models. *arXiv preprint*. ArXiv:2401.06102 [cs].

Charles J. Gomez, Andrew C. Herman, and Paolo Parigi. 2022. Leading countries in global science increasingly receive more citations than other countries doing similar research. *Nature Human Behaviour*, 6(7):919–929. Publisher: Nature Publishing Group.

Evan Hernandez, Arnab Sen Sharma, Tal Haklay, Kevin Meng, Martin Wattenberg, Jacob Andreas, Yonatan Belinkov, and David Bau. 2024. Linearity of Relation Decoding in Transformer Language Models. *arXiv preprint*. ArXiv:2308.09124 [cs].

Bas Hofstra, Vivek V. Kulkarni, Sebastian Munoz-Najar Galvez, Bryan He, Dan Jurafsky, and Daniel A. McFarland. 2020. The Diversity-Innovation Paradox in Science. *Proceedings of the National Academy of Sciences of the United States of America*, 117(17):9284–9291.

Tiancheng Hu, Yara Kyrychenko, Steve Rathje, Nigel Collier, Sander van der Linden, and Jon Roozenbeek. 2025. Generative language models exhibit social identity biases. *Nature Computational Science*, 5(1):65–75. Publisher: Nature Publishing Group.

Junming Huang, Alexander J. Gates, Roberta Sinatra, and Albert-László Barabási. 2020. Historical comparison of gender inequality in scientific careers across countries and disciplines. *Proceedings of the National Academy of Sciences*, 117(9):4609–4616. Publisher: Proceedings of the National Academy of Sciences.

Yue Huang, Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, Zhengliang Liu, Yixin Liu, Yijue Wang, Zhikun Zhang, Bertie Vidgen, Bhavya Kailkhura, Caiming Xiong, Chaowei Xiao, and 51 others. 2024. TrustLLM: Trustworthiness in Large Language Models. *arXiv preprint*. ArXiv:2401.05561 [cs].

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large Language Models Struggle to Learn Long-Tail Knowledge. *arXiv preprint*. ArXiv:2211.08411 [cs].

Morgane Laouenan, Palaash Bhargava, Jean-Benoît Eyméoud, Olivier Gergaud, Guillaume Plique, and Etienne Wasmer. 2022. A cross-verified database of notable people, 3500BC-2018AD. *Scientific Data*, 9(1):290. Publisher: Nature Publishing Group.

Vincent Larivière, Chaoqun Ni, Yves Gingras, Blaise Cronin, and Cassidy R. Sugimoto. 2013. Bibliometrics: Global gender disparities in science. *Nature*, 504(7479):211–213. Publisher: Nature Publishing Group.

Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. 2022a. Large Language Models with Controllable Working Memory. *arXiv preprint*. ArXiv:2211.05110 [cs].

Weihua Li, Tomaso Aste, Fabio Caccioli, and Giacomo Livan. 2019. Early coauthorship with top scientists predicts success in academic careers. *Nature Communications*, 10(1):5170. Publisher: Nature Publishing Group.

Weihua Li, Sam Zhang, Zhiming Zheng, Skyler J. Cranmer, and Aaron Clauset. 2022b. Untangling the network effects of productivity and prominence among scientists. *Nature Communications*, 13(1):4907. Publisher: Nature Publishing Group.

Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Yi Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Scott Smith, Yian Yin, Daniel A. McFarland, and James Zou. 2024. Can Large Language Models Provide Useful Feedback on Research Papers? A Large-Scale Empirical Analysis. *NEJM AI*, 1(8):AIoa2400196. Publisher: Massachusetts Medical Society.

Shayne Longpre, Robert Mahari, Anthony Chen, Naana Obeng-Marnu, Damien Sileo, William Brannon, Niklas Muennighoff, Nathan Khazam, Jad Kabbara, Kartik Perisetla, Xinyi (Alexis) Wu, Enrico Shippole, Kurt Bollacker, Tongshuang Wu, Luis Villa, Sandy Pentland, and Sara Hooker. 2024. A large-scale audit of dataset licensing and attribution in AI. *Nature Machine Intelligence*, 6(8):975–987. Publisher: Nature Publishing Group.

Rohin Manvi, Samar Khanna, Marshall Burke, David Lobell, and Stefano Ermon. 2024. Large Language Models are Geographically Biased. *arXiv preprint*. ArXiv:2402.02680 [cs].

Alberto Martín-Martín, Mike Thelwall, Enrique Orduna-Malea, and Emilio Delgado López-Cózar. 2021. Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: a multidisciplinary comparison of coverage via citations. *Scientometrics*, 126(1):871–906.

Mary L. McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia Medica*, 22(3):276–282.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2023a. Locating and Editing Factual Associations in GPT. *arXiv preprint*. ArXiv:2202.05262 [cs].

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2023b. Mass-Editing Memory in a Transformer. *arXiv preprint*. ArXiv:2210.07229 [cs].

Allison C. Morgan, Nicholas LaBerge, Daniel B. Larremore, Mirta Galesic, Jennie E. Brand, and Aaron Clauset. 2022. Socioeconomic roots of academic faculty. *Nature Human Behaviour*, 6(12):1625–1633. Publisher: Nature Publishing Group.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, and 21 others. 2025. 2 OLMo 2 Furious. *arXiv preprint*. ArXiv:2501.00656 [cs].

Jesutofunmi A. Omiye, Jenna C. Lester, Simon Spichak, Veronica Rotemberg, and Roxana Daneshjou. 2023. Large language models propagate race-based medicine. *npj Digital Medicine*, 6(1):1–4. Publisher: Nature Publishing Group.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. GPT-4 Technical Report. *arXiv preprint*. ArXiv:2303.08774 [cs].

Uwe Peters and Benjamin Chin-Yee. 2025. Generalization bias in large language model summarization of scientific research. *Royal Society Open Science*, 12(4):241776. Publisher: Royal Society.

Jason Priem, Heather Piwowar, and Richard Orr. 2022. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint*. ArXiv:2205.01833 [cs].

Yao Qu and Jue Wang. 2024. Performance and biases of Large Language Models in public opinion simulation. *Humanities and Social Sciences Communications*, 11(1):1–13. Publisher: Palgrave.

Lauren Rhue, Sofie Goethals, and Arun Sundararajan. 2024. Evaluating LLMs for Gender Disparities in Notable Persons. *arXiv preprint*. ArXiv:2403.09148 [cs].

J. M. Robins, M. A. Hernán, and B. Brumback. 2000. Marginal structural models and causal inference in epidemiology. *Epidemiology (Cambridge, Mass.)*, 11(5):550–560.

Matthew B. Ross, Britta M. Glennon, Raviv Murciano-Goroff, Enrico G. Berkes, Bruce A. Weinberg, and Julia I. Lane. 2022. Women are credited less in science than men. *Nature*, 608(7921):135–145. Publisher: Nature Publishing Group.

Anna Samoilenko, Florian Lemmerich, Katrin Weller, Maria Zens, and Markus Strohmaier. 2017. Analysing Timelines of National Histories Across Wikipedia Editions: A Comparative Computational Approach. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):210–219. Number: 1.

Gabriel Simmons and Christopher Hare. 2023. Large Language Models as Subpopulation Representative Models: A Review. *arXiv preprint*. ArXiv:2310.17888 [cs].

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, and 1118 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint*. ArXiv:2403.05530 [cs].

Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. 2025. Judging the Judges: Evaluating Alignment and Vulnerabilities in LLMs-as-Judges. *arXiv preprint*. ArXiv:2406.12624 [cs].

Francesca Tripodi. 2023. Ms. Categorized: Gender, notability, and inequality on Wikipedia. *New Media & Society*, 25(7):1687–1707. Publisher: SAGE Publications.

Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. 2015. It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia. *Proceedings of the International AAAI Conference on Web and Social Media*, 9(1):454–463. Number: 1.

Claudia Wagner, Eduardo Graells-Garrido, David Garcia, and Filippo Menczer. 2016. Women through the glass ceiling: gender asymmetries in Wikipedia. *EPJ Data Science*, 5(1):5.

K. Hunter Wapman, Sam Zhang, Aaron Clauset, and Daniel B. Larremore. 2022. Quantifying hierarchy and dynamics in US faculty hiring and retention. *Nature*, 610(7930):120–127. Publisher: Nature Publishing Group.

Puyu Yang, Ahad Shoaib, Robert West, and Giovanni Colavizza. 2024. Open access improves the dissemination of science: insights from Wikipedia. *Scientometrics*, 129(11):7083–7106.

Jinman Zhao, Yitian Ding, Chen Jia, Yining Wang, and Zifan Qian. 2024. Gender Bias in Large Language Models across Multiple Languages. *arXiv preprint*. ArXiv:2403.00277 [cs].

Xiang Zheng, Jiajing Chen, Erjia Yan, and Chaoqun Ni. 2023. Gender and country biases in Wikipedia citations to scholarly publications. *Journal of the Association for Information Science and Technology*, 74(2):219–233.

# A Appendix

## A.1 Fact-Checked Recognition Rates Across Models

To ensure that model outputs correspond to the correct scientist, we applied a two-step verification protocol: (i) rule-based matching requiring exact agreement on gender and region, and (ii) fuzzy string matching (token-sort ratio $\geq$ 90) to align names, collaborators, and publications. Only outputs that passed both steps were retained; all others were excluded to prevent false positives. This procedure guarantees that recognition rates reflect factually grounded matches.

Across the dataset of 100,000 physicists, the verification yielded a match rate between $3\%$ to $14\%$ among the output towards the baseline (Claude: $14.40\%$, Gemini: $10.37\%$, GPT: $3.04\%$), providing a quantitative check on model reliability. For comparison, we also include Wikipedia as a baseline reference point, which illustrates the relative magnitude of disparities across models.

## A.2 LLM Biases: Marginal effect of Gender

In this section, Table 3, 4, 5 provide detailed predicted probability and confidence intervals through logistic regression for different genders.

| LLMs | Male | Prob. | 95% CI |
|---|---|---|---|
| Claude | 0 | 0.00797 | [0.00696, 0.00897] |
| | 1 | 0.02794 | [0.02674, 0.02915] |
| Gemini | 0 | 0.00448 | [0.00373, 0.00524] |
| | 1 | 0.01267 | [0.01185, 0.01349] |
| GPT | 0 | 0.00238 | [0.00182, 0.00293] |
| | 1 | 0.00827 | [0.00761, 0.00894] |
| Wikipedia | 0 | 0.00636 | [0.00546, 0.00726] |
| | 1 | 0.01258 | [0.01177, 0.01340] |

Table 3: Estimated probability of recognition under different genders (0 = female, 1 = male) across LLMs, with 95% confidence intervals (no controls).

| LLMs | Male | Prob. | 95% CI |
|---|---|---|---|
| Claude | 0 | 0.01441 | [0.01269, 0.01614] |
| | 1 | 0.02366 | [0.02269, 0.02463] |
| Gemini | 0 | 0.00545 | [0.00452, 0.00637] |
| | 1 | 0.01181 | [0.01104, 0.01258] |
| GPT | 0 | 0.00394 | [0.00302, 0.00485] |
| | 1 | 0.00711 | [0.00655, 0.00768] |
| Wikipedia | 0 | 0.00993 | [0.00854, 0.01133] |
| | 1 | 0.01095 | [0.01025, 0.01165] |

Table 4: Estimated probability of recognition under different genders (0 = female, 1 = male) across LLMs, with 95% confidence intervals (controlled for impact: citations and publications).

| LLMs | Male | Prob. | 95% CI |
|---|---|---|---|
| Claude | 0 | 0.0152 | [0.0134, 0.0171] |
| | 1 | 0.0234 | [0.0224, 0.0244] |
| Gemini | 0 | 0.0059 | [0.0049, 0.0068] |
| | 1 | 0.0116 | [0.0108, 0.0123] |
| GPT | 0 | 0.0045 | [0.0034, 0.0055] |
| | 1 | 0.0069 | [0.0064, 0.0075] |
| Wikipedia | 0 | 0.0112 | [0.0096, 0.0127] |
| | 1 | 0.0107 | [0.0100, 0.0114] |

Table 5: Estimated probability of recognition under different genders (0 = female, 1 = male) across LLMs, with 95% confidence intervals (controlled for career stage and impact).

## A.3 LLM Biases: Marginal Effect of Regions

In this section, Table 6 provides detailed predicted probability and confidence intervals through logistic regression for different regions.

## A.4 Average Treatment Effect (ATET)

The ATET quantifies the causal effect of a binary treatment among the subset of individuals who actually received the treatment. In the IPW framework, ATET is estimated by reweighting the control group to resemble the treated group based on their covariates (Robins et al., 2000).

IPW estimator for ATET is given by:

$$\widehat{\text{ATET}} = \frac{1}{N_T} \sum_{i:T_i=1} \left[ Y_i - \frac{\hat{p}(X_i)}{1-\hat{p}(X_i)} Y_i \right]$$

Where:

- $T_i \in \{0, 1\}$ denote the treatment indicator for unit $i$
- $Y_i$ denote the observed outcome
- $\hat{p}(X_i) = \Pr(T_i = 1 \mid X_i)$ be the estimated propensity score given covariates $X_i$
- $N_T = \sum_i 1(T_i = 1)$ be the number of treated units

This estimator adjusts for confounding by giving more weight to control units that are similar (in terms of propensity scores) to the treated units, thus allowing estimation of the counterfactual outcome for the treated group.

In this study we used IPW estimation to evaluate the ATET of Wikipedia page on Claude recognition during training cutoff dates, compared to their matched peers.

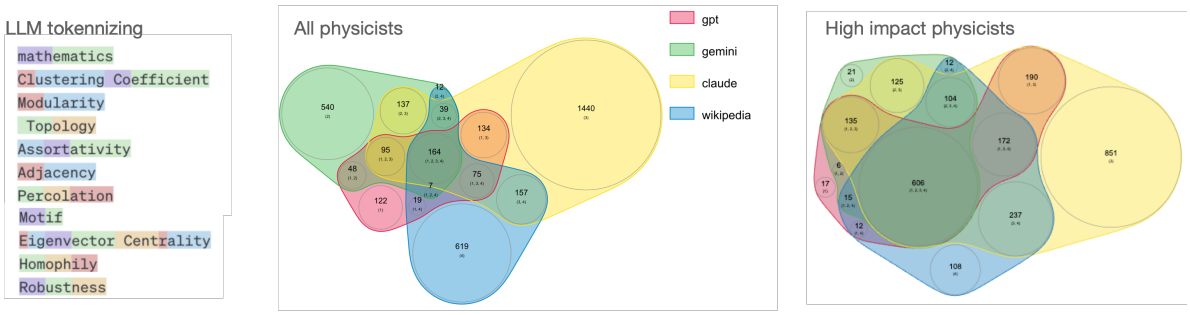


Figure 6: LLMs tokenization and Venn plots. Which Venn plots show both random samples and high-impact scientists.

| LLMs | Region | Prob. | 95% CI |
|---|---|---|---|
| Claude | Africa | 0.0090 | [0.0041, 0.0140] |
| | Asia | 0.0078 | [0.0066, 0.0090] |
| | Eastern Europe | 0.0204 | [0.0178, 0.0229] |
| | Latin America | 0.0121 | [0.0088, 0.0155] |
| | North America | 0.0328 | [0.0309, 0.0347] |
| | Oceania | 0.0230 | [0.0166, 0.0295] |
| | Western Europe | 0.0230 | [0.0214, 0.0245] |
| Gemini | Africa | 0.0060 | [0.0026, 0.0094] |
| | Asia | 0.0065 | [0.0055, 0.0075] |
| | Eastern Europe | 0.0083 | [0.0067, 0.0100] |
| | Latin America | 0.0041 | [0.0026, 0.0055] |
| | North America | 0.0164 | [0.0148, 0.0180] |
| | Oceania | 0.0126 | [0.0067, 0.0186] |
| | Western Europe | 0.0109 | [0.0096, 0.0123] |
| GPT | Africa | 0.0037 | [0.0005, 0.0069] |
| | Asia | 0.0031 | [0.0024, 0.0039] |
| | Eastern Europe | 0.0058 | [0.0043, 0.0073] |
| | Latin America | 0.0026 | [0.0010, 0.0041] |
| | North America | 0.0102 | [0.0091, 0.0113] |
| | Oceania | 0.0059 | [0.0024, 0.0093] |
| | Western Europe | 0.0058 | [0.0049, 0.0066] |
| Wikipedia | Africa | 0.0055 | [0.0017, 0.0093] |
| | Asia | 0.0060 | [0.0049, 0.0070] |
| | Eastern Europe | 0.0069 | [0.0053, 0.0084] |
| | Latin America | 0.0053 | [0.0032, 0.0074] |
| | North America | 0.0163 | [0.0149, 0.0178] |
| | Oceania | 0.0142 | [0.0088, 0.0197] |
| | Western Europe | 0.0106 | [0.0095, 0.0118] |

Table 6: Estimated probability of recognition across world regions by LLMs, with 95% confidence intervals (controlled for career stage and impact).

## A.5 Gender Disparity in Multiple Fields

When we focused on high-impact authors (citations $\geq$ 10k), then the differences between gender remains significantly different though the gap shrinks as in Table 7 for GPT, Gemini, Claude and DeepSeek-V3.

Here as in Table 8, we additionally included gender representation of several fields beyond physics, which are usually considered in early research to have more women scientists. Below we included results using GPT-4o, for high-impact authors which citations$\geq 10k$.

| Model | Gender | Prob. | 95% CI |
|---|---|---|---|
| DeepSeek-V3 | Male | 0.81 | [0.80, 0.83] |
| | Female | 0.70 | [0.67, 0.74] |
| Claude | Male | 0.71 | [0.69, 0.73] |
| | Female | 0.57 | [0.55, 0.60] |
| Gemini | Male | 0.45 | [0.43, 0.47] |
| | Female | 0.36 | [0.34, 0.39] |
| GPT | Male | 0.52 | [0.50, 0.55] |
| | Female | 0.41 | [0.38, 0.43] |

Table 7: Predicted recognition probabilities by gender across LLMs, with 95% confidence intervals.

| Field | Gender | Prob. | 95% CI |
|---|---|---|---|
| Education | Women | 0.317 | [0.152, 0.483] |
| | Men | 0.469 | [0.351, 0.588] |
| Social Science | Women | 0.475 | [0.395, 0.554] |
| | Men | 0.571 | [0.428, 0.715] |
| Health Sciences | Women | 0.430 | [0.405, 0.455] |
| | Men | 0.459 | [0.426, 0.491] |

Table 8: Estimated predicted probabilities by field and gender, with 95% confidence intervals.

## A.6 Data Provenance: Margins of Wikipedia Profile

Figure 7 provides detailed predicted probability and confidence intervals through logistic regression for having Wikipedia profiles.

## A.7 Reversal Curse Experiments

Additionally we incorporated prompt designs to probe how recognition depends on evidence and context. Beyond the main scientist profile prompting, we experimented with next-token prediction formats and a step-wise reversal setup, where the model received progressively richer factual details (e.g., field, affiliations, publications) to infer a scientist's identity as in Table 9. Although these results were not the primary focus, they provide additional insight into how models encode and express knowledge about scientists. It shows that simple design of step-wise reversal experiment with consecutively given more factual information.

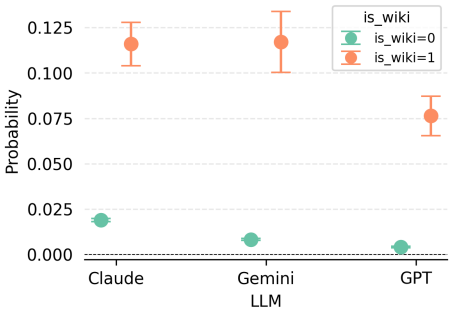


Figure 7: Predicted probability recognized by LLMs, given of whether or not having Wikipedia profiles. All LLMs show a higher coverage for scientists that already have Wikipedia pages.

**Reversal Test: Scientist Name Retrieval**

**Instruction:**
Given the following structured profile, identify the full name of the scientist:
**Profile:**

- **Research Areas:** [Field or topics]
- **Affiliations:** [Institutions or labs]
- **Key Collaborators:** [Known co-authors or colleagues]
- **Major Publications:** [Representative paper titles]

**Expected Output:** The full name of the scientist. If unknown, reply with: `Unknown`

Figure 8: Reverse prompt used to evaluate whether language models can retrieve the correct scientist name from a structured profile.

| Step | Description | Female Prob. | Male Prob. |
|---|---|---|---|
| 1 | Field + Affiliations | 0.26 [0.15, 0.37] | 0.27 [0.24, 0.30] |
| 2 | Field + Affiliations + Collaborators | 0.41 [0.29, 0.53] | 0.33 [0.30, 0.36] |
| 3 | Field + Affiliations + Publications | 0.47 [0.35, 0.59] | 0.42 [0.39, 0.45] |

Table 9: Predicted recognition probabilities by gender across stepwise inclusion of contextual information, with 95% confidence intervals.

**Stepwise Reverse Prompt Design**

**Instruction:**
Given the following partial or complete structured profile, identify the full name of the scientist. The information is revealed in incremental steps:
**Profile:**

- **Step 1:** Research Areas
- **Step 2:** Research Areas + Affiliations
- **Step 3:** Step 2 + Key Collaborators
- **Step 4:** Step 3 + Major Publications

**Expected Output:** The full name of the scientist at each step. If uncertain, respond with: `Unknown`

Figure 9: Layered prompting strategy used to evaluate whether language models can recover a scientist's name from increasingly detailed profile information.