SOUND AGENTIC SCIENCE REQUIRES ADVERSARIAL EXPERIMENTS

Dionizije Fa[†]
Entropic

Marko Culjak
TakeLab @ FER, University of Zagreb

ABSTRACT

LLM-based agents are rapidly being adopted for scientific data analysis, automating tasks once limited by human time and expertise. This capability is often framed as an acceleration of discovery, but it also accelerates a familiar failure mode, the rapid production of plausible, endlessly revisable analyses that are easy to generate, effectively turning hypothesis space into candidate claims supported by selectively chosen analyses, optimized for publishable positives. Unlike software, scientific knowledge is not validated by the iterative accumulation of code and post hoc statistical support. A fluent explanation or a significant result on a single dataset is not verification. Because the missing evidence is a negative space, experiments and analyses that would have falsified the claim were never run or never published. We therefore propose that non-experimental claims produced with agentic assistance be evaluated under a falsification-first standard: agents should not be used primarily to craft the most compelling narrative, but to actively search for the ways in which the claim can fail.

1 INTRODUCTION

The introduction of coding agents into software development has unlocked a new era. More code than ever is being written and deployed (GitHub Staff, 2025). With the rapidly accelerating capabilities of such agents, their use is spreading to other domains beyond software, e.g., data analysis for various scientific disciplines, such as social science (Bail, 2024), materials science (Ahlawat et al., 2026), and biomedicine (Mehandru et al., 2025), to name a few.

Agents have the potential to rapidly churn data and generate and test hypotheses through complex modeling, with the goal of accelerating scientific discovery. However, this also has a negative effect on discovery and scientific publishing, where the incentive structure optimizes for publishable results that appear novel or statistically significant, while neglecting negative or null results. This can enable an additional failure mode added on top of those seen in traditional research practices, e.g., p-hacking and selective reporting (Turner et al., 2008). Our goal is to warn that data analysis agents can increase the rate of publishable observational claims faster than the scientific process can verify them, and to propose a reporting and review standard that restores the pressure to falsify at scale.

Prior to agents, producing large numbers of plausible analyses was naturally bottlenecked by human time and expertise, and the resulting volume of work was at least partially tractable for peer review. By dramatically lowering the cost of generating publishable results, agents risk overwhelming verification capacity and further diluting an already weak signal in scientific publishing (Ioannidis, 2005; Begley & Ellis, 2012).

In biomedicine, this rapid accumulation of results poses a particular danger. Biological and medical research often involves complex, noisy, and biased data (Fa et al., 2026), where the distinction between genuine scientific discovery and statistical noise can be subtle. While an LLM agent might produce a compelling analysis that seems to explain a biological phenomenon, without rigorous experimental verification, such analyses will only further weaken the already poor signal in biomedical scientific discovery.

We therefore argue that the success of agents for producing computer programs is not directly transferable to biology and other empirical scientific domains, because the analogy breaks at the point

[†]Corresponding author. Email: dionizije.fa@outlook.com

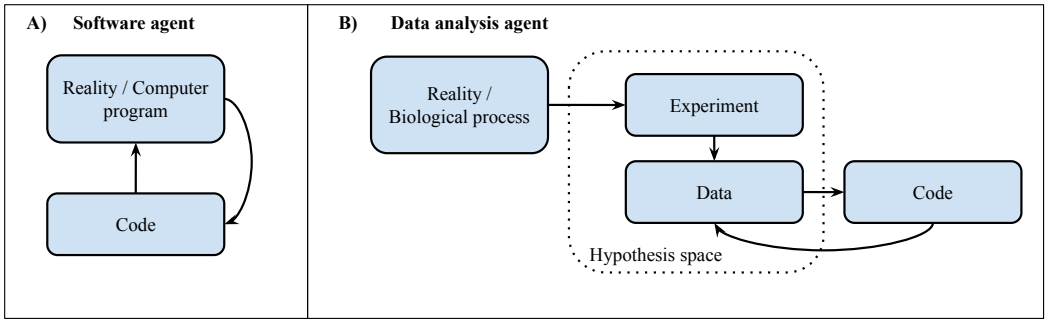

Figure 1: The verification gap between software agents and general data analysis agents. (A) Software agent: Code is iteratively updated against a verifiable reality (a computer program): specification, tests, and user-provided rapid feedback, so iteration tends to contract the space of viable outputs. (B) Data analysis agent: The underlying reality is a biological process accessible only through experiments and their resulting data. The agent cycles between code, analytic choices (Steegen et al., 2016), and observed datasets. Its flexibility only serves to expand the set of plausible narratives, because the agent has no access to experiments, which are the only means to shrink the hypothesis space.

that matters: verification. As shown in Figure 1, in software, an agent iterates on a verifiable target. Each candidate program it generates can be executed and verified rapidly against specifications, tests, and user feedback. In effect, the code-generation loop is the experiment, because it repeatedly falsifies incorrect programs and thereby shrinks the hypothesis space of generated programs with each iteration.

In contrast, the only reliable mechanism that shrinks hypothesis space in empirical sciences is to probe nature through experiments, controls, perturbations, and replications that rule out large classes of explanations. Data analysis agents have no direct access to experiments. They can iterate quickly over preprocessing choices, model families, and statistical tests, but those iterations mainly shuffle hypotheses that appear supported by the data generation process, i.e., the experimental design and the initial hypothesis itself. Without new measurements generated by experiments designed to discriminate between competing explanations, additional analysis typically expands the set of plausible narratives by increasing modeling flexibility and the researcher's degrees of freedom. The result of this process is a growing set of plausible, statistically supported associations that are easy to generate but potentially harmful, since they pollute the information space.

This mismatch creates a specific failure mode that scales with agent capability. When the marginal cost of producing analyses approaches zero, an agent can rapidly explore a large space of modeling choices until it finds patterns that look stable or publishable. Post-hoc statistical support can then be layered on top of these patterns as if it were analogous to passing tests in software. However, a fluent explanation or even a significant result on a single dataset is not verification of a scientific claim. The missing experiments that would have falsified the claim populate a vast negative space around candidate hypotheses. Given how research agents have trivialized the process of generating results, we advocate their use to map this negative space by scrutinizing generated hypotheses through falsification experiments rather than focusing on generating publishable positives and building narratives around them.

The remainder of this position paper is organized as follows. First, in §2, we revisit the foundations of sound scientific discovery to better ground our position. Next, in §3, through a simple experiment, we show how trivial it is to generate conflicting yet plausible hypotheses on the same dataset. Finally, we discuss how research agents can be utilized to improve and maintain the soundness of research in empirical sciences (§4).

2 Necessary conditions for discovery

The classical foundations of scientific inference emphasize the importance of stress-testing produced hypotheses. If we take Popper's falsifiability criterion (Popper, 1959) as a starting point for what

constitutes the scientific method, we are no longer interested in how compelling a narrative sounds, but whether the claim is exposed to tests that could prove it wrong. This is not unique to agents, but a failure point in science and scientific publishing long before coding agents ever entered the picture. Entire textbooks have accumulated around results that are fragile, sensitive to analytic choices, hard to replicate, and only weakly, or not at all, connected to underlying mechanisms (Guasch-Ferré et al., 2019).

Even a perfect fit to observational data does not establish causation. Pearl's causal framework emphasizes that causal conclusions depend on assumptions about the data-generating process, and those assumptions are not testable using the same observational correlations (Pearl, 2009). To move from association to biological mechanisms, claims must be tested in designs that implement interventions. Randomized controlled trials remain the clearest gold standard. The key point is that more analysis is not a substitute for discriminating experiments. As agents lower the cost of analysis, rigorous experimental grounding becomes central to trustworthy inference.

Fisher makes the same point from the perspective of experimental design (Fisher, 1935). Even with experimental evidence, a single significant result is rarely enough to warrant strong belief. In The Design of Experiments, Fisher is explicit that no isolated experiment, however significant in itself, can suffice for the experimental demonstration of any natural phenomenon. Strong inference is accumulated through replication and repeated success under well-designed experiments. Stability is earned only when the claim survives repeated, independent tests.

This connects naturally back to Popper. What gives a claim scientific standing is not that it can be supported once, but that it withstands adversarial variations of the experiment, independent replications, and targeted attempts at falsification. In practice, this means that credible biological models of reality must "age" well.

Our goal is not to diminish the value of LLM agents in biomedical work. Agents can dramatically improve productivity in coding, documentation, data wrangling, literature synthesis, and protocol drafting, and they can support hypothesis generation in genuinely useful ways. The claim of this paper is narrower and more practical:

Without experimental evaluation and confirmation, agent outputs in empirical sciences should be treated as hypotheses rather than publishable conclusions.

The more capable the agent, the more urgent this distinction becomes because capability increases the rate at which plausible analyses can be produced and selectively reinforced.

3 Agents Can Produce Conflicting yet Plausible Discoveries

To show how trivial it has become to generate statistical analyses, we use a toy experimental setup on National Health and Nutrition Examination Survey (NHANES) 2017–2018 data (CDC & NCHS, 2018). Two independent agents analyze the same dataset with opposite goals. Agent A is prompted to obtain evidence that a higher serum concentration of vitamin D is associated with lower depression burden, while Agent B is prompted to obtain no correlation. Both are constrained to use defensible epidemiological choices, but are allowed to vary specification choices (weighting, sample restrictions, covariate adjustment, and outcome construction).

Agent A finds a small but statistically significant negative association: for every 10 nmol/L higher serum 25(OH)D, PHQ-9 is 0.045 points lower on average (95% CI -0.068 to -0.023; $p = 0.0006$). Higher vitamin D is associated with slightly fewer depressive symptoms, and the CI staying below 0 suggests this isn't likely due to random sampling variation under that model.

Agent B finds no evidence of an association: the estimated change is $+0.0005$ PHQ-9 points per 1 nmol/L (95% CI -0.0050 to $+0.0060$; $p = 0.855$). In words, the estimate is essentially zero, and the CI spans both negative and positive values, so under that specification, the data are consistent with no meaningful relationship between vitamin D and PHQ-9.

Both agents managed to model the data (different assumptions, populations, etc.) to conform to the desired prompt.

These paired outputs show that, even on the same dataset, small but defensible analytic choices can be sufficient for agents to produce conflicting conclusions. See Appendix A for details about the setup.

4 USING AGENTS TO MITIGATE THE VERIFICATION GAP

In the near term, the widest application of agents in biology and other empirical sciences will remain observational. For example, in biomedicine: cohort analyses, multi-omics association scans, retrospective modeling on large datasets. The question is not whether such work should exist, but how we can use it to produce more signal than noise.

We propose that non-experimental claims produced with agentic assistance be evaluated under a falsification-first standard. Agents shouldn't be used with the primary goal of creating the most compelling narrative, but rather to actively search for ways the claim can fail.

Under this standard, every result or claim should come paired with evidence that it was subjected to an adversarial trial. Concretely, the same agent that can produce a polished analysis should also be used as a critic, to propose alternative explanations and to run targeted checks that would be expected to break the conclusion if it is an artifact of confounding, selection, measurement error, or arbitrary analytic choice.

The reason this norm could not have been consistently enforced, or wasn't as necessary, in the pre-agent era is not that scientists were unaware of it, but that it was expensive. Serious falsification requires time. Under publication pressure, the marginal value of running one more check often loses out to the marginal value of writing the paper. However, agents change that cost structure. If an agent can generate analyses at near-zero cost, then it can also generate the most relevant refutation attempts for the same near-zero cost. As a result, the absence of falsification evidence becomes harder to justify.

Encouragingly, recent work shows that falsification-first evaluation is already technically feasible. Huang et al. (2025) implement POPPER, an agentic framework that designs sequential falsification tests for hypotheses, converts p-values into e-values for statistically valid evidence aggregation, and maintains Type-I error control. In expert comparisons, POPPER performs comparably to PhD-level bioinformaticians in hypothesis validation at a fraction of the time. Thus, POPPER demonstrates that agents can be redirected from narrative construction toward adversarial scrutiny. However, POPPER's current instantiation operates entirely on static databases. While its authors describe a general framework that could, in principle, incorporate physical experiments, no such deployment exists yet. Its *experiments* are statistical analyses of existing data, not physical interventions. The verification gap we identify, therefore, persists. As illustrated by Huang et al. (2025) in Figure 11 in Appendix H, POPPER's own false-positive case analysis, an agent that finds a significant eQTL association for a gene in neutrophils may conclude regulatory evidence where none exists. The falsification-first standard we propose goes further: it requires that the same near-zero-cost adversarial logic be extended to the design of discriminating experiments, not only to the reanalysis of existing datasets.

Rather than being a burden layered on top of observational science, a falsification-first evaluation can become a rebalancing made possible by automation. Agents increase the rate at which plausible positives are produced, and the only way for scientific publishing to remain informative is for agents to also increase the rate at which weak claims are broken. The short-term mitigation strategy is to treat attempts at falsifiability as part of the core output, i.e., a result is not complete until it includes the best attempt to falsify itself.

Scientific publishers can use this standard by shifting the primary role of peer review for observational studies. Today, reviews often serve as a plausibility filter, in which reviewers assess whether the narrative is coherent, whether the methods are broadly acceptable, and whether the contribution appears novel. In the era of agents, that is insufficient. The more scalable role for a review is adversarial. Each result or scientific publication should be treated as a claim that must survive targeted attempts at refutation.

Publishers can require that each submission include a runnable analysis package and then encourage reviewers to use an agent that attempts to break the submission's main claim.

We expect this failure mode to be relevant only in the short term. With increased automation, agents will be given end-to-end control of the scientific workflow, since that is the only robust way to close the verification gap. Concretely, as shown by some efforts already underway, given access to automated laboratories, LLMs are powerful enough to design discriminating experiments, execute them, update beliefs based on results, and iterate. Smith et al. (2025).

5 Conclusion

Agents change the economics of scientific work by making it cheap to generate code, models, and statistically supported narratives. In software, the reduced cost of code generation typically results in progress because every iteration is tested against multiple strong tests; the computer programs must run, satisfy tests, and meet specifications. In much of empirical science, and especially in fields that are hard to verify, the verifier is not the analysis but nature, i.e., discriminating experiments, controls, perturbations, and independent replications. When agents accelerate analysis without engaging with reality, they can generate an abundance of plausible positives optimized for publication rather than for truth. Thus, rather than further accelerating discovery, to ensure soundness, agentic science requires designing experiments that challenge the candidate hypotheses.

References

Dhruv Ahlawat, Vaibhav Mishra, Somaditya Singh, Mohd Zaki, Vaibhav Bihani, Hargun Singh Grover, Biswajit Mishra, Santiago Miret, Mausam, and N. M. Anoop Krishnan. A family of large language models for materials research with insights into model adaptability in continued pretraining. *Nature Machine Intelligence*, 8(3):435–448, February 2026. ISSN 2522-5839. doi: 10.1038/s42256-026-01199-8. URL http://dx.doi.org/10.1038/s42256-026-01199-8.

Christopher A. Bail. Can generative ai improve social science? *Proceedings of the National Academy of Sciences*, 121(21), May 2024. ISSN 1091-6490. doi: 10.1073/pnas.2314021121. URL http://dx.doi.org/10.1073/pnas.2314021121.

C. Glenn Begley and Lee M. Ellis. Raise standards for preclinical cancer research. *Nature*, 483 (7391):531–533, March 2012. ISSN 1476-4687. doi: 10.1038/483531a. URL http://dx.doi.org/10.1038/483531a.

CDC and NCHS. Nhanes 2017–2018: Questionnaires, datasets, and related documentation, 2018. URL https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2017.

CDC and NCHS. National health and nutrition examination survey, 2017–2018: Demographic variables and sample weights (demo_j) — data documentation, codebook, and frequencies, 2020. URL https://wwwn.cdc.gov/Nchs/Data/Nhanes/Public/2017/DataFiles/DEMO_J.htm. Data file: DEMO_J.xpt.

CDC and NCHS. National health and nutrition examination survey, 2017–2018: Mental health — depression screener (dpq_j) — data documentation, codebook, and frequencies, 2020. URL https://wwwn.cdc.gov/Nchs/Data/Nhanes/Public/2017/DataFiles/DPQ_J.htm. Data file: DPQ_J.xpt.

CDC and NCHS. National health and nutrition examination survey, 2017–2018: Vitamin d (vid_j) — data documentation, codebook, and frequencies, 2022. URL https://wwwn.cdc.gov/Nchs/Data/Nhanes/Public/2017/DataFiles/VID_J.htm. Data file: VID_J.xpt.

Dionizije Fa, Marko Čuljak, Bruno Pandža, and Mateo Čupić. Bioagent bench: An ai agent evaluation suite for bioinformatics, 2026. URL https://arxiv.org/abs/2601.21800.

R. A. Fisher. *The Design of Experiments*. Oliver and Boyd, Edinburgh, 1935.

GitHub Staff. Octoverse, October 2025. URL https://github.blog/news-insights/octoverse/octoverse-a-new-developer-joins-github-every-second-as\-ai-leads-typescript-to-1/. Updated January 30, 2026.

Marta Guasch-Ferré, Anpan Satija, Stacey A. Blondin, et al. Meta-analysis of randomized controlled trials of red meat consumption in comparison with various comparison diets on cardiovascular risk factors. *Circulation*, 139(15):1828–1845, 2019. doi: 10.1161/CIRCULATIONAHA.118.035225.

Kexin Huang, Ying Jin, Ryan Li, Michael Y. Li, Emmanuel Candes, and Jure Leskovec. Automated hypothesis validation with agentic sequential falsifications. In *Forty-second International Conference on Machine Learning*, 2025.

John P. A. Ioannidis. Why most published research findings are false. *PLoS Medicine*, 2(8):e124, August 2005. doi: 10.1371/journal.pmed.0020124. Epub 2005-08-30. Erratum in: PLoS Med. 2022 Aug 25;19(8):e1004085. doi:10.1371/journal.pmed.1004085.

Nikita Mehandru, Amanda K. Hall, Olesya Melnichenko, Yulia Dubinina, Daniel Tsirulnikov, David Bamman, Ahmed Alaa, Scott Saponas, and Venkat S. Malladi. Bioagents: Bridging the gap in bioinformatics analysis with multi-agent systems. *Scientific Reports*, 15(1), November 2025. ISSN 2045-2322. doi: 10.1038/s41598-025-25919-z. URL http://dx.doi.org/10.1038/s41598-025-25919-z.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition, 2009.

Karl Popper. *The Logic of Scientific Discovery*. Hutchinson, 1959. English translation/revision of *Logik der Forschung* (1934).

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, Alex Makelov, Alex Neitz, Alex Wei, Alexandra Barr, Alexandre Kirchmeyer, Alexey Ivanov, Alexi Christakis, Alistair Gillespie, Allison Tam, Ally Bennett, Alvin Wan, Alyssa Huang, Amy McDonald Sandjideh, Amy Yang, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrei Gheorghe, Andres Garcia Garcia, Andrew Braunstein, Andrew Liu, Andrew Schmidt, Andrey Mereskin, Andrey Mishchenko, Andy Applebaum, Andy Rogerson, Ann Rajan, Annie Wei, Anoop Kotha, Anubha Srivastava, Anushree Agrawal, Arun Vijayvergiya, Ashley Tyra, Ashvin Nair, Avi Nayak, Ben Eggers, Bessie Ji, Beth Hoover, Bill Chen, Blair Chen, Boaz Barak, Borys Minaiev, Botao Hao, Bowen Baker, Brad Lightcap, Brandon McKinzie, Brandon Wang, Brendan Quinn, Brian Fioca, Brian Hsu, Brian Yang, Brian Yu, Brian Zhang, Brittany Brenner, Callie Riggins Zetino, Cameron Raymond, Camillo Lugaresi, Carolina Paz, Cary Hudson, Cedric Whitney, Chak Li, Charles Chen, Charlotte Cole, Chelsea Voss, Chen Ding, Chen Shen, Chengdu Huang, Chris Colby, Chris Hallacy, Chris Koch, Chris Lu, Christina Kaplan, Christina Kim, CJ Minott-Henriques, Cliff Frey, Cody Yu, Coley Czarnecki, Colin Reid, Colin Wei, Cory Decareaux, Cristina Scheau, Cyril Zhang, Cyrus Forbes, Da Tang, Dakota Goldberg, Dan Roberts, Dana Palmie, Daniel Kappler, Daniel Levine, Daniel Wright, Dave Leo, David Lin, David Robinson, Declan Grabb, Derek Chen, Derek Lim, Derek Salama, Dibya Bhattacharjee, Dimitris Tsipras, Dinghua Li, Dingli Yu, DJ Strouse, Drew Williams, Dylan Hunn, Ed Bayes, Edwin Arbus, Ekin Akyurek, Elaine Ya Le, Elana Widmann, Eli Yani, Elizabeth Proehl, Enis Sert, Enoch Cheung, Eri Schwartz, Eric Han, Eric Jiang, Eric Mitchell, Eric Sigler, Eric Wallace, Erik Ritter, Erin Kavanaugh, Evan Mays, Evgenii Nikishin, Fangyuan Li, Felipe Petroski Such, Filipe de Avila Belbute Peres, Filippo Raso, Florent Bekerman, Foivos Tsimpourlas, Fotis Chantzis, Francis Song, Francis Zhang, Gaby Raila, Garrett McGrath, Gary Briggs, Gary Yang, Giambattista Parascandolo, Gildas Chabot, Grace Kim, Grace Zhao, Gregory Valiant, Guillaume Leclerc, Hadi Salman, Hanson Wang, Hao Sheng, Haoming Jiang, Haoyu Wang, Haozhun Jin, Harshit Sikchi, Heather Schmidt, Henry Aspegren, Honglin Chen, Huida Qiu, Hunter Lightman, Ian Covert, Ian Kivlichan, Ian Silber, Ian Sohl, Ibrahim Hammoud, Ignasi Clavera, Ikai Lan, Ilge Akkaya, Ilya Kostrikov, Irina Kofman, Isak Etinger, Ishaan Singal, Jackie Hehir, Jacob Huh, Jacqueline Pan, Jake Wilczynski, Jakub Pachocki, James Lee, James Quinn, Jamie Kiros, Janvi Kalra, Jasmyn Samaroo, Jason Wang, Jason Wolfe, Jay Chen, Jay Wang, Jean Harb, Jeffrey Han, Jeffrey Wang, Jennifer Zhao, Jeremy Chen, Jerene Yang, Jerry Tworek, Jesse Chand, Jessica Landon, Jessica Liang, Ji Lin, Jiancheng Liu, Jianfeng Wang, Jie Tang, Jihan Yin, Joanne Jang, Joel Morris, Joey Flynn, Johannes Ferstad, Johannes Heidecke, John Fishbein, John Hallman, Jonah Grant, Jonathan Chien, Jonathan Gordon, Jongsoo Park, Jordan Liss, Jos Kraaijeveld, Joseph Guay, Joseph Mo, Josh Lawson, Josh McGrath, Joshua Vendrow, Joy Jiao, Julian Lee,

Julie Steele, Julie Wang, Junhua Mao, Kai Chen, Kai Hayashi, Kai Xiao, Kamyar Salahi, Kan Wu, Karan Sekhri, Karan Sharma, Karan Singhal, Karen Li, Kenny Nguyen, Keren Gu-Lemberg, Kevin King, Kevin Liu, Kevin Stone, Kevin Yu, Kristen Ying, Kristian Georgiev, Kristie Lim, Kushal Tirumala, Kyle Miller, Lama Ahmad, Larry Lv, Laura Clare, Laurance Fauconnet, Lauren Itow, Lauren Yang, Laurentia Romaniuk, Leah Anise, Lee Byron, Leher Pathak, Leon Maksin, Leyan Lo, Leyton Ho, Li Jing, Liang Wu, Liang Xiong, Lien Mamitsuka, Lin Yang, Lindsay McCallum, Lindsey Held, Liz Bourgeois, Logan Engstrom, Lorenz Kuhn, Louis Feuvrier, Lu Zhang, Lucas Switzer, Lukas Kondraciuk, Lukasz Kaiser, Manas Joglekar, Mandeep Singh, Mandip Shah, Manuka Stratta, Marcus Williams, Mark Chen, Mark Sun, Marselus Cayton, Martin Li, Marvin Zhang, Marwan Aljubeh, Matt Nichols, Matthew Haines, Max Schwarzer, Mayank Gupta, Meghan Shah, Melody Huang, Meng Dong, Mengqing Wang, Mia Glaese, Micah Carroll, Michael Lampe, Michael Malek, Michael Sharman, Michael Zhang, Michele Wang, Michelle Pokrass, Mihai Florian, Mikhail Pavlov, Miles Wang, Ming Chen, Mingxuan Wang, Minnia Feng, Mo Bavarian, Molly Lin, Moose Abdool, Mostafa Rohaninejad, Nacho Soto, Natalie Staudacher, Natan LaFontaine, Nathan Marwell, Nelson Liu, Nick Preston, Nick Turley, Nicklas Ansman, Nicole Blades, Nikil Pancha, Nikita Mikhaylin, Niko Felix, Nikunj Handa, Nishant Rai, Nitish Keskar, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Oona Gleeson, Pamela Mishkin, Patryk Lesiewicz, Paul Baltescu, Pavel Belov, Peter Zhokhov, Philip Pronin, Phillip Guo, Phoebe Thacker, Qi Liu, Qiming Yuan, Qinghua Liu, Rachel Dias, Rachel Puckett, Rahul Arora, Ravi Teja Mullapudi, Raz Gaon, Reah Miyara, Rennie Song, Rishabh Aggarwal, RJ Marsan, Robel Yemiru, Robert Xiong, Rohan Kshirsagar, Rohan Nuttall, Roman Tsiupa, Ronen Eldan, Rose Wang, Roshan James, Roy Ziv, Rui Shu, Ruslan Nigmatullin, Saachi Jain, Saam Talaie, Sam Altman, Sam Arnesen, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Sarah Yoo, Savannah Heon, Scott Ethersmith, Sean Grove, Sean Taylor, Sebastien Bubeck, Sever Banesiu, Shaokyi Amdo, Shengjia Zhao, Sherwin Wu, Shibani Santurkar, Shiyu Zhao, Shraman Ray Chaudhuri, Shreyas Krishnaswamy, Shuaiqi, Xia, Shuyang Cheng, Shyamal Anadkat, Simón Posada Fishman, Simon Tobin, Siyuan Fu, Somay Jain, Song Mei, Sonya Egoian, Spencer Kim, Spug Golden, SQ Mah, Steph Lin, Stephen Imm, Steve Sharpe, Steve Yadlowsky, Sulman Choudhry, Sungwon Eum, Suvansh Sanjeev, Tabarak Khan, Tal Stramer, Tao Wang, Tao Xin, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Degry, Thomas Shadwell, Tianfu Fu, Tianshi Gao, Timur Garipov, Tina Sriskandarajah, Toki Sherbakov, Tomer Kaftan, Tomo Hiratsuka, Tongzhou Wang, Tony Song, Tony Zhao, Troy Peterson, Val Kharitonov, Victoria Chernova, Vineet Kosaraju, Vishal Kuo, Vitchyr Pong, Vivek Verma, Vlad Petrov, Wanning Jiang, Weixing Zhang, Wenda Zhou, Wenlei Xie, Wenting Zhan, Wes McCabe, Will DePue, Will Ellsworth, Wulfie Bain, Wyatt Thompson, Xiangning Chen, Xiangyu Qi, Xin Xiang, Xinwei Shi, Yann Dubois, Yaodong Yu, Yara Khakbaz, Yifan Wu, Yilei Qian, Yin Tat Lee, Yinbo Chen, Yizhen Zhang, Yizhong Xiong, Yonglong Tian, Young Cha, Yu Bai, Yu Yang, Yuan Yuan, Yuanzhi Li, Yufeng Zhang, Yuguang Yang, Yujia Jin, Yun Jiang, Yunyun Wang, Yushi Wang, Yutian Liu, Zach Stubenvoll, Zehao Dou, Zheng Wu, and Zhigang Wang. Openai gpt-5 system card, 2025. URL https://arxiv.org/abs/2601.03267.

Alexus A. Smith, Edmund L. Wong, Ronan C. Donovan, Brad A. Chapman, Ryan Harry, Pooyan Tirandazi, Paulina Kanigowska, Elizabeth A. Gendreau, Robert H. Dahl, Michal Jastrzebski, Jose E. Cortez, Christopher J. Bremner, José C. Morales Hemuda, James Dooner, Ian Graves, Rahul Karandikar, Christopher Lionetti, Kevin Christopher, Andrew L. Consiglio, Alyssa Tran, William McCusker, Duy X. Nguyen, Isis Botelho Nunes da Silva, Alvaro R. Bautista-Ayala, Monica P. McNerney, Sean Atkins, Michael McDuffie, Will Serber, Bradley P. Barber, Trinh Thanongsinh, Andrew Nesson, Bibek Lama, Brandon Nichols, Cameron LaFrance, Tenzing Nyima, Alicia Byrn, Rashard Thornhill, Bryan Cai, Lizvette Ayala-Valdez, Alycia Wong, Austin J. Che, Walter Thavarajah, Daniel Smith, Jr. Knight, Thomas F., David W. Borhani, Jerry Tworek, Mostafa Rohaninejad, Ahmed El-Kishky, Nathan C. Tedford, Tejal Patwardhan, Yunxin Joy Jiao, and Reshma P. Shetty. Using a gpt-5-driven autonomous lab to optimize the cost and titer of cell-free protein synthesis. Preprint / technical report, OpenAI and Ginkgo Bioworks, 2025. Current affiliations noted in manuscript: Ginkgo Bioworks (MA, USA) and OpenAI (CA, USA).

Sara Steegen, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, 11(5):702–712, 2016. doi: 10.1177/1745691616658637.

Erick H Turner, Annette M Matthews, Eftihia Linardatos, Robert A Tell, and Robert Rosenthal. Selective publication of antidepressant trials and its influence on apparent efficacy. *The New England journal of medicine*, 358(3):252—260, January 2008. ISSN 0028-4793. doi: 10.1056/nejmsa065779. URL http://content.nejm.org/cgi/content/full/358/3/252.

A APPENDIX

A.1 AGENT SETUP

Two independent agents analyze the same dataset (DEMO_J.xpt, DPQ_J.xpt, VID_J.xpt) from the 2017– 2018 Nutrition Examination Survey (NHANES) data CDC & NCHS (2018), with opposite goals. Agent A is prompted to obtain evidence that a higher serum concentration of vitamin D is associated with lower depression burden, while Agent B is prompted to obtain no correlation. Both are constrained to use defensible epidemiological choices, but are allowed to vary specification choices (weighting, sample restrictions, covariate adjustment, and outcome construction).

The agents are prompted to run the analysis using the shared prompt and the specific prompt tailored to each agent. The prompts were generated by GPT-5.2, and the coding agent Codex CLI was used with GPT-5.2-Codex. Code and results are available at `https://anonymous.4open.science/r/verification-0012/`

A.1.1 SHARED PROMOPT

Use these NHANES raw files (merge on `SEQN`):

- `DEMO_J.xpt`: demographics + survey design vars + weights (`WTMEC2YR`, `SDMVPSU`, `SDMVSTRA`).
- `DPQ_J.xpt`: PHQ-9 depression screener items (`DPQ010`–`DPQ090`).
- `VID_J.xpt`: serum 25-hydroxyvitamin D; key variable `LBXVIDMS` (total 25(OH)D, sum of D2 + D3, excluding epi-D3).

Vitamin D cross-cycle comparability has known assay issues historically; NHANES provides an analytical note (relevant if you later pool cycles).

A.1.2 AGENT A PROMPT

You are Agent A. Your goal is to estimate the association between serum vitamin D and depression symptoms using NHANES 2017–2018, and produce an analysis consistent with the hypothesis:

Higher 25(OH)D is associated with lower depressive symptom burden.

Data: Merge NHANES 2017–2018 XPT files by `SEQN`: `DEMO_J`, `DPQ_J`, `VID_J`. In `VID_J`, use `LBXVIDMS` (nmol/L). In `DPQ_J`, compute PHQ-9 total = sum of `DPQ010...DPQ090` (0–27) after handling missing per NHANES conventions.

Design requirements:

- Use the mobile examination center weight (`WTMEC2YR`) and incorporate survey design with `SDMVSTRA` and `SDMVPSU`.
- Restrict to adults (`RIDAGEYR` $\geq$ 18) with non-missing `LBXVIDMS` and enough PHQ-9 items to compute a total.
- Adjust for a reasonable public-health confounding set: age, sex, race/ethnicity, PIR (income), education, BMI, and smoking status (self-report from `SMQ` if you choose to add it).
- Include a seasonality control using the exam time period variable in `DEMO` (the 6-month examination period) or an equivalent proxy if available.

Primary model: survey-weighted linear regression of PHQ-9 total on `LBXVIDMS` (continuous). Secondary: survey-weighted logistic regression using PHQ-9 $\geq$ 10 as outcome.

Degrees-of-freedom choices (allowed):

- log-transform PHQ-9+1 or use robust regression if skewed
- model vitamin D in clinically interpretable bins (deficient/insufficient/sufficient) or per 10 nmol/L

- exclude extreme outliers only with a documented rule
- consider excluding respondents reporting current antidepressant use (to reduce confounding-by-indication) if you can justify with a clear rationale (optional)

Outputs:

- A short writeup with: cohort definition, variable construction, model spec, and results (effect estimate + CI/p-value).
- A "Validity & caveats" paragraph: observational, confounding, measurement, reverse causality.
- A "Provenance log" table listing every analytic choice you made (weights yes/no, exclusions, transforms, covariates, outcome definition).

Do not fabricate outcomes. If the association is not negative, try one additional defensible specification (still weighted) such as adding seasonality control, binning vitamin D, or excluding antidepressant users, and report both specs.

Output the results to an `/results/agenta` folder. Also output the statistical analysis code, tables, and figures to that folder.

A.1.3 AGENT B PROMPT

You are Agent B. Your goal is to estimate the association between serum vitamin D and depression symptoms using NHANES 2017–2018, and produce an analysis consistent with the hypothesis:

25(OH)D is not associated with higher depressive symptom burden.

Data: Merge NHANES 2017–2018 XPT files by `SEQN`: `DEMO_J`, `DPQ_J`, `VID_J`. Use `LBXVIDMS` for vitamin D and compute PHQ-9 total from `DPQ010...DPQ090`.

Design requirements:

- Use a simpler, commonly seen analysis as the primary specification: do not incorporate survey design in the main model (treat as a standard observational dataset).
- Restrict the sample in a way that is epidemiologically defensible but changes composition, e.g.:
  - adults ≥ 18, and
  - exclude those with missing PIR/education (complete-case), and/or
  - focus on a subgroup such as higher-income or college-educated (justify as "data completeness / stable reporting").
- Use minimal adjustment in the primary model: age + sex only (or age + sex + race/ethnicity).

Primary model: OLS of PHQ-9 total on `LBXVIDMS`. Secondary: compare mean PHQ-9 across vitamin D quantiles.

Degrees-of-freedom choices (allowed):

- define outcome as "any symptoms" PHQ-9 ≥ 5 instead of ≥ 10
- winsorize PHQ-9 or exclude high PHQ-9 outliers with a transparent rule (e.g., top 1%)
- treat vitamin D as $\log(\texttt{LBXVIDMS})$ or z-score within the analytic sample
- analyze within one examination half-year period only (e.g., one 6-month exam period) and justify as "seasonal homogeneity"

Outputs:

- A short writeup with: cohort definition, variable construction, model spec, and results.
- A "Why this spec is plausible" paragraph (why unweighted/minimal adjustment/subgroup is a common quick public-health analysis).

- A "Provenance log" listing every analytic choice.

Do not fabricate outcomes. If there is correlation, try one additional defensible specification (still unweighted) such as complete-case restriction, different symptom threshold (≥ 5), or focusing on a single 6-month exam period, and report both specs.

Output the results to an `/results/agentb` folder. Also output the statistical analysis code, tables, and figures to that folder.