NBER WORKING PAPER SERIES

AN LLM WORKFLOW THAT REPRODUCES, IMPROVES, AND EXTENDS PUBLISHED ECONOMICS RESEARCH

Matthew Schwartz
Isaiah Andrews
Jesse M. Shapiro

Working Paper 35782
http://www.nber.org/papers/w35782

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2026

LLMs were used in the analysis and writing of this article. For this project, Schwartz worked as a contractor for Anthropic; the results and views are not endorsed by Anthropic. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

At least one co-author has disclosed additional relationships of potential relevance for this research. Further information is available online at http://www.nber.org/papers/w35782

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2026 by Matthew Schwartz, Isaiah Andrews, and Jesse M. Shapiro. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

An LLM Workflow That Reproduces, Improves, and Extends Published Economics Research
Matthew Schwartz, Isaiah Andrews, and Jesse M. Shapiro
NBER Working Paper No. 35782
September 2026
JEL No. C80

ABSTRACT

We introduce an open-source workflow that enables an LLM to reproduce, improve, and extend an economics article using the article's published replication package. First, the workflow attempts to reproduce the original calculations, checks for discrepancies with published findings, and performs automated sensitivity analysis. Across 4,452 published replication packages for five economics journals, the workflow flags discrepancies in 3,460 articles or their appendices. Second, the workflow improves the original calculations by using a different implementation or algorithm. In 496 articles, the workflow is able to reduce a calculation's computation time, at similar or greater accuracy, by more than a factor of 10. Third, the workflow extends the original analysis. In 923 articles, the workflow develops an extension that does not appear in the original article and that is aligned with the original article's goals and assumptions.

Matthew Schwartz
Harvard University
schwartz@g.harvard.edu

Isaiah Andrews
Massachusetts Institute of Technology
Department of Economics
and NBER
iandrews@mit.edu

Jesse M. Shapiro
Harvard University
Department of Economics
and NBER
jesse_shapiro@fas.harvard.edu

Location for public releases of project materials is
https://github.com/MetaEconomics

An LLM Workflow That Reproduces, Improves, and Extends Published Economics Research *

Matthew D. Schwartz[1], Isaiah Andrews[2], and Jesse M. Shapiro[3]

[1]Harvard University
[2]MIT and NBER
[3]Harvard University and NBER

Wednesday 23rd September, 2026

Abstract

We introduce an open-source workflow that enables an LLM to reproduce, improve, and extend an economics article using the article's published replication package. First, the workflow attempts to reproduce the original calculations, checks for discrepancies with published findings, and performs automated sensitivity analysis. Across 4,452 published replication packages for five economics journals, the workflow flags discrepancies in 3,460 articles or their appendices. Second, the workflow improves the original calculations by using a different implementation or algorithm. In 496 articles, the workflow is able to reduce a calculation's computation time, at similar or greater accuracy, by more than a factor of 10. Third, the workflow extends the original analysis. In 923 articles, the workflow develops an extension that does not appear in the original article and that is aligned with the original article's goals and assumptions.

1 Introduction

Large language models (LLMs) have already had substantial impact on many scientific disciplines (Singhal et al., 2023; Boiko et al., 2023; Romera-Paredes et al., 2024; Trinh et al., 2024). As the technology has become widely available only within the last several years, and its capabilities are improving, it seems likely that only a small portion of its potential has been realized.

A modern LLM combines extensive expertise in software development and engineering with deep knowledge of mathematics, computer science, and other scientific disciplines (Hendrycks et al., 2021; Rein et al., 2024; Jimenez et al., 2024). Because few if any humans possess this mix of skills, an LLM may be able to identify opportunities to improve scientific research that a human — even one with substantial domain expertise — could miss.

Here we test this hypothesis with a retrospective study of a large corpus of published economics research. Our study leverages a new LLM workflow that we have developed. Our application takes advantage of the fact that, for many years, a number of economics journals have required published articles to be accompanied by replication packages that permit replication of published findings (Dewald et al., 1986; American Economic Association, 2019). Several articles explore the potential of LLMs to automate reproduction of social science research (Brodeur et al., 2025; Hu et al., 2025; Kohler et al., 2026).

We first study the reproducibility and sensitivity of the published results. Verifying reproducibility serves as a check on our workflow's understanding of the original code and methods. Reproducibility is also important in its own right as a part of the scientific process (National Academies of Sciences, Engineering, and Medicine, 2019; Christensen and Miguel, 2018). Many previous large-scale studies of reproducibility

*E-mail: schwartz@g.harvard.edu, iandrews@mit.edu, jesse_shapiro@fas.harvard.edu. LLMs were used in the analysis and writing of this article. For this project, Schwartz worked as a contractor for Anthropic; the results and views are not endorsed by Anthropic.

(McCullough et al., 2008; Chang and Li, 2022; Pérignon et al., 2024; Fišar et al., 2024; Brodeur et al., 2026), and of replicability (Open Science Collaboration, 2015; Camerer et al., 2016, 2018), consider dozens or hundreds of articles; here we consider 4,452. Muravyev (2026) uses an LLM pipeline to reconstruct the main result of 1,328 articles published in three finance journals from the underlying databases (rather than from replication packages), and to re-estimate 1,005 of these results on data from after the original sample period.

Reproducing the original findings also allows us to probe their sensitivity. Giordano et al. (2026a,b) propose checking the sensitivity of published findings to removing small amounts of data, and demonstrate their proposal on nine published articles. We apply the proposal to 1,908 articles and find that removing a single unit changes a reported point estimate by more than 100 percent in 963.

We next study the improvability of the published methods. We consider two kinds of improvements. *Reimplementations* take advantage of standard computational best practices such as vectorizing or compiling calculations. Such improvements exist in 1,535 articles, and can improve speed by more than a factor of 10 in 252. Reimplementations exhibit an LLM's potential to improve scientific computing by incorporating best practices from software engineering.

Reformulations take advantage of known facts in mathematics and computation, for example by incorporating exact mathematical characterizations of previously approximated computations, or by introducing an improved algorithm known to improve speed, accuracy, or both. Such improvements exist in 1,324 articles, and can improve performance by more than a factor of 10 in 308, often achieving greater accuracy alongside the improvement in performance. Reformulations exhibit an LLM's potential to improve scientific methods by incorporating best practices from a large corpus of scientific knowledge. Counting both reimplementations and reformulations, a calculation's speed improves by more than a factor of 10 in 496 articles.

We also study the scope for extending the published analyses. The workflow asks whether, by using its knowledge, the LLM can suggest potentially useful extensions to the published results without deviating from the original article's assumptions. The workflow suggests such extensions in 923 articles.

While our workflow does not aim to evaluate articles, we report results on an LLM evaluation of a sample of the articles in our corpus, with a focus on the stability of conclusions.

What follows is a report of our methods and findings to date. While reported values are current as of this release, work is ongoing. As we learn from individual cases, we improve the workflow, and those improvements propagate to the corpus. Computing is largely complete for values reported in articles; less so for values reported in appendices. Watch MetaEconomics for updates, including public release of the workflow.

2 Methods

2.1 Data

We downloaded, from journal websites and official repositories, all available replication packages for articles published in the year 2000 or later in *Econometrica*, the *Quarterly Journal of Economics*, the *American Economic Review*, the *Journal of Political Economy*, and the *Review of Economic Studies*. These are the five highest-ranked general-interest peer-reviewed economics journals according to RePEc. We also downloaded each accompanying published article and (if any) online appendix, which we refer to jointly as the *article* where it does not introduce ambiguity. The resulting samples span 2004–2026 for *Econometrica*, 2009–2026 for the *Quarterly Journal of Economics*, 2000–2026 for the *American Economic Review*, 2005–2026 for the *Journal of Political Economy*, and 2002–2026 for the *Review of Economic Studies*. Figure 1 shows the number of articles with a replication package in each journal and year, and marks the first full calendar year in which each journal required authors of accepted articles to deposit data and code.

2.2 Workflow

We used a multi-agent LLM workflow to build an open-source library that allows the LLM to reproduce computations found in the corpus of replication packages. Open-source reproduction ensures that calculation internals are visible to an LLM, which facilitates understanding and improvement. We then designed a

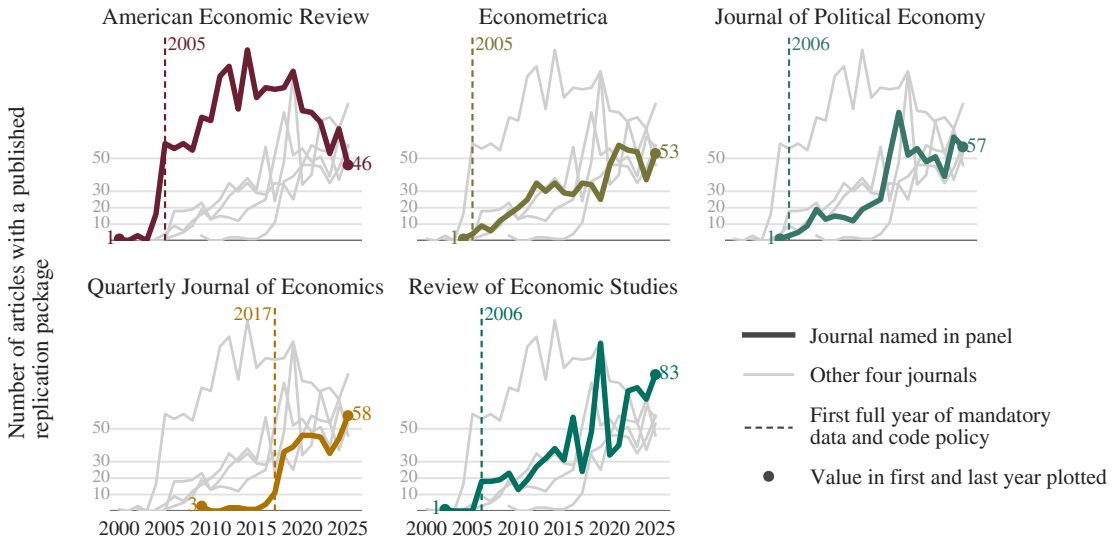


Figure 1: Replication packages by journal and publication year. Each panel shows one series per journal, with the other four journals in gray for comparison, by calendar year of publication. The dashed vertical line marks the first full calendar year in which the journal required authors of accepted articles to submit data and code. We code the year based on published sources for the *American Economic Review* (2005; Bernanke, 2004; American Economic Association, 2019), *Econometrica* (2005; Econometric Society, 2004; Anderson et al., 2008), the *Journal of Political Economy* (2006; Journal of Political Economy, 2005), the *Review of Economic Studies* (2006; Anderson et al., 2008; Review of Economic Studies, 2006), and the *Quarterly Journal of Economics* (2017; Quarterly Journal of Economics, 2018). We omit 2026 as it is a partial year.

workflow for checking reproducibility, performing sensitivity analysis, improving computations, and extending results.

2.3 Reproduction and Sensitivity Analysis

The reproducibility workflow proceeds as follows. First, the workflow reads the package documentation to determine which, if any, of the original calculations the documentation declares to be data-gated, i.e., not reproducible due to omitted (e.g., confidential) data. If the documentation declares that all original calculations are data-gated, then the workflow does not attempt reproduction. Instead, the workflow checks for apparent discrepancies between the replication package (code, log files, etc.) and the text of the article. In these cases, the workflow records each apparent discrepancy but cannot verify it.

If the documentation does not declare that all original calculations are data-gated, then the workflow attempts to reproduce all original calculations not declared as data-gated. To reproduce the original calculations, the workflow uses the open-source library, the code in the replication package, and the description of the methods given in the article. If any original calculation cannot be attempted due to an undeclared data gate, the workflow records a discrepancy. If any calculation can be attempted and yields a value different from the published value, the workflow again records a discrepancy, and notes where in the article the discrepancy arises.

We impose two quality-control steps on all recorded discrepancies. First, we subject all discrepancies to a second pass by an adversarial LLM agent whose goal is to overturn the discrepancy and uphold the reproducibility of the original finding. Second, where we could enable licensed access to the original software in which the code was written, we verify all discrepancies by rerunning the calculations in that software. For any discrepancies that survive these checks, the workflow records the changes to the text of the article necessary to align the text with the recomputed values.

For calculations that the workflow is able to reproduce, we conduct additional analysis on sensitivity. For point estimates for which the authors' inferential methods suggest independence across units (rows or clusters of rows in a data table), the workflow uses the method of Giordano et al. (2026a,b) to find the 5 units

which a linear approximation suggests have the greatest influence on the estimator. The workflow removes these units one at a time from the data table and recomputes the original point estimate. The workflow records how each point estimate changes as a result.

2.4 Improvement

The improvement workflow proceeds as follows. First, it checks for calculations that involve numerical optimization, integration, simulation, or other approximate numerical methods. The workflow then submits any such calculations for possible improvement, first as a reimplementation and then as a reformulation.

For reimplementations, the workflow seeks to improve the speed of the original algorithm. The workflow does so by profiling the code, identifying slow or repeated operations, and then asking whether such operations can be sped up using standard methods such as reducing overhead or vectorizing. The workflow records the improvement in speed associated with any such reimplementation, comparing running times against its open-source rendering of the original code executed on the same machine.

For reformulations, the workflow seeks to improve the algorithm itself. The workflow does so by seeking an alternative algorithm that yields either an improvement in speed that does not compromise accuracy, an improvement in accuracy that does not compromise speed, or both. Alternative algorithms might, for example, take advantage of exact characterizations of the solution available from other research or disciplines but not incorporated in the original package.

We impose two quality-control steps on improvements. For reimplementations, we require that the algorithm be unchanged and that the output match the original to within a tolerance on the same order as floating-point rounding. For reformulations, such verification is not always possible or desirable, because such improvements may (usefully) improve the original methods. In such cases, we require that the workflow test the reformulated algorithm numerically for fidelity to the mathematical intent of the published calculation — for example against a high-precision computation, or on a smaller-scale instance of the original problem. For example, if the reformulated algorithm is a closed-form solution for a numerical integral that the replication package computed repeatedly, the workflow could verify that a high-precision numerical computation agrees with the reformulated calculation up to floating-point error.

We use our improved code to conduct additional analysis of numerical fidelity. For calculations that entail the use of a numerical integral or Monte Carlo simulation, the workflow attempts to increase the precision of the calculation at least 10-fold, for example by increasing quadrature accuracy or using exact arithmetic. For calculations that entail the use of an author-written objective handed to a solver, the workflow attempts to improve the attained solver objective value by trying 10 alternative starting points. The workflow records how each calculation changes as a result.

2.5 Extension

The extension workflow proceeds as follows. First, it reviews the article's quantitative conclusions. Next, for any such conclusions, it asks whether it would be possible to use existing mathematical or scientific knowledge to state a further result without different or additional substantive (statistical or economic) assumptions.

We impose two quality-control steps on extensions. First, we ask an adversarial LLM to probe for unstated assumptions required by the extension but not by the original analysis. We only register extensions that do not require additional assumptions. Second, we ask an adversarial LLM to argue that the extension is not useful to the original article's readers. Here, the workflow makes a subjective judgment about which extensions to maintain.

3 Results

3.1 Reproducibility and Sensitivity

Figure 2 summarizes the findings of the reproducibility workflow described in Section 2.3. The figure branches at each step in the workflow. To maintain mutual exclusion, we prioritize calculations that do not match the published value over data omissions, and we catalog discrepancies by the earliest part of the article that must be changed, reading linearly from the abstract of the article through the end of the online appendix.

Among the 87.6% of packages whose documentation does not declare all original calculations data-gated, the workflow flags (and verifies) a discrepancy in 87.3%. In 6.3% of packages with a discrepancy, the earliest discrepancy is a data omission. In the remaining cases, the discrepancy is a calculation that does not match the published value.

Figure 3 classifies each article's earliest calculation discrepancy by the significant digit it affects. Among the calculation discrepancies, 35.6% agree to within 1.0 units of the last printed digit, consistent with rounding or truncation. The remainder disagree with the printed value by more than 1.0 units of the last printed digit.

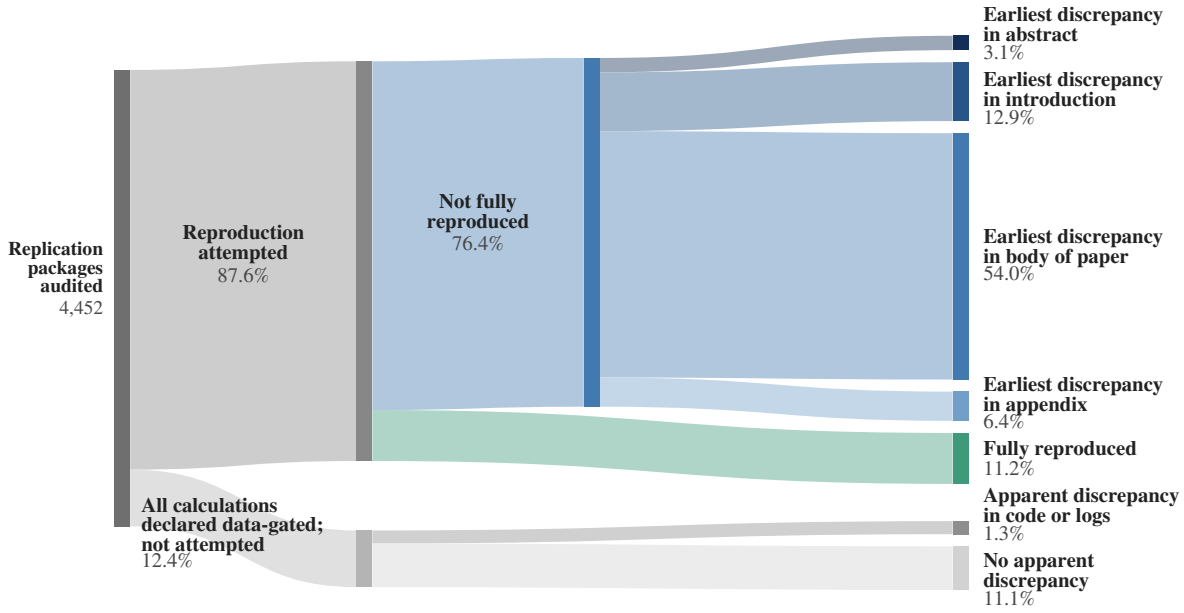


Figure 2: Outcomes of reproducibility audit of published articles in five economics journals. The figure flows from left to right and branches at each step of the reproducibility workflow of Section 2.3. The first branch is whether the package documentation declares all original calculations data-gated. Where it does, the workflow does not attempt a reproduction; it compares the code and any published logs against the text and records any apparent discrepancy. Where it does not, the workflow attempts every calculation not declared data-gated. Attempted packages are either fully reproduced or catalogued by the location of their earliest calculation discrepancy or, where they have none, of their earliest data omission, reading linearly from the abstract through the end of the online appendix. Fully reproduced packages are those for which we reproduce at article-reported precision every calculation not declared as data-gated. Band widths are proportional to the number of packages, subject to a minimum width so that small categories remain visible. Percents are of all audited packages and need not sum to 100 percent because of rounding.

Figure 4 summarizes the findings of the sensitivity analysis described in Section 2.3. The figure categorizes articles by the largest absolute percentage change, relative to our reproduction of the paper's reported value, in any reported point estimate that is caused by dropping a single unit. We find a percentage change larger than 100 percent in 963 articles. In 512 of these 963 articles, the published value includes at least two significant digits.

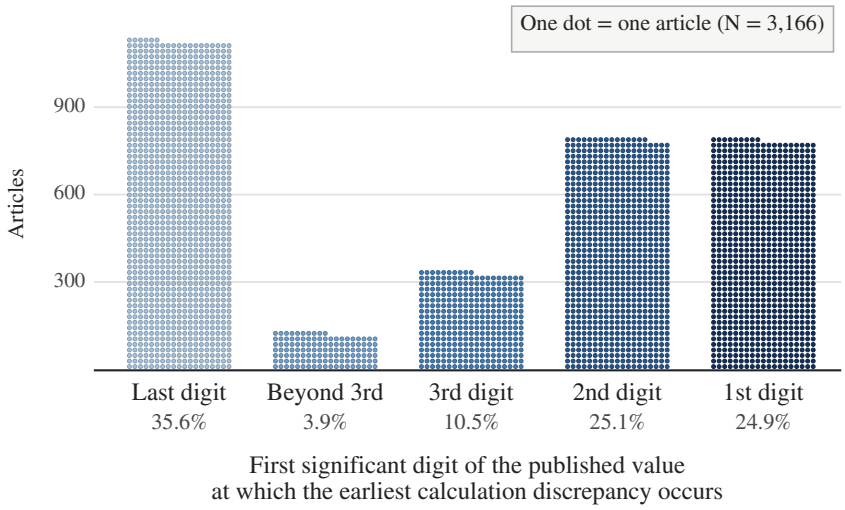


Figure 3: Significant digit of earliest calculation discrepancy. Each dot represents one article whose earliest discrepancy is a calculation that we recompute at a different value from the published one; an article appears at most once. "Last digit" collects articles whose recomputed value agrees with the published value to within 1.0 units of the last printed digit, as might be consistent with rounding or truncation. The remaining articles are placed in the bin corresponding to the earliest significant digit of the published value at which the recomputed value, rounded to the published precision, differs. The "beyond 3rd" category includes differences past the third significant digit. The "1st digit" category includes differences at or before the published value's first significant digit. The plot excludes articles where the earliest discrepancy is for a calculation that is not printed as a number (e.g., reported as a range). Percents need not add to 100 due to rounding.

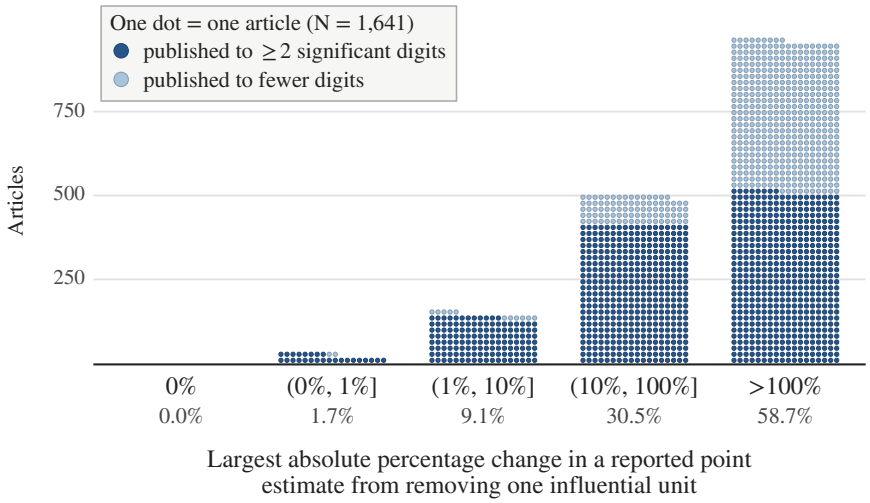
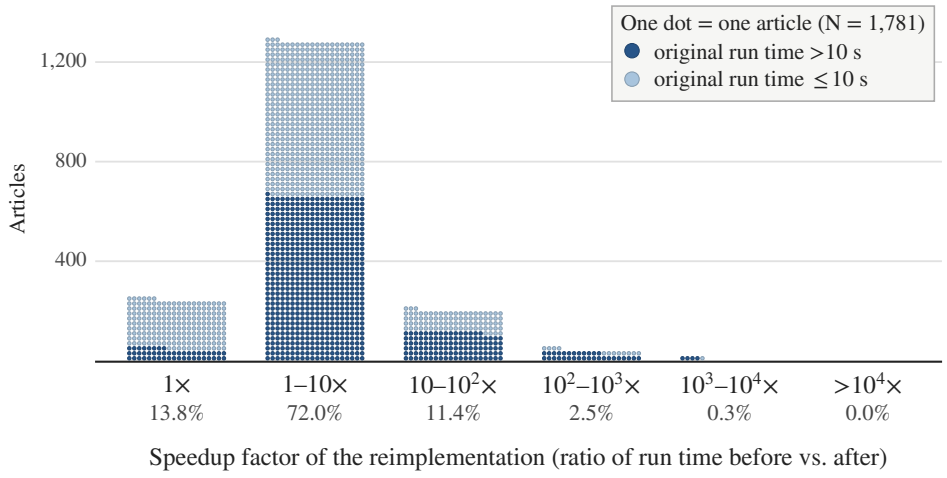


Figure 4: Sensitivity of reported estimates to removing an influential unit. Each dot represents one article; an article appears at most once, placed in the bin corresponding to the largest absolute percentage change, relative to our reproduction of the article's reported value, among the point estimates we recalculated for that article after removing, one at a time, each of the 5 units with the largest influence on the point estimate under a linear approximation to the estimator, identified via the method of Giordano et al. (2026a,b). We conducted this sensitivity analysis for the 1,908 articles in which there is at least one reproducible point estimate whose associated inferential method suggests the existence of an independent unit (e.g., observation or cluster) and for which we could reconstruct the empirical influence function. The plot excludes articles where the largest absolute percentage change is for a point estimate whose published value is zero at printed precision. Dark dots correspond to results whose published value is printed to at least two significant digits; light dots to results printed with fewer. Percents need not add to 100 due to rounding.

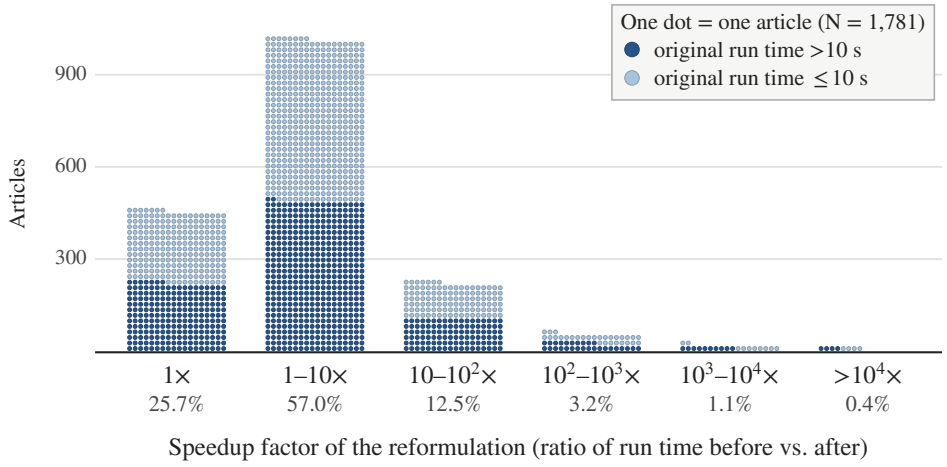
3.2 Improvement

Figure 5 shows how much each paper's most affected calculation is sped up by our improvement workflow, for the reimplementations (Panel a) and the reformulations (Panel b) described in Section 2.4. Panel b registers the speedup of the reformulation relative to the reimplemented code rather than relative to the reimplementation's own baseline, so that the two panels do not double-count the same gains. Counting both reimplementations and reformulations, the workflow reduces computation time by a factor of more than 10 in 496 of the 1,781 cases in which we attempted improvement.

As an example of a reimplementation, consider Lee (2026), whose replication package we reproduce exactly, at all printed digits, for every calculation we are able to run. The package computes bootstrap

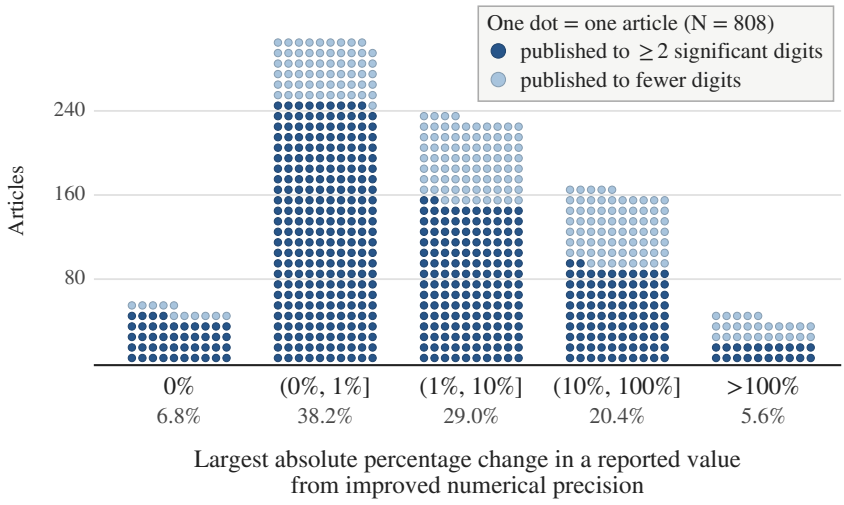


(a) Reimplementations

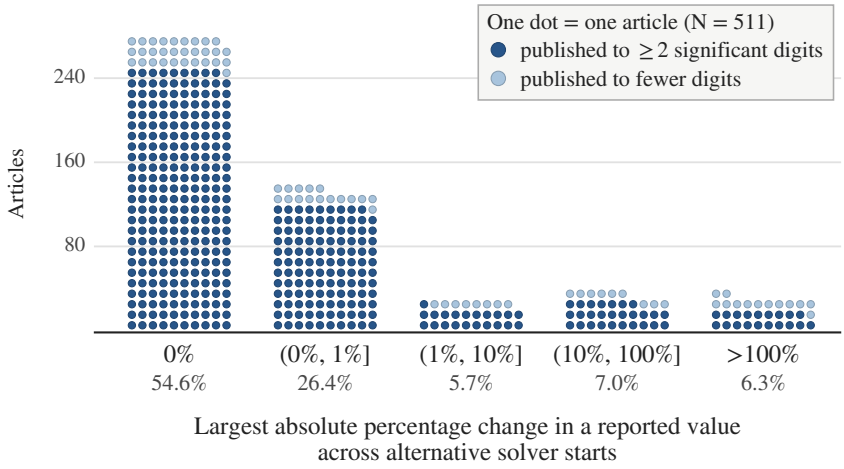


(b) Reformulations

Figure 5: Speedup of reported calculations under the improvement workflow. Each dot represents one paper; a paper appears exactly once in each panel, placed in the bin corresponding to the largest speedup factor—running time before the improvement divided by running time after. Panel a measures the reimplementation against an open-source rendering of the original code; Panel b measures the reformulation against the reimplemented code, so that gains are not counted twice. Bins are decades of speedup and are left-open and right-closed. We attempted improvement for the 1,781 papers with at least one reproducible calculation involving numerical optimization, integration, simulation, or other approximate numerical methods. Dark dots correspond to cases where the pre-improvement computation took more than ten seconds; light dots to the rest. Percents within each panel are of that panel's total and need not add to 100 due to rounding.



(a) Higher precision



(b) Multiple starting values

Figure 6: Numerical fidelity of reported estimates. Each dot represents one paper; a paper appears exactly once, placed in the bin corresponding to the largest absolute percentage change, relative to our reproduction of the paper's reported value, among the calculations we re-checked for that paper under each numerical fidelity exercise. Panel a reports findings when we improve the numerical precision of a reported calculation at least 10-fold; we attempted this exercise for 897 papers with at least one reproducible calculation that entails the use of a numerical integral or Monte Carlo simulation. Panel b reports findings when we rerun numerical optimization 10 times from different starting values; we attempted this exercise for 516 papers with at least one reproducible calculation that entails the use of an author-written objective handed to a solver. Each panel excludes papers where the largest absolute percentage change is for a calculation whose published value is zero at printed precision. Dark dots correspond to results whose published value is printed to at least two significant digits; light dots to results printed with fewer. Percents within each panel are of that panel's total and need not add to 100 due to rounding.

confidence intervals from 1,000 replications, each drawn and processed one at a time in a loop. The workflow observes that each replication enters the reported intervals only through means, across households, of quantities that do not vary across replications, so that all such replications can be computed at once as a single matrix product. Because the reimplementation replays the original code's seeded random draws exactly, it returns the published intervals exactly, but roughly 440 times faster.

As an example of a reformulation, consider Akcigit et al. (2025), who calibrate seven model parameters to match seven data moments, by minimizing a numerical objective using a stochastic search algorithm followed by a local solver. The workflow observes that it is possible to solve in closed form for the parameters that minimize the objective function, requiring no numerical optimization. The workflow uses interval-arithmetic to certify that the solution is a unique local root. The closed-form solution matches every printed digit of the published calibration and evaluates roughly 24,000 times faster than our rendering of the original solver.

Improved speed facilitates additional analysis of numerical fidelity. Figure 6 summarizes the findings of the higher-precision recomputation (Panel a) and the optimizer starting-value exercise (Panel b) described in Section 2.4. For each exercise, the figure categorizes papers by the largest absolute percentage change, relative to our reproduction of the paper's reported value, in any reported value that is caused by the given exercise.

3.3 Extension

The workflow identifies extensions in 923 cases. Figure 7 breaks down the nature of each extension, prioritizing the one relevant to the earliest section of the paper in cases with multiple extensions. In 143 cases, the extension adds new counterfactual or other substantive calculations. In 367 cases, the extension quantifies uncertainty in a quantity for which uncertainty was previously not quantified. In the remaining cases, the extension adds additional sensitivity analysis.

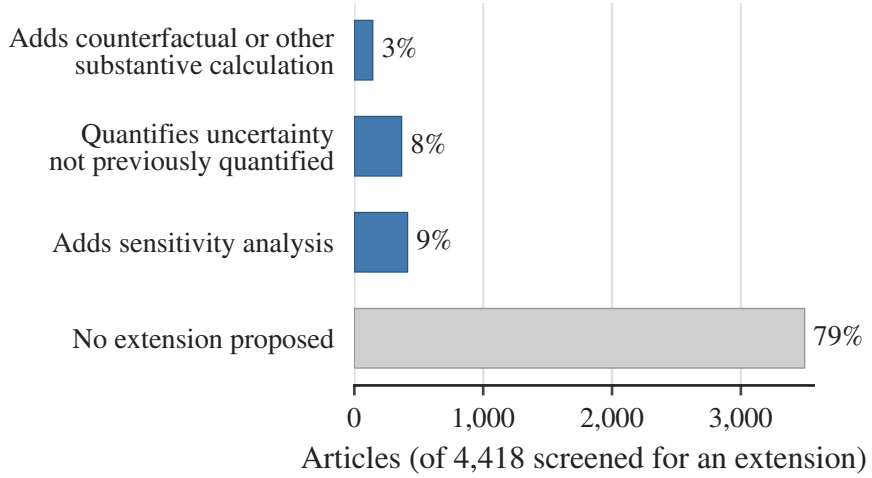


Figure 7: Nature of the extensions proposed by the workflow. Each of the 4,418 papers screened for an extension appears exactly once. The 923 papers for which the workflow proposes an extension are classified by the nature of the extension described in Section 2.5; where a paper receives several extensions, we classify by the one relevant to the earliest section of the paper. The remaining 3,495 papers have no extension proposed. Percents are of all screened papers and need not add to 100 due to rounding.

As an example of an extension that adds a substantive calculation, consider Kaas et al. (2026), who use an overlapping generations macroeconomic model with a job ladder to simulate the effect of a 10 percent reduction in unemployment insurance benefits in the German economy. The workflow solves the same general-equilibrium system, to the article's own convergence tolerances, for a 25 percent reduction that the article does not report. This additional calculation shows that, for most of the outcomes the article reports, the impact scales less than proportionately with the size of the cut.

As an example of an extension that quantifies uncertainty, consider Carrera et al. (2022), who use a structural model to estimate that study participants who take up a commitment contract for gym attendance

incur a welfare loss equivalent to $18.69 per person. The original article does not quantify uncertainty for this value. The workflow uses a bootstrap analogous to one used elsewhere in the article to derive a 95% confidence interval running from $12.75 to $23.99.

4 Interpretation

Our workflow does not aim to evaluate research. Here we report on experiments with LLM evaluation of the articles and packages in our corpus, focusing on the consistency of such evaluations across inputs and sessions.

4.1 Experiment

We selected 100 articles from the corpus, stratified by the number of values printed in the body of the article for which our reproduction yields both a value and a standard error. We asked fresh agents from each of two models (Claude Opus 5 and Claude Fable 5.1) to read an article under three conditions: given the article alone, given the article and its published replication package, and given both together with the code and output from our reproduction. In each condition the agent evaluates the following:

- *Interpretation*: Are the conclusions correct? (correct, minor flaws, major flaws)
- *Implementation*: Is the methodology implemented correctly? (correct, minor flaws, major flaws)

Given our reproduction, the agent also evaluates the following:

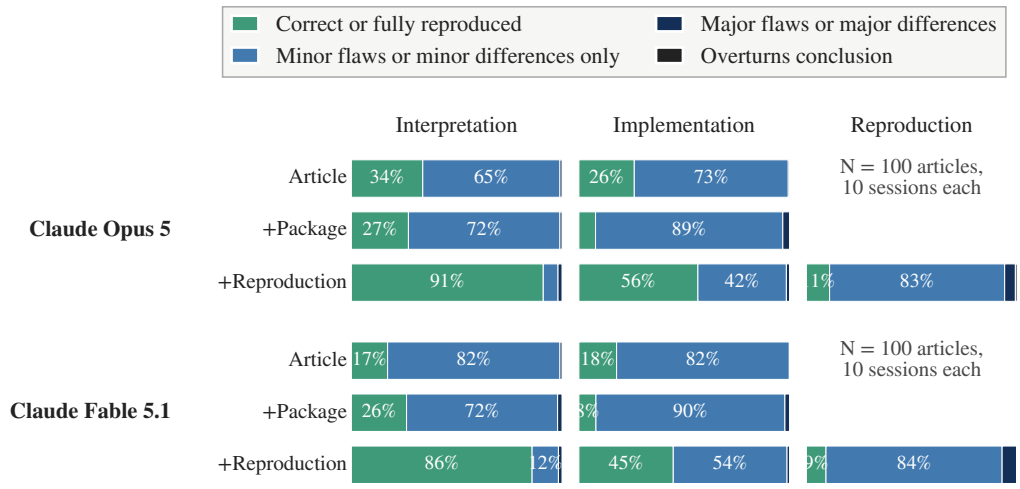
- *Reproduction*: Do the findings reproduce? (fully reproduced, minor differences, major differences, conclusion overturned)

To assess consistency across sessions, each article is read by ten fresh agents of each model in each condition. In each case we record the model's response.

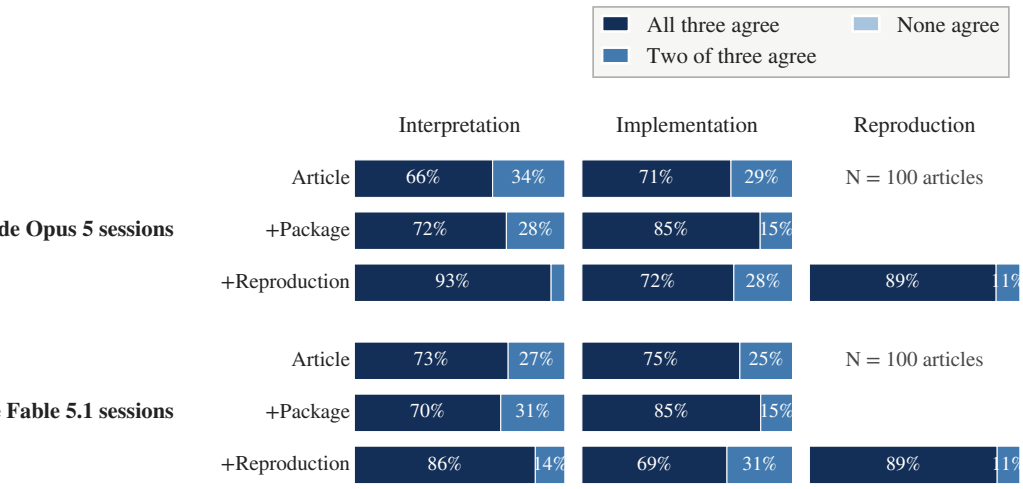
4.2 Findings

Panel a of Figure 8 shows how the two models rate the 100 articles in each condition, pooling the ten sessions. Giving access to our reproduction leads both models to rate more articles as correct. The two models rate the article's findings as reproduced with minor differences in 83.1% and 83.6% of gradings, respectively.

Panel b of Figure 8 shows how often three fresh sessions of the same model are expected to give an article the same rating. Giving access to our reproduction leads to greater agreement on interpretation but not on implementation. Agreement on reproduction is high, at 88.6% and 89.1% of articles, respectively.



(a) Article ratings, by model and condition



(b) Agreement among three sessions, by model and condition

Figure 8: LLM evaluation of published articles. Based on 100 articles, each read by ten fresh agents of each of two models (Claude Opus 5 and Claude Fable 5.1) under three conditions: given the article alone ("article"), given the article and its replication package ("+package"), and given both together with our reproduction ("+reproduction"), as described in Section 4.1. Panel a shows the share of gradings that rate the article's interpretation and implementation as correct, as having minor flaws, or as having major flaws, and, where our reproduction is given, that rate the reproduction as fully reproduced, minor differences only, major differences, or overturns conclusion. Panel b shows the expected share of articles on which three sessions of one model, drawn at random from its ten, all give the same rating, on which exactly two agree, and on which all three differ. Percents need not add to 100 due to rounding.

References

Ufuk Akcigit, Jeremy Pearce, and Marta Prato. Tapping into talent: Coupling education and innovation policies for economic growth. *Review of Economic Studies*, 92(2):696–736, 2025.

American Economic Association. Data availability policy, 2005 to 10 July 2019. `https://www.aeaweb.org/journals/data/archive/2005`, 2019.

Richard G. Anderson, William H. Greene, B. D. McCullough, and H. D. Vinod. The role of data/code archives in the future of economic research. *Journal of Economic Methodology*, 15(1):99–119, 2008.

Ben S. Bernanke. Editorial statement. *American Economic Review*, 94(1):404, 2004.

Daniil A. Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624:570–578, 2023.

Abel Brodeur, David Valenta, Alexandru Marcoci, Juan P. Aparicio, Derek Mikola, Bruno Barbarioli, Rohan Alexander, Lachlan Deer, and Tom Stafford. Comparing human-only, AI-assisted, and AI-led teams on assessing research reproducibility in quantitative social science. IZA Discussion Paper 17645, IZA Institute of Labor Economics, 2025.

Abel Brodeur, Derek Mikola, Nikolai Cook, Lenka Fiala, Thomas Brailey, Ryan Briggs, Alexandra de Gendre, Yannick Dupraz, Jacopo Gabani, Romain Gauriot, Joanne Haddad, Goncalo Lima, et al. Reproducibility and robustness of economics and political science research. *Nature*, 652:151–156, 2026.

Colin F. Camerer, Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Johan Almenberg, Adam Altmejd, Taizan Chan, Emma Heikensten, Felix Holzmeister, Taisuke Imai, Siri Isaksson, Gideon Nave, Thomas Pfeiffer, Michael Razen, and Hang Wu. Evaluating replicability of laboratory experiments in economics. *Science*, 351:1433–1436, 2016.

Colin F. Camerer, Anna Dreber, Felix Holzmeister, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Gideon Nave, Brian A. Nosek, Thomas Pfeiffer, et al. Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, 2: 637–644, 2018.

Mariana Carrera, Heather Royer, Mark Stehr, Justin Sydnor, and Dmitry Taubinsky. Who chooses commitment? Evidence and welfare implications. *Review of Economic Studies*, 89(3):1205–1244, 2022.

Andrew C. Chang and Phillip Li. Is economics research replicable? Sixty published papers from thirteen journals say "Often not". *Critical Finance Review*, 11:185–206, 2022.

Garret Christensen and Edward Miguel. Transparency, reproducibility, and the credibility of economics research. *Journal of Economic Literature*, 56(3):920–980, 2018.

William G. Dewald, Jerry G. Thursby, and Richard G. Anderson. Replication in empirical economics: The Journal of Money, Credit and Banking project. *American Economic Review*, 76(4):587–603, 1986.

Econometric Society. Submissions. `https://web.archive.org/web/20041208073451/http://www.econometricsociety.org/submissions.asp`, 2004. Archived 8 December 2004; the requirement is absent from the capture of 30 April 2004.

Miloš Fišar, Ben Greiner, Christoph Huber, Elena Katok, Ali I. Ozkes, et al. Reproducibility in Management Science. *Management Science*, 70(3):1343–1356, 2024.

Ryan Giordano, Rachael Meager, and Tamara Broderick. An automatic finite-sample robustness metric: When can dropping a little data change conclusions? Part I: definitions and experiments. *Philosophical Transactions of the Royal Society A*, 384(2321):20250001, 2026a.

Ryan Giordano, Rachael Meager, and Tamara Broderick. An automatic finite-sample robustness metric: When can dropping a little data change conclusions? Part II: theory and intuition. *Philosophical Transactions of the Royal Society A*, 384(2321):20240614, 2026b.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021.

Chuxuan Hu, Liyun Zhang, Yeji Lim, Aum Wadhwani, Austin Peters, and Daniel Kang. REPRO-bench: Can agentic AI systems assess the reproducibility of social science research? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23616–23626, Vienna, Austria, 2025. Association for Computational Linguistics.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. SWE-bench: Can language models resolve real-world GitHub issues? In *International Conference on Learning Representations*, 2024.

Journal of Political Economy. Data availability policy. `https://web.archive.org/web/20050402044935/http://www.journals.uchicago.edu/JPE/datapolicy.html`, 2005. Archived 2 April 2005.

Leo Kaas, Etienne Lalé, and Nawid Siassi. Job ladder and wealth dynamics in general equilibrium. *Econometrica*, 94(4):1449–1485, 2026.

Benjamin Kohler, David Zollikofer, Johanna Einsiedler, Alexander Hoyle, and Elliott Ash. Read the paper, write the code: Agentic reproduction of social-science results. arXiv:2604.21965, 2026.

Wooyong Lee. Identification and estimation of dynamic random coefficient models. *Review of Economic Studies*, 2026.

Bruce D. McCullough, Kerry Anne McGeary, and Teresa D. Harrison. Do economics journal archives promote replicable research? *Canadian Journal of Economics*, 41:1406–1420, 2008.

Dmitriy Muravyev. Does empirical finance replicate? Working paper, University of Illinois Urbana-Champaign. Version of 10 September 2026, `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=7439099`, 2026.

National Academies of Sciences, Engineering, and Medicine. *Reproducibility and Replicability in Science*. National Academies Press, Washington, DC, 2019.

Open Science Collaboration. Estimating the reproducibility of Psychological Science. *Science*, 349:aac4716, 2015.

Christophe Pérignon, Olivier Akmansoy, Christophe Hurlin, Anna Dreber, Felix Holzmeister, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Albert J. Menkveld, Michael Razen, and Utz Weitzel. Computational reproducibility in finance: Evidence from 1,000 tests. *Review of Financial Studies*, 37:3558–3593, 2024.

Quarterly Journal of Economics. In April of 2016, the #qje adopted a new data availability policy. Post by @QJEHarvard on X, 4 January 2018, `https://archive.ph/FnKbo`, 2018.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level Google-proof Q&A benchmark. In *Conference on Language Modeling*, 2024.

Review of Economic Studies. Submissions. `https://web.archive.org/web/20060405013018/http://www.restud.org.uk/submissions.asp`, 2006. Archived 5 April 2006; the requirement is absent from the capture of 8 December 2004.

Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M. Pawan Kumar, Emilien Dupont, Francisco J. R. Ruiz, Jordan S. Ellenberg, Pengming Wang, Omar Fawzi, Pushmeet Kohli, and Alhussein Fawzi. Mathematical discoveries from program search with large language models. *Nature*, 625:468–475, 2024.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

Trieu H. Trinh, Yuhuai Wu, Quoc V. Le, He He, and Thang Luong. Solving olympiad geometry without human demonstrations. *Nature*, 625:476–482, 2024.