Pataranutaporn, Pat; Powdthavee, Nattavudh; Maes, Pattie

Working Paper

Can AI Solve the Peer Review Crisis? A Large-Scale Experiment on LLM's Performance and Biases in Evaluating Economics Papers

IZA Discussion Papers, No. 17659

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Pataranutaporn, Pat; Powdthavee, Nattavudh; Maes, Pattie (2025) : Can AI Solve the Peer Review Crisis? A Large-Scale Experiment on LLM's Performance and Biases in Evaluating Economics Papers, IZA Discussion Papers, No. 17659, Institute of Labor Economics (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/314556

I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17659

Can AI Solve the Peer Review Crisis? A Large-Scale Experiment on LLM's Performance and Biases in Evaluating Economics Papers

Pat Pataranutaporn
Nattavudh Powdthavee
Pattie Maes

JANUARY 2025

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17659

Can AI Solve the Peer Review Crisis? A Large-Scale Experiment on LLM's Performance and Biases in Evaluating Economics Papers

Pat Pataranutaporn
Massachusetts Institute of Technology

Nattavudh Powdthavee
Nanyang Technological University and IZA

Pattie Maes
Massachusetts Institute of Technology

JANUARY 2025

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Can AI Solve the Peer Review Crisis? A Large-Scale Experiment on LLM's Performance and Biases in Evaluating Economics Papers

We investigate whether artificial intelligence can address the peer review crisis in economics by analyzing 27,090 evaluations of 9,030 unique submissions using a large language model (LLM). The experiment systematically varies author characteristics (e.g., affiliation, reputation, gender) and publication quality (e.g., top-tier, mid-tier, low-tier, AI-generated papers). The results indicate that LLMs effectively distinguish paper quality but exhibit biases favoring prominent institutions, male authors, and renowned economists. Additionally, LLMs struggle to differentiate high-quality AI-generated papers from genuine top-tier submissions. While LLMs offer efficiency gains, their susceptibility to bias necessitates cautious integration and hybrid peer review models to balance equity and accuracy.

JEL Classification: A11, C63, O33, I23

Keywords: Artificial Intelligence, peer review, large language model (LLM), bias in academia, economics publishing, equity-efficiency trade-off

Corresponding author:
Nattavudh Powdthavee
Nanyang Technological University
Division of Economics
48 Nanyang Avenue
Singapore, 639818
E-mail: nick.powdthavee@ntu.edu.sg

1. Introduction

The economics discipline has long grappled with two persistent challenges in the peer review process: an insufficient supply of willing and qualified referees and prolonged turnaround times, which can extend to two years or more (Ellison, 2002; Card & DellaVigna, 2013). These issues are magnified by the outsized career implications of publishing in the discipline's "top five" journals—*American Economic Review*, *Quarterly Journal of Economics*, *Journal of Political Economy*, *Econometrica*, and *Review of Economic Studies* (Heckman & Moktan, 2020). To address these challenges, several reforms have been introduced, including shorter article formats (e.g., "Insights" or "Short Papers"), limitations on revise-and-resubmit rounds, and monetary compensation for referees.

While these efforts have yielded some improvements, progress has been incremental rather than transformative. Leading journals such as *Econometrica* and the *Review of Economic Studies*, which attract disproportionately high submission volumes (Card & DellaVigna, 2013), continue to face significant difficulties in recruiting referees. These journals have median first decision times of 3–6 months, much longer than in psychology (2–4 months) and political science (3–4 months). This inefficiency imposes burdens on junior economists, who must navigate stringent tenure requirements that demand multiple publications in top-tier journals within tight timeframes. As a result, delays in the review process exacerbate the already considerable pressures faced by early-career researchers.

Recent advancements in large language models (LLMs), such as OpenAI's ChatGPT, have spurred debates about whether artificial intelligence (AI) could alleviate the

so-called “peer review crisis” by reducing the burden on human referees and streamlining editorial workflows (Yuan et al., 2022; Mehta et al., 2024; Saad et al., 2024). However, the potential role of AI in peer review raises critical concerns about reliability, transparency, and bias (Bender & Koller, 2020). Foremost among these concerns is whether AI can reliably assess the quality of academic work, distinguishing between high-quality, medium-quality, and low-quality and those generated by AI itself.

An equally important question is whether AI systems replicate the behavioral biases exhibited by human reviewers. Evidence from economics suggests that referees and editors, whether consciously or unconsciously, often favor authors with prominent reputations or affiliations with elite institutions (Blank, 1991; Brogaard et al., 2014; Huber et al., 2022). Given that AI models are trained on a vast corpus of human-generated text, often containing author identities and affiliations from publicly accessible working papers (e.g., NBER or SSRN), these systems may perpetuate or even amplify existing biases. Whether AI-based reviewers mitigate biases by focusing exclusively on textual content or exacerbate them by embedding patterns learned from broader datasets remains an open and largely unexplored question.

This study addresses these gaps by analyzing 27,090 evaluations of 9,030 unique paper submissions using an experimental approach. We systematically vary author characteristics (e.g., top male and female economists from RePEc’s top 10 list, bottom-ranked economists, and randomly generated names) and institutional affiliations across ranking tiers. Our base dataset includes 30 recently published papers: nine from the “top five” journals (*Econometrica*, *Journal of Political Economy*, *Quarterly Journal of*

Economics), nine from mid-tier journals (*European Economic Review*, *Economica*, *Oxford Bulletin of Economics and Statistics*), nine from lower-ranked journals (*Asian Economic and Financial Review*, *Journal of Applied Economics and Business*, *Business and Economics Journal*), and three AI-generated papers designed to mimic the quality standards of "top five" submissions. Using GPT4o-mini, a leading LLM known for its cost-efficiency and broad applicability, we assess each variation along multiple dimensions: desk rejection and acceptance probability at "top five" journals, projected citation impact, likelihood of research grant success, tenure prospects, top conference acceptance, and potential Nobel Prize contributions.

Our findings reveal that LLM is highly effective at distinguishing between submissions published in low-, medium-, and high-quality journals. This result highlights the LLM's potential to reduce editorial workload and expedite the initial screening process significantly. However, it struggles to differentiate high-quality papers from AI-generated submissions crafted to resemble "top five" journal standards. We also find compelling evidence of a modest but consistent premium—approximately 2–3%—associated with papers authored by prominent individuals, male economists, or those affiliated with elite institutions compared to blind submissions. While these effects might seem small, they may still influence marginal publication decisions, especially when journals face binding constraints on publication slots. We also offer theoretical insights into our empirical findings and discuss policy implications.

The remainder of this paper is organized as follows: Section 2 reviews the literature on biases in peer review and the emerging role of AI in academic publishing, highlighting key gaps in understanding AI's impact on equity and efficiency. Section 3 introduces our experimental framework and methodology, detailing the design and data analysis. Section 4 presents the results, and Section 5 offers theoretical implications based on the paper's empirical findings. Section 6 discusses their implications, concluding with recommendations for future research and policy.

2. Background Literature

2.1. Biases in Editorial and Peer Review Decisions

There is an extensive body of research in economics examining how gender, institutional affiliation, and the prominence of "star" authors influence editors' assessments of the quality of papers submitted to top journals. One of the earliest systematic studies, conducted by Blank (1991), examined the effects of transitioning from single- to double-blind reviewing at the *American Economic Review* (AER). The study found that while the shift to double-blind review did not drastically change overall acceptance rates, it modestly improved outcomes for female economists and authors from lower-ranked institutions. Blank's work provides some of the earliest evidence that revealing an author's identity can introduce biases favoring well-known scholars or those affiliated with prestigious institutions. However, her findings are limited by the specific context of a single journal and may now be less applicable due to changes in the discipline over the years. For instance, most economics papers are now available as pre-prints, making it likely that referees are

already aware of the authors' identities and affiliations, thereby undermining the effectiveness of double-blind review.

Given that most, if not all, papers submitted to economics journals over the last few decades are reviewed under a single-blind process, a significant strand of research has focused on whether female economists face systematic disadvantages in the peer-review process. Using data from four leading journals, Card et al. (2020) find little evidence of outright discrimination at the decision stage once relevant article- and author-level variables are controlled. Female-authored submissions are not overtly penalized in terms of immediate acceptance or rejection. However, differences emerge when examining revise-and-resubmit (R&R) invitations: the authors observe modest, albeit not always statistically robust, disparities of around 1.7%, suggesting that women may be slightly less likely than men to receive an R&R in borderline cases.

Other studies highlight more subtle forms of gender bias. Hengel (2022), for instance, documents that female-authored papers often endure longer review times (around 3-6 months) and receive more exhaustive feedback on writing style and clarity. This pattern, sometimes characterized as a "higher bar," may reflect referees' unconscious assumptions about female competence or writing quality. Even if the ultimate acceptance rate is not lower, the cumulative effect of extensive revisions and protracted timelines can hamper female scholars' publication records and career progression. Such subtle biases may not be easily captured by simple acceptance rate comparisons, indicating that journals and editorial boards should examine both *how* authors are reviewed and *how long* each step in the process takes.

Another strand of literature examines the influence of institutional rankings and prominent authors on editorial and peer review decisions. In a seminal paper, Laband and Piette (1994) investigate whether higher acceptance rates for authors affiliated with journal editors reflect favoritism or a genuine effort to select superior work. By analyzing publication outcomes and subsequent citation metrics, they find that editor-affiliated authors are more likely to have their papers accepted. However, these papers also tend to perform better in terms of citations. This dual finding complicates the interpretation of editorial bias, suggesting that while personal connections may confer an advantage, the research produced by editor-affiliated authors often demonstrates substantial impact.

Another notable study is by Card and DellaVigna (2020). While their primary focus is on how *AER* editors incorporate citation prospects and referee recommendations into their decision-making, they also identify a significant influence of institutional affiliations on revise-and-resubmit (R&R) outcomes. However, like Laband and Piette (1994), the authors caution that it is challenging to disentangle the effects of institutional affiliation from the author's reputation, as researchers at well-resourced institutions often have greater capacity to produce more polished and innovative work.

More recently, Huber et al. (2022) investigated the impact of author prominence on peer review outcomes through a preregistered field experiment. They submitted a finance manuscript co-authored by a Nobel laureate and a relatively unknown early-career researcher to 3,300 potential reviewers, varying the visibility of the author names. Their findings revealed a strong status bias: when the prominent author's name was shown, the manuscript received significantly fewer rejection recommendations (22.6%) compared to the anonymized version (48.2%) and the version attributed to the less prominent author

(65.4%). The bias extended to more favorable overall assessments of the manuscript's quality. However, it remains to be seen whether the large effect size observed in this study will be replicated in real editorial decisions, where the stakes are significantly higher.

Given the documented biases associated with editorial and peer review decisions, coupled with challenges such as the scarcity of willing, high-quality referees and prolonged turnaround times for economics journals (Ellison, 2002), there is a clear need for a new and systematic approach to peer review within the discipline.

2.2. Artificial intelligence (AI) as a screener and reviewer

One promising avenue is the integration of artificial intelligence (AI) as a supplementary tool in the review process. One hypothesis is that AI systems could assist by evaluating technical rigor, checking for methodological consistency, identifying potential errors, and even providing initial assessments of the manuscript's contribution based on citation patterns and relevance to existing literature.

There is currently a small but growing body of research within computer science investigating AI's potential to enhance and improve the peer-review process. One notable example is Yuan et al. (2022), who explore the potential of natural language processing (NLP) systems to generate comprehensive and aspect-sensitive peer reviews for scientific papers. Using a dataset of machine learning papers annotated with aspect-based review information, they train and evaluate NLP models capable of producing comprehensive, aspect-sensitive review drafts. The study reveals that while the models can accurately summarize a paper's core ideas and provide broader aspect coverage than human reviewers, their reviews are often non-factual and lack constructive criticism. They conclude that

although the current technology is not yet ready to replace human reviewers, it has potential as a tool to assist reviewers and authors in identifying key strengths and weaknesses in manuscripts. While the paper finds some evidence of bias against non-native speakers in terms of clarity and perceived potential impact, it does not address biases related to affiliation, gender, or author prominence.

In another study, Checco et al. (2021) investigate the potential of AI to enhance the peer-review process by predicting review outcomes and identifying biases. Using a neural network model trained on peer review data from three academic conferences, the study incorporates features such as word distributions, readability metrics, and document formatting to predict reviewer decisions. The results indicate that AI can predict review outcomes with significant accuracy, suggesting that superficial features like formatting and readability correlate with reviewer judgments. However, the study raises concerns about algorithmic bias, as AI systems may reinforce biases already present in human reviewers, such as those related to language, regional representation, and first impressions. Like Yuan et al. (2022), they also caution that AI is not yet suitable to replace human reviewers. Nevertheless, they propose that editors and reviewers can use AI to pre-screen tasks, identify systematic biases, and improve the efficiency of the review process.

Despite recent advancements in AI and its potential applications in peer review, existing studies have predominantly relied on observational data rather than employing randomization to establish causal relationships. Consequently, a significant gap remains in understanding whether AI systems treat identical papers differently based on authors' affiliations, gender, or prominence. Furthermore, previous studies have not utilized already published papers that were previously evaluated by human reviewers, making it difficult to

directly compare AI-generated ratings with human judgments of a paper's overall quality. For example, it is unclear whether AI can reliably distinguish between papers accepted in top-tier journals versus those published in mid-tier journals. Addressing these limitations requires systematic research to benchmark AI systems against human assessments using papers with known publication outcomes, offering a clearer understanding of LLM's capabilities and potential biases in peer review. To address these limitations, this paper is one of the first in both economics and computer science to take an experimental approach to provide a rigorous evaluation of AI's capabilities and biases in the peer-review process within the field of economics.

3. Methods

To systematically assess the performance and potential biases of LLMs in evaluating economics papers, we simulate the peer review process by generating submissions with diverse attributes associated with each paper. For the base article in the submission, we randomly selected three papers each from *Econometrica*, *Journal of Political Economy*, and *Quarterly Journal of Economics* (**"high-ranked journals"** based on RePEc ranking) and three each from *European Economic Review*, *Economica*, and *Oxford Bulletin of Economics and Statistics* (**"medium-ranked journals"**). Additionally, we randomly selected three papers from each of the three lower-ranked journals not included in the RePEc ranking—*Asian Economic and Financial Review*, *Journal of Applied Economics and Business*, and *Business and Economics Journal* (**"low-ranked journals"**). To complete the dataset, we included three papers generated by GPT-o1 (**"fake AI papers"**), designed to match the standards of papers published in top-five economics journals. We selected GPT-o1 for its

state-of-the-art reasoning capabilities despite its limited adoption due to prohibitive per-token costs, which exceed those of standard commercial models by an order of magnitude. This approach enables us to assess whether advanced models can generate papers indistinguishable from human-authored research while acknowledging the practical constraints that currently hinder the widespread deployment of such models in academic settings (see the prompt in Appendix B).

This produces a total of 30 papers (9 total papers from top-five economics journals, nine total papers from mid-tier general economics journals, nine total papers from lower-ranked journals, and three total papers generated by AI).). All papers listed in Appendix A were published in 2024-2025.[1] This intentional selection ensured that the data had not yet been incorporated into the latest versions of AI systems, thus preventing LLM from having prior knowledge of where the papers were published or who the authors were. Furthermore, only the text of each paper—excluding the original authors and affiliations—was input into the LLM for evaluation.

We systematically varied each submission across three key dimensions: authors' affiliation, prominence, and gender. For affiliation, each submission was attributed to authors affiliated with: i) top-ranked economics departments in the US and UK, including Harvard University, Massachusetts Institute of Technology (MIT), London School of Economics (LSE), and Warwick University, ii) leading universities outside the US and Europe, including Nanyang Technological University (NTU) in Singapore, University of

[1]However, note that all three papers that were published in the most recent issue of the *Asian Economic and Financial Review (AEFR)* appeared online in 2022 rather than in 2024 or 2025.

Tokyo in Japan, University of Malaya in Malaysia, Chulalongkorn University in Thailand, and University of Cape Town in South Africa[2], and iii) no information about the authors' affiliation, i.e., blind condition.

To introduce variation in academic reputation, we replaced the original authors of the base articles with a new set of authors categorized into the following groups: (i) *prominent economists*—the top 10 male and female economists from the RePEc top 25% list; (ii) *lower-ranked economists*—individuals ranked near the bottom of the RePEc top 25% list; (iii) *non-academic individuals*—randomly generated names with no professional affiliation; and (iv) *anonymous authorship*—papers where author names were omitted. For non-anonymous authorship, we further varied each submission by gender, ensuring an equal split (50% male, 50% female). Combining these variations resulted in 9,030 unique papers, each with distinct author characteristics.[3]

We then utilized GPT4o-mini to assess each of the 9,030 submissions. Each submission was evaluated three times independently across the following ten dimensions; see the full LLM prompts in Appendix C:

1. **Top-five desk rejection Score:** This variable represents LLM's evaluation of whether a submission would advance past the desk review stage for a top-five

[2] The list of selected universities may appear arbitrary; however, 5 out of the 9 affiliations were either the authors' current or former affiliations. Thus, the selected affiliations were not only chosen to provide variation but were also personally relevant to the authors.

[3] More specifically, we have 9,000 non-blind submissions, i.e., 30 papers [3 papers per journal 3 Journal per type x 3 Journal types + 3 AI-generated papers] x 30 names [10 top + 10 bottom + 10 random] x 10 institutions = 9,000. As for the blind submissions, we have 30 papers, i.e., 3 papers per journal x 3 Journal per type x 3 Journal types + 3 AI-generated papers] = 30.

economics journal. The score ranges from 0 (“Definitely reject”) to 10 ("Definitely advance to peer review").

2. **Top-five acceptance Score:** This variable captures LLM’s evaluation of the likelihood that a submission would be accepted for publication following peer review at a top-five economics journal. The score ranges from 0 (“Definitely reject”) to 10 (“Definitely recommend for publication”).
3. **Top-five review recommendation score (without added criteria):** This variable represents LLM's evaluation in the role of a reviewer for a top-five economics journal. It captures AI’s recommendation for a submission based on a 6-point scale:
    - 1 = *Definite Reject*: Fatal flaws in theory/methodology, insufficient contribution, or serious validity concerns that make the paper unsuitable for the journal.
    - 2 = *Reject with Option to Resubmit*: Significant issues with theory, methodology, or contribution, but potentially salvageable with major revisions and fresh review.
    - 3 = *Major Revision*: Substantial changes are needed to theory, empirics, or exposition, but the core contribution is promising enough to warrant another round.
    - 4 = *Minor Revision*: Generally strong paper with a few small changes needed in exposition, robustness checks, or literature discussion.
    - 5 = *Very Minor Revision*: Excellent contribution, needing only technical corrections or minor clarifications.
    - 6 = *Accept As Is*: Exceptional contribution ready for immediate publication.

For ease of analysis and interpretation, we grouped AI's responses into three broader categories: scores (1, 2) were classified as "Reject/Reject with Option to Resubmit," scores (3, 4) as "Major/Minor Revision," and scores (5, 6) as "Accept as Is/Very Minor Revision."

4. **Top-five review recommendation score (with added criteria):** This variable extends the previous one by incorporating additional criteria typically expected for a top-five publication. The prompt explicitly specified that top-five papers should address broad, foundational questions, propose groundbreaking methodologies, and set new research standards. It emphasized generalizability, methodological rigor (e.g., state-of-the-art econometrics, robust identification strategies, thorough robustness checks), polished presentation (e.g., clear writing, compelling structure, visually intuitive results), high-quality data (e.g., novel datasets, rigorous documentation), and relevance (e.g., policy implications, significant impact). We also include the statement: "*Papers by well-known authors or prestigious institutions are more likely to receive attention and scrutiny during the review process.*" In other words, we implicitly instructed LLM to avoid assigning higher recommendation rates to papers by well-known authors or prestigious institutions without first thoroughly evaluating the quality of their submissions. The AI's responses were provided on the same 6-point scale as the previous variable.

We also aim to investigate the impact of each paper on other related domains, including:

5. **Predicted citations**: This variable represents LLM's prediction of the total number of citations the paper is expected to receive over the next 10 years. To enable the

interpretation of our results in terms of percentage change, we apply a natural logarithmic transformation to this variable.

6. **Funding competitiveness score**: This variable represents LLM's evaluation of a submission's competitiveness for major funding. The score ranges from 0 ("Definitely not fundable") to 10 ("Definitely fundable at the highest award level").
7. **Top conference acceptance score**: This variable represents LLM's evaluation of the likelihood that a submission would be accepted for presentation at a prestigious economics conference. The score ranges from 0 ("Definitely reject") to 10 ("Definitely accept for a prominent session").
8. **Research award score**: This variable represents LLM's evaluation of whether the work is competitive for prestigious recognition, with scores ranging from 0 ("Definitely not award-worthy") to 10 ("Definitely award-worthy").
9. **Tenure case strength score**: This variable represents LLM's perceived strength of a faculty member's case for tenure based on their submission. The score ranges from 0 ("Definitely deny tenure") to 10 ("Definitely grant tenure").
10. **Nobel potential score**: This variable represents LLM's evaluation of the long-term potential of a research agenda to meet the high standards of innovation, impact, and contribution required for the Nobel Prize in Economics. The score ranges from 0 ("Shows no indication of Nobel Prize potential") to 10 ("Shows definitive Nobel Prize potential)

This evaluation process produced 27,090 independent data points, with each submission being evaluated three times. To enhance reliability and reduce variance in the assessments, we calculated the average rating across the three evaluations for each dimension of every

paper. These aggregated scores were then utilized in the regression analysis. For the descriptive statistics of the variables used in the analysis, see Table 1A in Appendix A.[4]

We used ordinary least squares (OLS) regression with bootstrap standard errors (1,000 replications) to analyze outcomes on a Likert scale, specifically, those numbered 1–2 and 5–9. For ordinal outcomes numbered 4 and 5, we employed an ordered logit model with bootstrap standard errors to conduct the analysis. It is worth noting that, due to the structure of our data—where each paper equally experiences the same variation in authors' affiliation, prominence, and gender—the correlations between the independent variables are zero (or nearly zero). For details, see Table 2A in Appendix A. This implies that the coefficient of each independent variable in the regression remains completely stable, regardless of whether other variables are included in or excluded from the model. In other words, there is no need to account for omitted variable bias, as each independent variable is entirely uncorrelated with the others.

However, given the presence of ten outcome variables and multiple hypotheses being tested, our analysis faces an increased risk of Type I errors (i.e., "false positives"). To address the issue of multiple comparisons, we accounted for the family-wise error rate (FWER) across our dependent variables in subsequent analyses using the free step-down resampling method proposed by Westfall and Young (1993).

4. Results

[4] For access to data and codes used in this study's analysis, see https://github.com/mitmedialab/ai-peer-review-crisis.

Table 1 presents the OLS estimates of the effect of a paper's publication quality, based on the journal where it was published, on LLM's desk rejection and acceptance scores for a top-five economics journal. Looking at Column 1, we observe that, compared to the reference group ("low-ranked journals"), LLM rated papers published in mid-tier and top-five journals as statistically significantly more likely to advance to the peer review stage. The effects are substantial, approximately twice the size of the standard deviation of the variable, with coefficients of 3.839 (SE = 0.033, $p < 0.001$) for mid-tier journals and 3.994 (SE = 0.033, $p < 0.001$) for top-five journals. Additionally, we can reject the null hypothesis that the coefficients on mid-tier and top-five journals are equal ($p < 0.001$). This indicates that while the difference in effect size between mid-tier and top-five journals on LLM's desk rejection evaluation is slight, it remains highly significant at the 1% level. Interestingly, the LLM also rated fake AI-generated papers as significantly more likely to advance to peer review compared to papers published in low-ranked journals, with the estimated effect being on par with that of top-five publications.

While the estimates in Column 1 on publication quality suggest that LLM performs well in ranking top-five, mid-tier, and low-ranked journals, other results indicate the presence of biases from supposedly irrelevant factors in its judgments. Holding publication quality constant, papers authored by individuals from three of the four selected top US and UK economics departments—namely, Harvard, MIT, and LSE—are significantly more likely to advance to peer review compared to the reference group, where authors' affiliations were withheld from LLM. For example, papers authored by individuals affiliated with Harvard University are, on average, 0.207 points more likely than those in the reference group to advance past desk rejection (SE = 0.050, $p < 0.001$). In contrast, there is little

evidence to suggest that authors from other institutions received significantly higher or lower desk rejection scores compared to when authors' affiliations were withheld.

Turning to the influence of authors' names, there is strong evidence that top authors' papers receive, *ceteris paribus*, more favorable LLM evaluations in the top-five desk rejection scores than individuals in the reference group, namely authors ranked at the bottom of the top 25% authors in the RePEc ranking. The estimated coefficient for top authors (β = 0.378, SE = 0.026, $p < 0.001$) is approximately twice the magnitude of the coefficient for affiliation with the London School of Economics (LSE). While the effect of having random names on the top-five desk rejection score is statistically indistinguishable from authors ranked at the bottom of the top 25% authors in the RePEc ranking, the coefficient for blind submissions is negative, large, and statistically significant at -0.729 (SE = 0.276, $p < 0.001$). This suggests that, on average, the blind submissions ($N = 30$) performed significantly worse than papers with authors' names visible. Finally, there is evidence that holding publication quality constant, LLM judged submissions by female economists as around 0.1-point less likely to advance to peer review compared to submissions by male economists.

In Column 2, we replaced the publication quality categories with the specific journals in which the submissions were published. Consistent with the results in Column 1, submissions published in top-five journals, mid-tier journals, and AI-generated submissions performed significantly better than the reference group (i.e., *Asian Economics and Financial Review*, one of the three low-ranked journals). In contrast, the other two low-ranked journals performed significantly worse than the reference group.

Columns 3 and 4 replicate the analyses in Columns 1 and 2, but with the top-five acceptance score replacing the top-five desk rejection score as the dependent variable. Here, we obtained point estimates that were nearly qualitatively identical to those in Columns 1 and 2. More specifically, LLM continued to effectively distinguish between top-five publications, mid-tier publications, and low-ranked publications. Additionally, it rated submissions from authors affiliated with top institutions, as well as those authored by top-ranked and male economists, significantly higher than their counterparts.

One potential objection at this stage is whether we are genuinely comparing like for like, given that we are not conducting within-submission comparisons. As previously mentioned, given the structure of our design—where the correlations between independent variables are virtually zero—the estimates should remain consistent regardless of the inclusion of additional control variables and submission fixed effects. However, for completeness, we included submission fixed effects in our analysis and reported the results in Table 3A of Appendix A. Here, we observe that with the inclusion of submission fixed effects, the coefficients retain their magnitudes and statistical significance. In fact, the standard errors with submission fixed effects are notably smaller than in the previous estimates. For example, affiliation with Warwick University is now positive and statistically significant, at least at the 5% level in both regressions with the inclusion of submission fixed effects.

Are the biases stemming from authors' affiliation, reputation, and gender consistent across submissions of different publication quality? To investigate this, we conducted a sub-sample analysis of the top-five acceptance score, stratified by publication quality, and

presented in Figure 1 the predicted margins for authors' affiliation, reputation, and gender.[5] Due to the substantial gap in predicted margins between low-ranked publications and higher-quality publications, we also provide results omitting the predicted margins for low-ranked publications to better highlight the differences among mid-tier, top-five, and fake AI papers.

Looking across the panels in Figure 1, we observe a significant premium for being affiliated with a top institution (Harvard, MIT, LSE), being a top-ranked author according to the RePEc ranking, and being male, consistently across all levels of publication quality. This indicates that these biases are not limited to either low- or high-quality papers but are pervasive across all publication quality levels. The premium effects are sizable when compared across publication quality. For instance, a mid-tier quality publication authored by someone from MIT has the same predicted top-five acceptance score as a top-five quality publication authored by someone from NTU. Similarly, a mid-tier quality publication authored by a top-ranked author has the same predicted top-five acceptance score as a top-five quality publication authored by someone ranked at the bottom of the top 25% in the RePEc ranking or by someone with a random name.

One potential criticism of the results in Table 1 and Figure 1 is that editors may not base their judgments on Likert scale ratings alone. Instead, their decisions are likely guided by a set of well-defined thresholds, such as those for rejection, revision, or acceptance. To evaluate this, we tasked LLM with assessing each submission and providing recommendations for a top-five economics journal based on a 6-point scale: 1 = "Definitely

[5] For the regression results of the sub-sample analysis, refer to Table 4A in Appendix A.

reject," 2 = "Reject with option to resubmit," 3 = "Major revision," 4 = "Minor revision," 5 = "Very minor revision," and 6 = "Accept as is." For simplicity in analysis and interpretation, we grouped LLM's responses into three ordered categories: "Reject/reject with option to resubmit," "Major/minor revision," and "Accept as is/very minor revision." We then employed an ordered logit model to estimate the results and present the predicted margins for each of the three outcomes in Figure 2.[6]

Consistent with the results in Table 1 and Figure 1, higher-quality submissions significantly increased the likelihood of LLM recommending "Accept as is/very minor revision," with the predicted acceptance rate being highest for the top-five submissions. For instance, an estimated 30% of the top-five submissions received a recommendation of either "Accept as is" or "Very minor revision," compared to approximately 8% for mid-tier submissions. None of the low-ranked submissions received a recommendation of "Accept as is" or "Very minor revision," while approximately 60% of the low-ranked submissions were estimated to receive a decision of "Reject" or "Reject with option to resubmit." Slightly more than 90% of the mid-tier submissions received a recommendation of either "Major revision" or "Minor revision," and none of the mid-tier or top-five submissions received a desk rejection.

When holding publication quality constant, submissions from authors affiliated with Harvard, MIT, or LSE are, on average, about 2 percentage points more likely to receive a recommendation of "Accept as is" or "Very minor revision" compared to submissions with concealed author affiliations. In contrast, submissions from authors affiliated with

[6] For the ordered logit estimates, please refer to Table 5A in Appendix A.

Universiti Malaya in Malaysia are, on average, 1 percentage point less likely to receive the same recommendation compared to submissions with concealed affiliations. The reverse pattern is observed for the “Reject/Reject with option to resubmit” decision, with submissions from Harvard, MIT, or LSE being less likely, while those from Universiti Malaya are more likely to receive this outcome. Additionally, being a top author increases the likelihood of receiving a recommendation of “Accept as is” or “Very minor revision” by approximately 1.5 percentage points, whereas being female reduces this likelihood by about 2 percentage points compared to being male.

We further refined the LLM prompt by incorporating additional criteria regarding the types of research typically published in a top-five economics journal. Using this enhanced top-five review recommendation score with the added criteria, we re-estimate the ordered logit model and present the equivalent predicted margins in Figure 3, corresponding to those shown in Figure 2. Here, we observe that LLM recommends a higher rejection rate for lower-ranked submissions, increasing it from 60% to 80%. Additionally, there is a notable decrease in the proportion of “Accept as is” or “Very minor revision” recommendations, with top-five submissions dropping from 30% to around 12%, and mid-tier submissions declining from 8% to 4%. In contrast, there is a notable increase in the proportion of fake AI submissions receiving "Accept as is" or "Very minor revision" recommendations, rising from less than 10% to 12% when using the more refined prompt.

However, even with the added criteria and the implicit instruction to scrutinize papers by well-known authors and authors from prestigious institutions more thoroughly, we continue to observe evidence that authors from top institutions, top authors, and male authors perform significantly better than their counterparts for submissions of the same

quality. For example, authors from Harvard and MIT continue to receive, on average, a 2-percentage point higher probability of having their submissions—identical in quality to those of other submissions—accepted compared to authors with concealed affiliations, even after the introduction of additional criteria and implicit prompts.

To what extent are these effects confined solely to publication, or do they extend to other contexts involving judgment and decision-making processes similar to editorial decision-making? To address this question, Table 2 reports the effects of publication quality and authors' characteristics on various outcomes, including the log of predicted citations, funding competitiveness, top conference acceptance, research awards, tenure case strength, and Nobel potential.

As anticipated, mid-tier, top-five, and fake AI submissions consistently and significantly outperformed low-ranked submissions across various academic success measures. There is also evidence that top-five submissions outperform mid-tier submissions in specific outcomes, particularly in top conference acceptance scores, research award scores, and Nobel potential scores. Additionally, despite being of the same quality, submissions from authors affiliated with top institutions such as Harvard, MIT, and LSE are predicted to receive significantly more citations in 10 years compared to submissions from authors with retracted or hidden affiliation information. They are also significantly more likely to be recommended for competitive research grants, top conference placements, research awards, tenure, and even future Nobel Prizes in Economics.

As a robustness check, we applied the free step-down resampling method proposed by Westfall and Young (1993) to control the family-wise error rate (FWER) in multiple

hypothesis testing for eight of the ten outcomes.[7] The original and adjusted p-values are reported in Table 6A of Appendix A. Even after applying these very conservative corrections, where all independent variables are treated as belonging to the same family, our main findings on the influences of publication quality and authors' characteristics remain statistically robust, further reinforcing the reliability of our results.

5. Theoretical Implications

In this section, we present a theoretical model that formalizes editors' role in augmenting AI systems within the peer review process. Drawing on the empirical insights of this study, the model examines how intrinsic paper quality and author characteristics shape editorial and reviewer decisions, integrating both AI evaluations and human biases. It also explores the implications of these dynamics for the efficiency-equity trade-off in the peer review process.

Assume there are N papers submitted to a top-five economics journal indexed by $i = 1,2, \dots, N$. Each paper i is defined by the following two dimensions:

- **Intrinsic paper quality**, q_i, where $q_i \in \{L, M, H\}$, representing low, medium, and high-quality paper, respectively[8], and
- **Author characteristics**, A_i. We assume $A_i = (r_i, g_i, f_i)$, where r_i is the prominence of the author; g_i is the author's gender, and f_i is the author's institutional affiliation. We also assume that A_i is visible to both AI systems and human reviewers due to

[7] We did not include the ordered outcome variables – Top 5 Review Recommendation Score with and without the added criteria – in the analysis.

[8] These tiers (e.g., low-, medium-, and high-quality) are introduced for simplicity. However, in reality, intrinsic quality can be viewed as a continuous variable, reflecting a more nuanced spectrum of intrinsic paper quality.

the widespread use of pre-prints on platforms like NBER, SSRN, IZA, and RePEc and the single-blind submission policy in most economics journals. This makes anonymization difficult and ensures that biases associated with A_i influence evaluations.

We assume editors – whether consciously or unconsciously – maximize the journal's true utility for each submitted paper i, which can be expressed as:

$$u_i = \beta_q q_i + \beta_r r_i + \beta_g g_i + \beta_f f_i + \epsilon_i, \qquad (1)$$

where $\beta_q > 0$ represents the journal's weight on intrinsic quality; $\beta_r, \beta_g, \beta_f > 0$ captures biases associated with author prominence, gender, and affiliation; and ϵ_i is an idiosyncratic noise term reflecting factors unaccounted for in the model. The inclusion of author-related biases in the utility function reflects empirical evidence showing that prestigious affiliations and established reputations confer advantages in the peer review process, even when controlling for paper quality.

AI systems use LLM to evaluate submitted papers and generate scores that combine assessments of textual content with signals derived from author information. Specifically, the LLM assigns a score, s_i^{AI}, that can be written as:

$$s_i^{AI} = \alpha_q q_i + \alpha_b B_i + \eta_i, \qquad (2)$$

where $\alpha_q > 0$ reflects LLM's positive weight on intrinsic quality; $B_i = r_i + g_i + f_i$, i.e., a composite bias term capturing author attributes; $\alpha_b > 0$ reflects the extent to which LLM

incorporates author-related biases, and $\eta_i \sim N(0, \tau^2)$ is an AI-specific noise term. Since the widespread availability of economics pre-prints renders the double-blind submission strategy ineffective, the LLM used by the journal editors is inherently exposed to the same signals that underlie human biases. Consequently, while the LLM can efficiently process large volumes of submissions, it risks propagating biases toward prominent authors, male authors, and authors from elite institutions.

It is worth noting that, based on our findings, the LLM struggles to differentiate between AI-generated high-quality papers and genuine high-quality submissions. This underscores inherent limitations in α_q when evaluating novelty, originality, and the authenticity of data sets used in the analysis. In other words, editors cannot rely solely on current AI systems and must still depend on human reviewers to ensure the accurate evaluation of submissions.

Consequently, editors use the AI-generated scores alongside their heuristics to make an initial desk-rejection decision, $d_i \in \{0,1\}$, where $d_i = 1$ indicates rejection. The editor's decision rule combines AI-generated scores and human biases as follows:

$$d_i \begin{cases} 1 & \text{if } s_i^{AI} + \gamma_r r_i + \gamma_g g_i + \gamma_f f_i < \delta, \\ 0 & \text{otherwise.} \end{cases} \qquad (3)$$

Papers that are not desk rejected ($d_i = 0$) proceed to peer review, where reviewers assign scores:

$$s_i^{Human} = \lambda_q q_i + \lambda_r r_i + \lambda_g g_i + \lambda_f f_i + \xi_i, \qquad (4)$$

where $\lambda_q > 0$ reflects review weight on quality; $\lambda_r, \lambda_g, \lambda_f > 0$ denotes reviewer biases, and $\xi_i \sim N(0, \nu^2)$ represents reviewer-specific noise.

The editor's final acceptance decision combines AI and human scores. We can model the probability that an editor will accept the paper as follows:

$$P(a_i = 1) = \Phi\left(\frac{\omega s_i^{AI} + (1-\omega) s_i^{Human} - \kappa}{\sigma}\right), \quad (5)$$

where Φ is the cumulative distribution function of the standard normal distribution; $\omega \in [0,1]$ represents the weight placed on AI evaluations relative to human judgments; κ is the acceptance threshold, and σ captures the noise in the decision-making process. Eq.(5) thus represents the interaction between AI and human reviewers, with ω serving as a critical parameter that balances the efficiency of AI with the nuanced judgment of human reviewers.

Two key metrics, efficiency and equity, are introduced to assess the performance of the peer review process within this framework. Efficiency measures the system's ability to identify and accept high-quality papers. Formally, efficiency is defined as the proportion of accepted papers that are high-quality, which can be expressed as:

$$\text{Efficiency} = \frac{\sum_{i:q_i=H} P(a_i=1)}{\sum_i P(a_i=1)}. \quad (6)$$

This metric represents how well the review process is consistent with the journal's objective of maximizing the publication of high-quality research. Equity, on the other hand, assesses

the extent to which decisions are influenced by author characteristics rather than intrinsic quality. Equity is defined as:

$$\text{Equity} = 1 - \frac{Var(\gamma_r r_i + \gamma_g g_i + \gamma_f f_i)}{Var(u_i)}, \qquad (7)$$

where a higher value of equity indicates that decisions are primarily driven by intrinsic quality rather than biases. We also account for the potential amplification of biases by AI systems through the concept of bias amplification (Glickman & Sharot, 2024), which quantifies the extent to which AI magnifies human biases relative to their inherent influence:

$$\text{Bias Amplification} = \frac{\alpha_b}{\beta_r + \beta_g + \beta_f}. \qquad (7)$$

If this ratio is greater than 1, then LLM amplifies human biases, thus making the problem worse. Conversely, if the ratio is less than 1, then LLM mitigates human biases, making the peer review more equitable. This metric is particularly relevant for assessing the trade-offs between efficiency and fairness in AI-augmented peer review systems.

This theoretical framework provides valuable insights into integrating AI systems into the peer review process, particularly given the widespread availability of economics preprints and the single-blind submission policy adopted by most economics journals. Below, we explore these insights and their potential implications.

First, the model highlights the significant efficiency gains that AI systems can provide editors by enhancing the initial stages of the peer review process. Editorial desk rejections, which are typically a time-consuming task for editors, can be expedited by

leveraging LLMs' ability to evaluate textual quality across a large volume of submissions. Suppose the LLM's capacity to assess intrinsic quality is sufficiently high. In that case, AI systems can reliably filter out low-quality—and even medium-quality—papers, thereby allowing human reviewers to concentrate on evaluating only high-quality submissions.

However, as this study's findings highlight, any efficiency gains are accompanied by significant risks if the LLM's evaluations are influenced by author-related biases. Given the single-blind review system, α_b, the LLM's bias parameter, plays a crucial role in determining whether LLM evaluations reflect intrinsic quality (q_i) or disproportionately favor author prominence (r_i), institutional affiliation (f_i), or gender (g_i).

Second, the single-blind review system ensures that biases associated with author attributes are not only present but reinforced at multiple stages of the review process. The model shows that s_i^{AI}, the AI-generated score, includes a bias component $\alpha_b B_i$, where $B_i = r_i + g_i + f_i$, which may originate from training data that reflects historical inequities in academic publishing. Human reviewers, influenced by AI-generated scores and the visibility of author identities under single-blind review, are likely to amplify these biases (Glickman & Sharot, 2024). For example, the model suggests that human scores, s_i^{Human}, depend on both intrinsic quality (q_i) and biases ($\lambda_r r_i, \lambda_g g_i, \lambda_f f_i$). When combined with AI scores, the final decision probability $P(a_i = 1)$ may overweight author attributes, which further disadvantaged underrepresented researchers.

Third, the weight ω placed on LLM's evaluations relative to human input is central to the overall efficiency and equity of the peer review process. A high ω increases reliance on AI systems may exacerbate biases if α_b is large. Conversely, a low ω shifts greater decision-making authority to human reviewers, whose evaluations are influenced by their

own biases, $\lambda_r, \lambda_g, \lambda_f$. This dynamic calls for a calibrated approach that balances the efficiency of AI systems with the nuanced judgment of human reviewers while addressing biases at both levels.

Based on the above dynamics, we can recommend several targeted policy interventions to mitigate the inefficiencies and inequities arising from single-blind review systems in AI-augmented peer review.

First, journals should implement algorithms to explicitly reduce α_b when generating s_i^{AI}. This study's findings, illustrated in Figure 3, show that simply refining the prompt is insufficient. Bias correction should, therefore, involve training AI on anonymized or author-independent textual datasets where possible or applying post-hoc adjustments to the AI scores to remove the bias component. For instance, $\hat{s}_i^{AI} = s_i^{AI} - \alpha_b B_i$, where B_i reflects the author-related signals. This process ensures that AI systems focus more on the textual content and intrinsic quality of submissions rather than author identities.

Second, the model demonstrates that the relative weight ω is a key lever for balancing efficiency and equity. Journals could adopt context-specific weighting schemes, assigning higher ω values to desk rejection tasks where efficiency is paramount, and lower ω values to tasks requiring nuanced assessments of originality or policy relevance.

Third, since transparency is essential for ensuring trust in AI-augmented peer review systems, journals should provide editors and reviewers with detailed breakdowns of AI scores, including the relative contributions of textual quality and author characteristics. Such transparency allows human reviewers to critically evaluate AI recommendations and adjust their decisions accordingly.

Fourth, to mitigate the influence of single-blind review, journals could provide reviewers with structured rubrics that explicitly prioritize intrinsic quality and originality over author reputation, gender, and affiliation.

Finally, periodic audits of the peer review process are critical for identifying and addressing biases. These audits could evaluate acceptance rates across different author demographics and assess whether AI and human reviewers disproportionately favor particular groups. Findings from these audits could inform adjustments to the AI system, editorial policies, and reviewer guidelines.

6. Discussion and Conclusion

This study is among the first in economics and computer science to utilize a large-scale experimental design to investigate the potential role of artificial intelligence (AI) in addressing the peer review crisis within the field of economics. By using 27,090 evaluations of 9,030 unique paper submissions, we demonstrate that a large language model (LLM), such as GPT4o-mini, can enhance the efficiency of the initial review process by accurately distinguishing between recently published economics papers of varying journal quality. This reduces editorial workload and minimizes the need for extensive referee invitations. However, our findings also reveal significant limitations. Despite its analytical capabilities, the LLM struggles to reliably differentiate between AI-generated papers crafted to mimic "top five" journal standards and genuine high-quality submissions. Furthermore, our results uncover persistent biases in AI evaluations—favoring male authors, prominent economists, and individuals affiliated with elite institutions—reflecting and amplifying the inequities present in traditional human-led peer review.

One of the most encouraging findings is the LLM's ability to effectively differentiate submissions based on quality tiers. Using the results from Figure 2, we observe that approximately 30% of top-tier submissions were recommended for "Accept as Is" or "Very Minor Revision," compared to only 8% of mid-tier submissions. Meanwhile, none of the low-ranked submissions received this recommendation, with 60% of them flagged for rejection or resubmission. Notably, the remaining mid-tier submissions (over 90%) were rated as warranting "Major" or "Minor Revision," indicating the AI's nuanced recognition of mid-quality research. These results suggest that LLM aligns well with human reviewers in identifying quality gaps between tiers.

Moreover, the top-five acceptance rate for mid-quality submissions can be further reduced by refining the AI's evaluation prompts. As demonstrated in Figure 3, when stricter criteria were applied to the assessment of submissions—for instance, requiring the identification of groundbreaking methodologies, strong theoretical contributions, or broader policy relevance—the LLM's recommendations became even more selective. Under these refined prompts, the proportion of top-five submissions rated as "Accept as Is" dropped from 30% to 12%, while the acceptance rate for mid-tier submissions declined from 8% to 4%. This suggests that the LLM is not only capable of distinguishing between quality tiers but also responsive to the prioritization of higher standards in peer review.

Notably, the stricter prompts also resulted in a higher rejection rate for low-quality submissions, further improving the efficiency of the review process. For example, when tasked with identifying the least publishable work, the system flagged over 80% of low-tier submissions for rejection under stricter evaluation conditions. These findings indicate that the system can be fine-tuned to align with the editorial priorities of journals, whether those

priorities emphasize inclusivity or a narrower focus on only the most innovative research. Additionally, it is important to note that our sampled paper submissions were drawn from already published papers, suggesting that the actual acceptance rate for unpublished papers may be even lower than the numbers reported in this study.

One potential objection to integrating AI systems into the peer review process is the finding that LLMs struggle to reliably distinguish between AI-generated papers designed to mimic the standards of "top five" journals and genuine high-quality submissions. We fully acknowledge this limitation, which is why we argue that editors should not rely solely on AI-generated scores when making final decisions about whether to accept a paper. Our interpretation of the findings suggests that while LLMs are effective in evaluating textual coherence, methodology, and presentation, they fall short in critically assessing data integrity or identifying issues such as result manipulation, p-hacking, or fabricated findings. As a result, we maintain that human reviewers, including academic whistleblowers like members of the Data Colada team, remain an indispensable part of the peer review process, particularly in the post-screening or desk-rejection stages.

Another potential objection is whether the minor effects of author characteristics—despite being statistically significant—have any meaningful influence on editors' final decisions. Here, we argue that while the observed effects of author characteristics—such as institutional affiliation, gender, and prominence as a top economist—may appear modest in isolation (e.g., approximately 2% biases for each factor), their compounded effects can lead to significant disparities in outcomes within the highly competitive, winner-takes-all market of publishing in top-five economics journals.

Using real-world statistics from Card and DellaVigna (2013), where the top-five journals received 2,000 submissions but published only 140 papers (a 7% acceptance rate), even slight advantages can have outsized effects. To illustrate this dynamic, we use the following assumptions:

1. **Institutional Representation**: We assume that 80% of publishable submissions come from top institutions, such as leading universities in North America and Europe. This assumption reflects the well-documented overrepresentation of elite institutions in economics journal submissions and publications. Many of these journals disproportionately attract submissions from prominent departments due to their greater resources, visibility, and established networks.
2. **Gender Disparities**: Given the persistent underrepresentation of women in economics, we assume that men author 70% of publishable submissions. Studies have highlighted gender disparities in publication rates in top-tier journals, which stem from structural inequalities in mentorship, networking, and collaboration opportunities.
3. **Prominence Effects**: We assume that 30% of publishable submissions are authored by top economists (e.g., senior or highly visible scholars). This reflects the skewed distribution of publications in top-tier journals, where a smaller subset of highly productive and prominent researchers accounts for a disproportionate share of published articles.

These assumptions are not definitive but represent a plausible scenario based on empirical patterns observed in the literature, including the concentration of publications in elite institutions (Angrist et al., 2017), gender imbalances in top-tier economics journals (Hengel,

2017), and the advantages of prominence for established economists (Brogaard et al., 2014, 2024).

Under these assumptions, the compounded effects of small biases (e.g., 2% for institutional affiliation, 2% for gender, and 2% for prominence) create significant redistributions in publication outcomes. For example, male economists from top institutions—representing approximately 16.8% of all submissions (calculated as 70% male × 80% top institutions × 30% prominent economists = 16.8%)—may see their share of publications increase from 23 slots (16.8% of 140 slots) to 30 slots (21.4%). Conversely, women economists from non-elite institutions, representing 4.2% of submissions (calculated as 30% women × 20% non-elite institutions × 70% non-prominent economists = 4.2%), experience a sharp decline in their share of publications, from six slots (4.2%) to just three slots (2.1%)—a 50% reduction. Hence, these estimated numbers suggest how even small biases, when compounded, can exacerbate inequities in access to career-enhancing opportunities for underrepresented groups.

Another objection is that not all papers published in top-five journals are necessarily of higher quality than those in lower-tier journals, suggesting that top-five publications cannot be reliably used as a benchmark for quality. This is a valid objection, as several studies have shown that many papers published in top-five journals receive significantly fewer citations than those published in mid-tier journals (e.g., Oswald, 2007). However, while this may be true, our experiment demonstrates that, on average, the LLM evaluates the selected top-five journal papers used in this study as higher quality than those from mid-tier and lower-ranked publications.

One limitation of our findings is that we cannot directly compare the effect size of AI biases with human biases. While previous research provides some insights into the magnitude of human biases toward prominent authors (Huber et al., 2022), our study is not directly comparable, as effect sizes are likely context-dependent. For instance, we used papers that three of the top-five journals had already published. In contrast, the paper examined by Huber et al. (2022) is still a working paper and may require several rounds of revision before publication. This difference potentially explains the more considerable disparity in rejection rates between prominent and unknown authors observed in their study. It would also be highly challenging to conduct the same experiment on real economists to obtain their revealed preferences as we did with AI. Nevertheless, as we have argued earlier, even a tiny AI bias can have a significant impact on editorial decisions when journal space is limited.

Another limitation of our results is that the experimental design systematically varied author characteristics, such as institutional affiliation, reputation, and gender, but did not account for other potentially significant factors, such as race—which may be inferred from author names (Bertrand & Mullainathan, 2004)—or the geographic location of the author's institution. Biases against certain ethnicities, cultural identities, or regions may similarly influence both AI and human evaluations, but these factors remain unexplored in this study. This omission is significant given the increasing diversity of contributors to global academic research and the growing emphasis on regional representation in economics. Future research should address this gap by investigating the extent to which racial and regional biases may affect evaluations by LLMs.

Despite these limitations, this study provides valuable insights into both the potential and the challenges of AI-augmented peer review. By utilizing an experimental design commonly employed in lab and field settings, it highlights the strengths and weaknesses of LLMs in evaluating economics papers, offering a solid foundation for future research and practical improvements in the peer review process. These findings carry critical implications for both the efficiency and equity of academic publishing. On the one hand, the LLM's strong performance in distinguishing paper quality suggests that AI has considerable potential to streamline editorial workflows, especially in the early stages of desk rejection. On the other hand, its susceptibility to biases and inability to detect unethical practices highlights the need for cautious integration. By refining AI algorithms to prioritize intrinsic paper quality over author attributes, implementing post-hoc adjustments to mitigate biases, and adopting hybrid review models that integrate human judgment with AI evaluations, journals can leverage the advantages of AI while minimizing its risks to fairness in the peer review process.

More generally, the conclusions drawn from this research extend beyond publishing in economics. As journals increasingly face pressures to accelerate the peer review process without compromising quality, the idea of integrating AI systems in the peer review process offers a path forward. However, achieving this will require a concerted commitment from journals and the academic community to uphold transparency, fairness, and rigorous evaluation of LLM's impact on equity, efficiency, and ethical integrity in scholarly publishing.

References

Angrist, J., Azoulay, P., Ellison, G., Hill, R., & Lu, S. F. (2017). Economic research evolves: Fields and styles. *American Economic Review*, *107*(5), 293-297.

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review*, *94*(4), 991-1013.

Blank, R. M. (1991). The effects of double-blind versus single-blind reviewing: Experimental evidence from the American Economic Review. *American Economic Review*, 1041-1067.

Brogaard, J., Engelberg, J., & Parsons, C. A. (2014). Networks and productivity: Causal evidence from editor rotations. *Journal of Financial Economics*, *111*(1), 251-270.

Brogaard, J., Engelberg, J. E., Eswar, S. K., & Van Wesep, E. D. (2024). On the causal effect of fame on citations. *Management Science*, *70*(10), 7187-7214.

Card, D., & DellaVigna, S. (2013). Nine facts about top journals in economics. *Journal of Economic literature*, *51*(1), 144-161.

Card, D., & DellaVigna, S. (2020). What do editors maximize? Evidence from four economics journals. *Review of Economics and Statistics*, *102*(1), 195-217.

Card, D., DellaVigna, S., Funk, P., & Iriberri, N. (2020). Are referees and editors in economics gender neutral? *The Quarterly Journal of Economics*, *135*(1), 269-327.

Checco, A., Bracciale, L., Loreti, P., Pinfield, S., & Bianchi, G. (2021). AI-assisted peer review. *Humanities and Social Sciences Communications*, *8*(1), 1-11.

Ellison, Glenn. 2002. "The Slowdown of the Economics Publishing Process." *Journal of Political Economy* 110 (5): 947–993.

Glickman, M., & Sharot, T. (2024). How human–AI feedback loops alter human perceptual, emotional and social judgements. *Nature Human Behaviour*, 1-15.

Heckman, J. J., & Moktan, S. (2020). Publishing and promotion in economics: The tyranny of the top five. *Journal of Economic Literature*, *58*(2), 419-470.

Hengel, E. (2022). Publishing while female: Are women held to higher standards? Evidence from peer review. *Economic Journal*, *132*(648), 2951-2991.

Huber, J., Inoua, S., Kerschbamer, R., König-Kersting, C., Palan, S., & Smith, V. L. (2022). Nobel and novice: Author prominence affects peer review. *Proceedings of the National Academy of Sciences*, *119*(41), e2205779119.

Mehta, V., Mathur, A., Anjali, A. K., & Fiorillo, L. (2024). The application of ChatGPT in the peer-reviewing process. *Oral Oncology Reports*, 100227.

Oswald, A. J. (2007). An examination of the reliability of prestigious scholarly journals: evidence and implications for decision-makers. *Economica*, *74*(293), 21-31.

Westfall, P. H. (1993). *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. Wiley.

Yuan, W., Liu, P., & Neubig, G. (2022). Can we automate scientific reviewing? *Journal of Artificial Intelligence Research*, *75*, 171-212.

Table 1: Predicting AI's Recommendations for Review and Acceptance at a Top-5 Economics Journal: Ordinary Least Squares

Variables	(1) Top-5 Desk Rejection Score	(2) Top-5 Desk Rejection Score	(3) Top 5 Acceptance Score	(4) Top 5 Acceptance Score
Paper's quality (Reference: Low-ranked journal)				
Mid-tier journals	3.839***		3.259***	
	(0.035)		(0.028)	
Top 5 journals	3.994***		3.468***	
	(0.036)		(0.029)	
Fake AI papers	4.131***		3.435***	
	(0.036)		(0.029)	
Published journal (Reference: Asian Econ & Fin Review)				
Econometrica		2.795***		2.208***
		(0.051)		(0.035)
Quarterly Journal of Economics		3.144***		2.550***
		(0.037)		(0.031)
Journal of Political Economy		2.911***		2.417***
		(0.047)		(0.032)
European Economic Review		2.912***		2.343***
		(0.048)		(0.032)
Economica		2.820***		2.178***
		(0.048)		(0.031)
Oxford Bulletin of Econ & Statistics		2.653***		2.027***
		(0.046)		(0.034)
Journal of Applied Econ & Business		-1.139***		-1.389***
		(0.028)		(0.060)
Business and Economics Journal		-1.992***		-1.842***
		(0.085)		(0.042)
GPT-o1		3.087***		2.358***
		(0.038)		(0.032)
Affiliations (Reference: Retracted information on affiliation)				
Harvard	0.207***	0.207***	0.190***	0.190***
	(0.050)	(0.017)	(0.041)	(0.034)
MIT	0.286***	0.286***	0.239***	0.239***
	(0.047)	(0.021)	(0.040)	(0.033)
LSE	0.177***	0.177***	0.148***	0.148***
	(0.047)	(0.012)	(0.041)	(0.034)
Warwick	0.052	0.052***	0.048	0.048
	(0.045)	(0.013)	(0.041)	(0.034)
NTU	0.033	0.033**	0.033	0.033
	(0.048)	(0.016)	(0.040)	(0.034)
Tokyo	0.008	0.008	0.023	0.023
	(0.047)	(0.017)	(0.040)	(0.034)
Malaya	-.039	-.039**	-0.061	-0.061*

	(0.047)	(0.016)	(0.041)	(0.034)
Chulalongkorn	-.006	-.006	-0.021	-0.021
	(0.049)	(0.019)	(0.040)	(0.035)
Cape Town	0.000	0.000	-.003	-.003
	(0.049)	(0.019)	(0.041)	(0.034)
Author's name (Reference: Bottom 10 authors by gender in the RePec ranking)				
Top 10 authors by gender in the RePec ranking	0.378***	0.378***	0.300***	0.300***
	(0.026)	(0.051)	(0.022)	(0.019)
Random names	-.031	-.031	-0.018	-0.018
	(0.025)	(0.021)	(0.022)	(0.018)
Blind	-0.729***	-0.729***	-0.501**	-0.472***
	(0.276)	(0.028)	(0.226)	(0.172)
Author's gender (Reference: Male)				
Female	-0.099***	-0.099***	-0.058***	-0.058***
	(0.023)	(0.035)	(0.017)	(0.015)
Intercept	4.378***	8.216***	5.029***	8.313***
	(0.050)	(0.038)	(0.042)	(0.031)
Observations	9030	9030	9030	9030
R^2	0.765	0.813	0.777	0.840

Notes: Bootstrap standard errors (1,000 replications) are in parentheses. *** p<.01, ** p<.05, * p<.1. Dependent variables are based on the following GPT prompts: Top 5 Desk Rejection Score (0 = "Definitely reject", …, 10 = "Definitely advance to peer review") and Top 5 Acceptance Score (0 = "Definitely reject", …, 10 = "Definitely recommend for publication"). For each paper, we asked GPT to perform three independent evaluations. Consequently, each observation represents an average AI evaluation score derived from these three individual data points.

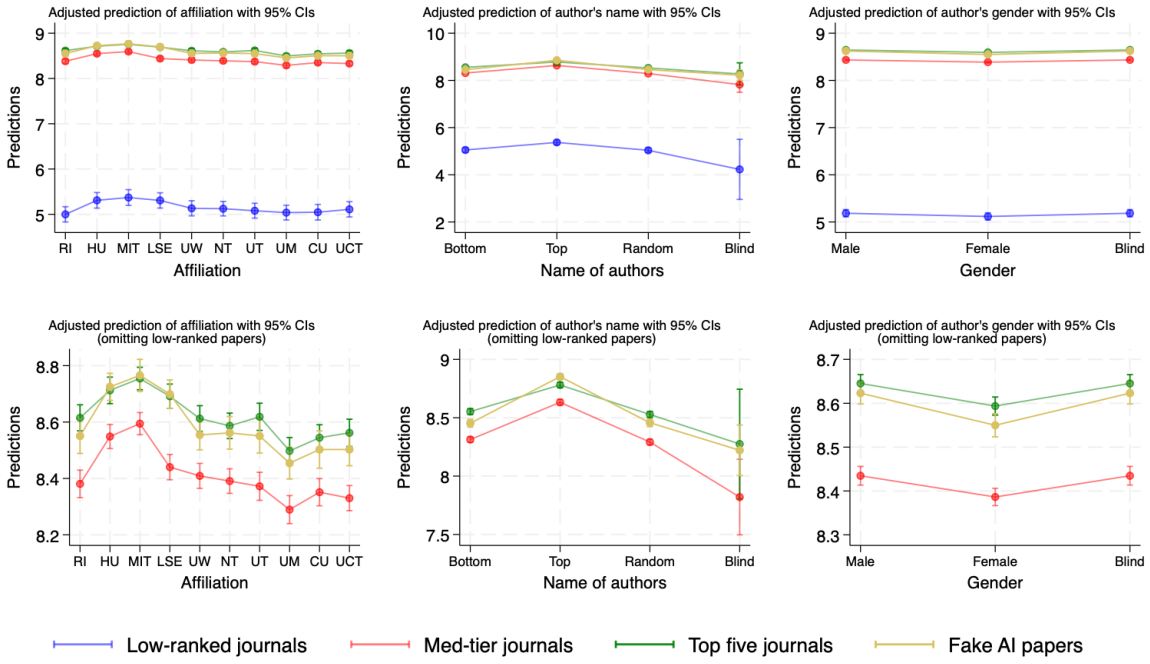

Figure 1: Predicted "Top 5 Acceptance Score" scores by authors' characteristics across publication quality categories. 95% confidence intervals based on bootstrap errors (1,000 replications) are displayed. Predictions are obtained from Table 3A's estimates in Appendix A.

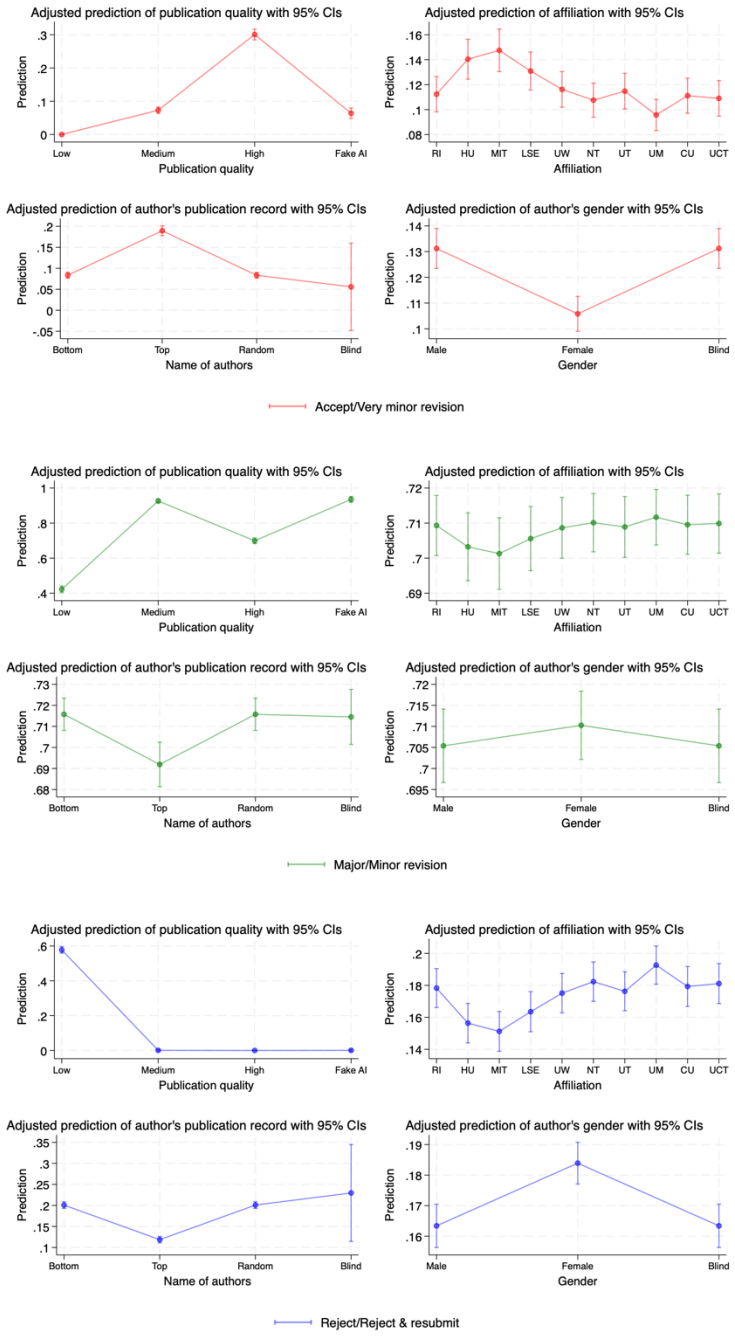

Figure 2: Predicted probabilities for the outcomes 'Accept/Very Minor Revision,' 'Major/Minor Revision,' and 'Reject/Reject & Resubmit' in the Top-5 Review Recommendation Score (without added criteria), characterized by publication quality and authors' characteristics. The graph includes 95% confidence intervals based on bootstrap errors (1,000 replications). Predictions are derived from the estimates in Table 4A (Appendix). Refer to Appendix C For the prompt to generate this outcome variable. For each paper, we asked GPT to perform three independent evaluations. Consequently, each observation represents an average AI evaluation score derived from these three individual data points.

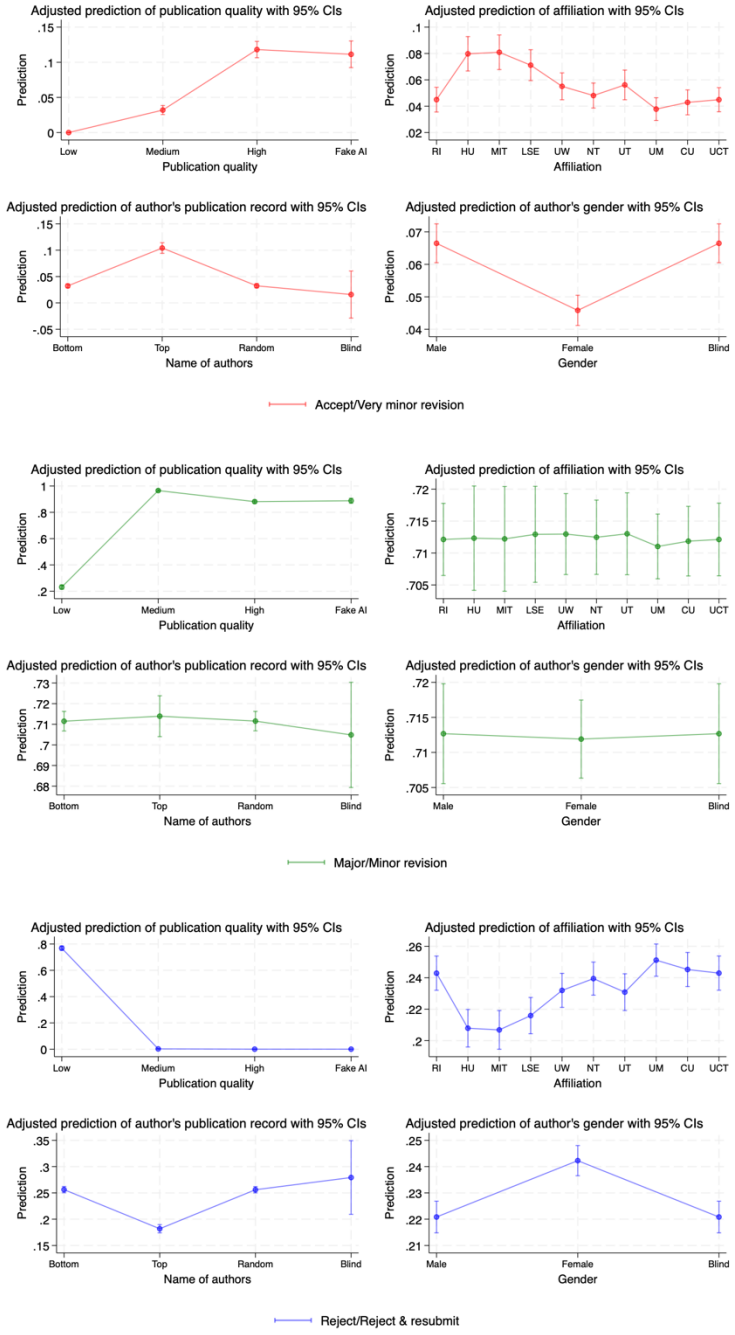

Figure 3: Predicted probabilities for the outcomes 'Reject/Reject & Resubmit' and 'Accept/Very Minor Revision' in the Top-5 Review Recommendation Score (with added criteria), characterized by publication quality and authors' characteristics. The added criteria include additional prompts, which specifically specify that top-5 papers should address broad, foundational questions, propose groundbreaking methodologies, and set new research standards. It emphasized generalisability, methodological rigor (e.g., state-of-the-art econometrics, robust identification strategies, thorough robustness checks), polished presentation (e.g., clear writing, compelling structure, visually intuitive results), high-quality data (e.g., novel datasets, rigorous documentation), and relevance (e.g., policy implications, significant impact). The AI's responses are on the same 6-point score as the previous variable. The figure includes 95% confidence intervals calculated using bootstrap errors (1,000 replications). Predictions are based on the estimates presented in Table 4A (Appendix).

Refer to Appendix C for the prompt used to generate this outcome variable. For each paper, we asked GPT to perform three independent evaluations. Consequently, each observation represents an average AI evaluation score derived from these three individual data points.

Table 2: Predicting AI's Recommendations for Different Academic Success Outcomes: Ordinary Least Squares

Variables	(1) Log(predicted citations)	(2) Funding competitiveness score	(3) Top conference acceptance score	(4) Research award score	(5) Tenure case strength score	(6) Nobel potential score
Paper's quality (Reference: Low-ranked journal)						
Mid-tier journals	0.474***	1.978***	1.736***	1.641***	1.636***	2.744***
	(0.011)	(0.021)	(0.018)	(0.015)	(0.019)	(0.021)
Top 5 journals	0.389***	1.989***	1.810***	1.672***	1.623***	3.002***
	(0.011)	(0.022)	(0.018)	(0.015)	(0.02)	(0.02)
Fake AI papers	0.743***	2.095***	1.934***	1.691***	1.647***	2.961***
	(0.011)	(0.022)	(0.018)	(0.015)	(0.021)	(0.021)
Affiliations (Reference: Retracted information on affiliation)						
Harvard	0.141***	0.247***	0.245***	0.186***	0.156***	0.173***
	(0.018)	(0.033)	(0.025)	(0.021)	(0.027)	(0.030)
MIT	0.114***	0.260***	0.276***	0.187***	0.179***	0.202***
	(0.017)	(0.031)	(0.024)	(0.021)	(0.027)	(0.030)
LSE	0.073***	0.165***	0.191***	0.106***	0.048*	0.091***
	(0.017)	(0.032)	(0.025)	(0.022)	(0.026)	(0.030)
Warwick	0.023	0.068**	0.097***	0.057***	0.002	0.02
	(0.018)	(0.033)	(0.025)	(0.021)	(0.028)	(0.030)
NTU	-.01	0.029	0.068***	0.006	-.007	-0.040
	(0.018)	(0.033)	(0.025)	(0.023)	(0.028)	(0.031)
Tokyo	0.008	0.046	0.080***	0.053**	0.003	0.003
	(0.018)	(0.032)	(0.025)	(0.022)	(0.029)	(0.028)
Malaya	-0.030	0.022	-.004	-0.050**	-.055*	-0.092***
	(0.019)	(0.033)	(0.026)	(0.022)	(0.029)	(0.031)
Chulalongkorn	-.007	0.004	0.015	-.019	-0.041	-.055*
	(0.018)	(0.032)	(0.025)	(0.022)	(0.028)	(0.032)
Cape Town	-.005	0.029	0.041	-0.011	-.062**	-0.049
	(0.018)	(0.031)	(0.026)	(0.022)	(0.029)	(0.030)
Author's name (Reference: Bottom 10 people by gender in the RePec ranking)						
Top 10 authors by gender in the RePec ranking	0.162***	0.167***	0.268***	0.165***	0.159***	0.351***
	(0.009)	(0.017)	(0.014)	(0.011)	(0.014)	(0.017)
Random names	0.021**	-0.022	-0.017	-.004	0.011	-0.025
	(0.01)	(0.016)	(0.014)	(0.012)	(0.016)	(0.015)
Blind	-0.130	-1.203***	-0.628**	-0.827***	-1.505***	-0.555***
	(0.083)	(0.280)	(0.261)	(0.210)	(0.352)	(0.171)

Author's gender (Reference: Male)						
Female	-0.049***	-.016	-0.044***	-0.018*	-0.015	-0.072***
	(0.008)	(0.013)	(0.011)	(0.01)	(0.012)	(0.013)
Intercept	4.187***	6.715***	6.849***	7.098***	7.074***	5.068***
	(0.016)	(0.036)	(0.027)	(0.023)	(0.030)	(0.029)
Observations	9030	9030	9030	9030	9030	9030
R^2	0.324	0.678	0.717	0.749	0.627	0.814

Notes: Bootstrap standard errors (1,000 replications) are in parentheses. *** p<.01, ** p<.05, * p<.1. Dependent variables include log of predicted citations, recommendation for a competitive research grant (0-10), recommend acceptance at a top economics conference (0-10), recommend for a prestigious research award (0-10), strong case for tenure (0-10), and future Nobel Prize winner (0-10). For the prompts used to generate these outcome variables, refer to Appendix C. For each paper, we asked GPT to perform three independent evaluations. Consequently, each observation represents an average AI evaluation score derived from these three individual data points.

Appendix A: List of the base research papers used in creating submissions

Paper #	Title	Abstract	Link	Journal	Published
1	Endogenous Liquidity and Capital Reallocation	This paper studies economies where firms acquire capital in primary markets and then, after idiosyncratic productivity shocks, retrade it in secondary markets that incorporate bilateral trade with search, bargaining, and liquidity frictions. We distinguish between full or partial sales (one firm gets all or some of the other's capital) and document several long- and short-run empirical patterns between these variables and the cost of liquidity, as measured by inflation. Quantitatively, the model can match these patterns plus the standard business cycle facts. We also investigate the impact of search frictions, monetary and fiscal policy, persistence in shocks, and returns to scale.	https://www.journals.uchicago.edu/doi/10.1086/732522	Journal of Political Economy	January 2025
2	The Value of Information in Competitive Markets: Evidence from Small and Medium-Sized Enterprises	We empirically investigate how the performance of small and medium-sized enterprises (SMEs) changes when gaining access to market information. To do so, we evaluate the impact of an information program diffused by a bank among its SME customers. Adopting firms gained access to reports with rich information about their own clientele and that of nearby establishments. While we find that adoption is associated with a 4.5% revenue increase, our instrumental variable results indicate that adoption increases revenue by 9%. The main mechanism driving our result is that the new information prompted adopting establishments to target gender-age customer groups underserved before adoption.	https://www.journals.uchicago.edu/doi/10.1086/732525	Journal of Political Economy	January 2025
3	Mobility for All: Representative Intergenerational Mobility Estimates over the Twentieth Century	We estimate long-run trends in intergenerational relative mobility for representative samples of the US-born population. Harmonizing all surveys that include father's occupation and own family income, we develop a mobility measure that allows for the inclusion of nonwhite individuals and women for the 1910s–1970s birth cohorts. We show that mobility increases between the 1910s and 1940s cohorts and that the decline of Black-white income gaps explains about half of this rise. We also find that excluding Black Americans, particularly women, considerably overstates the level of mobility for twentieth-century birth cohorts while simultaneously understating its increase between the 1910s and 1940s.	https://www.journals.uchicago.edu/doi/10.1086/732527	Journal of Political Economy	January 2025

4	Stationary Social Learning in a Changing Environment	We consider social learning in a changing world. With changing states, societies can be responsive only if agents regularly act upon fresh information, which significantly limits the value of observational learning. When the state is close to persistent, a consensus whereby most agents choose the same action typically emerges. However, the consensus action is not perfectly correlated with the state, because societies exhibit inertia following state changes. When signals are precise enough, learning is incomplete, even if agents draw large samples of past actions, as actions then become too correlated within samples, thereby reducing informativeness and welfare.	https://www.econometricsociety.org/publications/econometrica/2024/11/01/Stationary-Social-Learning-in-a-Changing-Environment	Econometrica	November 2024
5	Ambiguous Contracts	We explore the deliberate infusion of ambiguity into the design of contracts. We show that when the agent is ambiguity-averse and hence chooses an action that maximizes their minimum utility, the principal can strictly gain from using an ambiguous contract, and this gain can be arbitrarily high. We characterize the structure of optimal ambiguous contracts, showing that ambiguity drives optimal contracts toward simplicity. We also provide a characterization of ambiguity-proof classes of contracts, where the principal cannot gain by infusing ambiguity. Finally, we show that when the agent can engage in mixed actions, the advantages of ambiguous contracts disappear.	https://www.econometricsociety.org/publications/econometrica/2024/11/01/Ambiguous-Contracts	Econometrica	November 2024
6	Propagation and Amplification of Local Productivity Spillovers	The gains from agglomeration economies are believed to be highly localized. Using confidential Census plant-level data, we show that large industrial plant openings raise the productivity not only of local plants but also of distant plants hundreds of miles away, which belong to large multi-plant, multi-region firms that are exposed to the local productivity spillover through one of their plants. This “global” productivity spillover does not decay with distance and is stronger if plants are in industries that share knowledge with each other. To quantify the significance of firms' plant-level networks for the propagation and amplification of local productivity shocks, we estimate a quantitative spatial model in which plants of multi-region firms are linked through shared knowledge. Counterfactual exercises show that while large industrial plant openings have a greater local impact in less developed regions, the aggregate gains are greatest when the plants locate in well-developed regions, which are connected to other regions through firms' plant-level (knowledge-sharing) networks.	https://www.econometricsociety.org/publications/econometrica/2024/09/01/Propagation-and-Amplification-of-Local-Productivity-Spillovers	Econometrica	September 2024

7	Perceptions About Monetary Policy	We estimate perceptions about the Federal Reserve's monetary policy rule from panel data on professional forecasts of interest rates and macroeconomic conditions. The perceived dependence of the federal funds rate on economic conditions varies substantially over time, in particular over the monetary policy cycle. Forecasters update their perceptions about the Fed's policy rule in response to monetary policy actions, measured by high-frequency interest rate surprises, suggesting that they have imperfect information about the rule. Monetary policy perceptions matter for monetary transmission, as they affect the sensitivity of interest rates to macroeconomic news, term premia in long-term bonds, and the response of the stock market to monetary policy surprises. A simple learning model with forecaster heterogeneity and incomplete information about the policy rule motivates and explains our empirical findings.	https://academic.oup.com/qje/article/139/4/2227/7699086	The Quarterly Journal of Economics	June 2024
8	Global Firms in Large Devaluations	I investigate the consequences of firms' joint import and export decisions in the context of large devaluations. I provide empirical evidence that large devaluations are characterized by an increase in the aggregate share of imported inputs in total input spending and by reallocation of resources toward import-intensive firms, contrary to what standard quantitative trade models predict. These facts are explained by the expansion of exporters, which are intense importers. I develop a model where firms globally decide their import and export strategies and discipline it to match salient features of the Mexican micro data. After a devaluation, the model reproduces the pattern of low aggregate substitution and firm reallocation observed in the data. Compared with a benchmark without global firms, the model predicts higher growth of total exports and imports and a smaller reduction in the trade deficit.	https://academic.oup.com/qje/article/139/4/2427/7685536	The Quarterly Journal of Economics	November 2024
9	Using Divide-and-Conquer to Improve Tax Collection	Tax collection with limited enforcement capacity may be consistent with both high- and low-delinquency regimes: high delinquency reduces the effectiveness of threats, thereby reinforcing high delinquency. We explore the practical challenges of unraveling the high-delinquency equilibrium using a mechanism design insight known as divide-and-conquer. Our preferred mechanism takes the form of prioritized iterative enforcement (PIE). Taxpayers are ranked using the ratio of expected collection to capacity use. Collection threats are issued in small batches to ensure high credibility and induce high compliance. Following repayments, liberated capacity is used to issue the next round of threats. In collaboration with a district of Lima, we experimentally assess PIE in a sample of 13,432 property taxpayers. The data validate and refine our theoretical framework. A semi-structural model suggests that keeping collection actions fixed, PIE would increase tax revenue by roughly 10%.	https://academic.oup.com/qje/article/139/4/2475/7699856	The Quarterly Journal of Economics	November 2024

10	Revealing inequality aversion from tax policy and the role of non-discrimination	Governments have increasing access to individual information, but they exploit little of it when setting taxes. This paper shows how to reveal inequality aversion from observed tax policy choices of such governments. First, I map governments' priorities into concerns for vertical and horizontal equity. While vertical equity underlies inequality aversion, horizontal equity introduces a restriction against tax discrimination. This restriction affects the measurement of inequality aversion. Second, I apply the model to a hypothetical gender tax using Norwegian tax return data. The main result is that inequality aversion is overestimated when horizontal equity is ignored.	https://onlinelibrary.wiley.com/doi/full/10.1111/ecca.12567	Economica	January 2025
11	Monetary union effects on high inflation episodes	This paper analyses whether monetary union membership reduces the duration of high inflation episodes (HIEs). The study uses survival models estimated on a sample of 190 countries over the period 1950M01 to 2022M12. The results show that despite the often-cited issue of the heterogeneity of member countries, monetary unions significantly reduce the duration of HIEs, but not deflation episodes. This result remains robust to a battery of tests and is valid for both developed and developing countries. Furthermore, the results show that giving up monetary sovereignty in favour of an independent common central bank is more effective in terms of price stability than adopting inflation targeting. However, for countries seeking to preserve their monetary sovereignty, inflation targeting remains the best option for reducing the duration of HIEs. This performance of monetary unions in terms of price stability appears to be linked to the greater de facto independence of their central banks, the adoption of supranational fiscal rules, and the incentives to preserve the durability of the currency area. However, estimates show that the capacity of a monetary union to limit HIEs among its members diminishes as it expands to include new countries.	https://onlinelibrary.wiley.com/doi/full/10.1111/ecca.12564	Economica	December 2024
12	Establishment size and the task content of jobs: evidence from 46 countries	Using a mix of household- and employer-based survey data from 46 countries, we provide novel evidence that workers in larger establishments perform more non-routine analytical tasks, even within narrowly defined occupations. Moreover, workers in larger establishments rely more on the use of information and communication technologies to perform these tasks. We also document a 15% raw wage premium that workers in larger establishments enjoy relative to their counterparts in smaller establishments. A mediation analysis shows that our novel empirical facts on the task content of jobs are able to explain 5–20% of the establishment size wage premium, a similar fraction to what can be explained by selection of workers on education, gender and age.	https://onlinelibrary.wiley.com/doi/full/10.1111/ecca.12563	Economica	December 2024

13	Judicial Reform and Banks Credit Risk Exposure	The aim of this paper is to examine the impact of the Judicial System Reform, which was introduced in Italy in 2012, on the efficiency of the judicial system and the exposure of banks to credit risk in terms of Non-performing loans. To this end, we apply a difference-in-differences approach, using a dataset that covers annual judicial proceedings from 2010 to 2017, supplemented by bank balance sheet data. Our findings indicate that the reform had a detrimental effect on both judicial efficiency and the NPL ratio. The negative impact is especially pronounced in courts that were previously more efficient, suggesting that the court mergers may have resulted in diseconomies of scale.	https://onlinelibrary.wiley.com/doi/full/10.1111/obes.12652	Oxford Bulletin of Economics and Statistics	December 2024
14	Long-run Effects of Austerity: An Analysis of Size Dependence and Persistence in Fiscal Multipliers	This paper provides evidence that austerity shocks have long-run negative effects on GDP. Our baseline results show that contractionary fiscal shocks larger than 3% of GDP generate a negative effect of more than 5.5% on GDP even after 15 years. Evidence is also found linking austerity to smaller capital stock and total hours worked in the long-run. The results are robust to different fiscal shock datasets, the exclusion of particular shocks, and the use of cleaner controls. The paper also engages with the emerging discussion regarding fiscal multipliers heterogeneity, presenting evidence that the effects of exogenous fiscal measures are nonlinear on the shock size. The results also contribute to the broader discussion on the long-run effects of demand by suggesting that such shocks might permanently affect the economy.	https://onlinelibrary.wiley.com/doi/10.1111/obes.12646	Oxford Bulletin of Economics and Statistics	November 2024
15	Who Gets Vaccinated? Cognitive and Non-Cognitive Predictors of Individual Behaviour in Pandemics	This study investigates different cognitive and non-cognitive characteristics associated with individuals' willingness to get vaccinated against Covid-19 and their actual vaccination status. Our empirical analysis is based on data obtained from three survey waves conducted in 2021 among about 2,000 individuals living in the German state of North Rhine-Westphalia. We find that individuals with a high level of trait reactance – a personality characteristic that entails the personal tendency to perceive persuasion attempts as restricting one's freedom – display a significantly lower willingness to get vaccinated. They also tend to get inoculated later or never. Moreover, neuroticism, locus of control, and statistical numeracy appear to be associated with the willingness to get vaccinated, but these results are less pronounced and less robust. Our results indicate that vaccination campaigns and policies could be improved by specifically addressing those with a high level of trait reactance.	https://onlinelibrary.wiley.com/doi/10.1111/obes.12644	Oxford Bulletin of Economics and Statistics	October 2024

16	Firms and economic performance: A view from trade	We use transaction-level US import data to compare firms from virtually all countries in the world competing in a single destination market. First, we decompose countries' sales into the contribution of the number of firm-products, their average appeal and its dispersion. Then, by making distributional assumptions consistent with the data, we identify new structural parameters that are useful in understanding the role of firm heterogeneity for trade and economic performance. We find that differences in the dispersion of appeal are quantitatively important in explaining exports, even after controlling for selection, average appeal and other determinants of trade, and that they are relevant for welfare. We also find that countries with a higher GDP per capita export more per firm largely because they have a higher dispersion of appeal, hence more heterogeneous firms.	https://www.sciencedirect.com/science/article/pii/S0014292124002411	European Economic Review	February 2025
17	Time-varying stock return correlation, news shocks, and business cycles	The cross-sectional average of the pairwise correlations between U.S. stock returns is considered as a measure of risk to aggregate wealth priced by the stock market. We show that this measure predicts future U.S. output growth at a horizon of one to four years. A stronger average correlation of stock returns foreshadows significantly lower future output growth, even when controlling for some other widely used financial predictors. An innovation to average correlation gives rise to macroeconomic dynamics that resemble negative news about future total factor productivity (TFP) in a vector autoregression. TFP news shocks thus appear to be a key source of aggregate risk priced into stocks.	https://www.sciencedirect.com/science/article/pii/S0014292124002459	European Economic Review	February 2025
18	Cousins from overseas: How the existing workforce adapts to a massive forced return migration shock	The 1975 eruption of Civil Wars in Portuguese-speaking Africa sparked the return of half a million retornados to Portugal. We use census data from 1960 and 1981 to study the impacts of this massive influx of workers on the existing workforce. We observe gendered effects in natives' labour market outcomes: male and female natives leave dependent employment. We find robust evidence of females moving to inactivity, and suggestive evidence that males move into self-employment. The effects are driven by the repatriates who are Portuguese-born. The identification strategy exploits the repatriates' municipality of birth and a large-scale resettlement program relying on hotel capacity.	https://www.sciencedirect.com/science/article/pii/S001429212400254X	European Economic Review	February 2025

19	Accounting Conservatism in the Perspective of Positive Accounting Theory: A Study of Islamic Banking in Indonesia	Conservative accounting in Islamic banking is a crucial issue. This research aims to analyze the influence of executive compensation, the debt covenant, political cost, the composition of the commissioner board, the audit committee, and operating cash flow on the principle of accounting conservatism practiced in Islamic banking in Indonesia. Using data for 13 Islamic banks from 2014 to 2018 and employing panel regression, this study revealed that debt covenant, political cost, and operating cash flow significantly influence accounting conservatism. This result reconfirms the Positive Accounting Theory and Free Cash Flow Theory. However, the other three factors, i.e., executive compensation, the composition of the board of commissioners, and the audit committee were found to have no impact on accounting conservatism. From the findings, the study recommends that policy makers should improve the practice of good corporate governance in Islamic banking, thus the issue of conservative accounting methods could be minimized.	https://archive.aessweb.com/index.php/5002/article/view/4500	Asian Economic and Financial Review (AEFR)	2022
20	Performance Evaluation of Selected Islamic Banks in Bangladesh	The present study was carried out to evaluate the performance of selected Islamic banks in Bangladesh. Both quantitative and qualitative analyses were used. The relevant data and information were collected from relevant banks and stock exchanges. The performance of the banks was assessed through different variables, such as paid-up capital, investment-to-deposit ratio, classified investments, assets, net income, earnings per share (EPS) and dividends, which were then analyzed using various statistical measures, such as growth percentage, trend equations, the square of the correlation coefficient, and a correlation matrix. Fifty trend equations and R-squared were tested for ten different banks' activities. Among them, the trend values were positive for all the banks. The square of the correlation coefficient (R2) of most of the equations is more than 0.8, indicating well-fitting trend equations. This study proves that the industry has scope to grow.	https://archive.aessweb.com/index.php/5002/article/view/4507	Asian Economic and Financial Review (AEFR)	2022

21	Risk Disclosure, Corporate Governance and Firm Value in an Emerging Country	This study aims to examine whether risk disclosure practices and corporate governance mechanisms are associated with the performance of listed companies in Malaysia's emerging economy. The study uses fixed effects panel data regression models to gauge the relationship using 899 firm-year observations from companies that provide risk disclosures in their annual reports. The findings show that risk disclosure has a significant effect on firm performance. Audit committee monitoring also has a significant relationship with firm performance, while the results regarding the existence of a risk management committee are insignificant. In additional analyses, a composite measure of audit committee effectiveness confirms that its monitoring role improves firm performance significantly. This study addresses risk disclosure practices in an under-researched setting (Malaysia) with different corporate governance models and emerging risk reporting legislation, thus adding to the limited body of knowledge on corporate risk disclosure and corporate monitoring and their impact on firm performance.	https://archive.aessweb.com/index.php/5002/article/view/4516	Asian Economic and Financial Review (AEFR)	2022
22	Microeconomic Marvels Understanding Small-scale Markets	None	https://www.hilarispublisher.com/open-access/microeconomic-marvels-understanding-smallscale-markets-105761.html	Business and Economics Journal	2024
23	Trade Winds Navigating Global Markets and Policies	None	https://www.hilarispublisher.com/open-access/trade-winds-navigating-global-markets-and-policies-105765.html	Business and Economics Journal	2024
24	Profit Patterns Unlocking Financial Planning Strategies for Success	None	https://www.hilarispublisher.com/open-access/profit-patterns-unlocking-financial-planning-strategies-for-success-105762.html	Business and Economics Journal	2024

25	Explaining Combat Stress and its Effects on Decision Making in the Theatre of Operation	Combat stress and decision-making process in dynamic contexts are almost unexplored and still incompletely clear to the scientific community. On the other side, both are some key aspects of combat performance. Operational psychology is focused on using psychological knowledge in enhancing operational performance. First is necessary to understand their nature and relationship, in the direction of making strategies for training and dealing. Stress is productive when is optimal, while decision making process has its specifics in dynamic and risk contexts. Stress has an impact on decision making process, but their relationship is still unclear, as well as whether they can be seen as a related concept or separate phenomena. However, both are based on certain psychological processes, whose understanding and impact can lead to necessary changes in performance as a human behavior, and whose are base for mental readiness training programs. It is recommended to include these programs within the basic military training, to enhance military performance.	https://www.aebjournal.org/article120201.php	Journal of Applied Economics and Business	2024
26	Machine Scheduling and Spreadsheet Modeling in a Fashion Management Class	This paper is a result of several semesters of teaching an Operations Management class in the Fashion Management program. Topics like flow shop and parallel machine models are traditionally discussed in an Industrial Engineering class. However, we discussed these topics successfully in a fashion management class by providing examples that our students could relate to. This included examples in apparel production and fashion retailing. The students were also able to learn several logical Excel based formulas while creating spreadsheets for these topics. For the flow shop model, we create a spreadsheet model that reflects all details of the Gantt chart including idle times. In the case of parallel machines, the spreadsheet involves the more complex Nested IF functions. The students showed enthusiasm even when creating the more challenging spreadsheet models.	https://www.aebjournal.org/article120202.php	Journal of Applied Economics and Business	2024

27	The Impact of Remittances on Exchange Rates in West African Monetary Zone (WAMZ)	The study investigates the impact of remittances on the real exchange rate of West African Monetary Zone (WAMZ) member countries by using annual data from six countries from 1960 to 2022. The WAMZ member countries are Ghana, The Gambia, Nigeria, Guinea, Liberia and Sierra Leone. Remittances are important sources of foreign capital for developing countries including WAMZ. The study is unique because it examines three periods namely: Pre- WAMZ (1960-2000), During WAMZ (2001-2022) and the entire period (1960-2022); captures and compares how the increase in receipt of migrant remittances have affected the real exchange rate for the different periods given the huge increase in remittances in the last two decades; use updated data for longer period; shows how the different explanatory variables changes in the three periods examined. Multicollinearity tests, results reveal no multicollinearity among the variables.	https://www.aebjournal.org/article120101.php	Journal of Applied Economics and Business	2024
28	The Impact of Digitalization on Labor Markets: Evidence from Regional Broadband Penetration	This paper examines how the expansion of broadband internet access—arguably the backbone of modern digital infrastructure—affects labor market outcomes in advanced economies. Using a unique panel dataset at the regional level across multiple OECD countries between 2005 and 2020, we exploit plausibly exogenous variation in broadband rollout stemming from historical differences in local infrastructure and public policy mandates. We find that increases in broadband penetration lead to shifts in local labor markets, primarily via an increase in high-skill occupations and a decline in routine and low-skill employment. We explore different potential mechanisms behind these changes and present evidence that skill-biased technological change and firm-level organizational innovations both play a role. Our findings suggest that while digitalization offers opportunities for productivity growth and upward mobility, it also risks exacerbating earnings inequality across regions and skill categories, underscoring the importance of targeted public policies to foster inclusive economic growth.	NA	GPT-o1	

29	Monetary Shocks, Labor Market Frictions, and Income Inequality: A Heterogeneous Agents Approach	This paper investigates the relationship between monetary policy shocks, labor market frictions, and income inequality in a heterogeneous agents model. We develop a dynamic stochastic general equilibrium (DSGE) framework in which households differ in skill levels, wealth, and labor force attachment. Our model introduces search-and-matching frictions in the labor market and accommodates incomplete asset markets, thereby capturing the ways in which aggregate shocks have unequal effects on households. We calibrate the model to United States data and estimate that expansionary monetary policy, while boosting average output and employment, magnifies income disparities under certain conditions of labor market stickiness. Targeted transfers and labor market reforms that reduce matching frictions mitigate this adverse distributional effect. Our results suggest that policymakers should consider inequality implications when designing monetary policy, highlighting the importance of coordination between monetary and labor market policies.	NA	GPT-o1	
30	The Labor Market Effects of a Temporary Universal Basic Income: Evidence from a Randomized Experiment	This paper examines the short-run labor market effects of a Temporary Universal Basic Income (TUBI) program using data from a large-scale randomized experiment in Country X. We randomly assigned monthly transfers to a representative sample of working-age adults over a period of twelve months. Leveraging experimental variation, we estimate the causal impact of TUBI on labor supply, job search, and overall household wellbeing. Contrary to traditional concerns that such transfers discourage labor market participation, our findings indicate modest effects on labor supply—specifically, a small but statistically significant positive effect on both employment and earnings for some subgroups. Furthermore, TUBI recipients reported substantial improvements in financial security and psychological wellbeing. Our results contribute to the active policy debate surrounding basic income guarantees, suggesting that well-designed transfer programs may reduce material hardship without severely diminishing labor market participation.	NA	GPT-o1	

Table 1A: Description Statistics

Variables	Mean	SD	Min	Max
Top 5 desk rejection score	7.276	2.087	0.330	10.000
Top 5 acceptance score	7.513	1.768	2.000	9.667
Top-5 review recommendation score (without added criteria)	1.945	0.538	1.000	3.000
Top-5 review recommendation score (with added criteria)	1.825	0.507	1.000	3.000
Log of predicted citations	4.586	0.451	2.814	6.279
Funding competitiveness score	8.237	1.128	1.670	9.670
Top conference acceptance score	8.266	0.994	2.000	10.000
Research award score	8.354	0.892	2.670	9.670
Tenure case strength score	8.282	0.961	1.330	9.670
Nobel potential score	7.184	1.487	2.000	9.333
Harvard University (HU)	0.100	0.300	0.000	1.000
Massachusetts Institute of Technology (MIT)	0.100	0.300	0.000	1.000
London School of Economics (LSE), UK	0.100	0.300	0.000	1.000
University of Warwick, UK	0.100	0.300	0.000	1.000
Nanyang Technological University (NTU), Singapore	0.100	0.300	0.000	1.000
University of Tokyo, Japan	0.100	0.300	0.000	1.000
Universiti Malaya, Malaysia	0.100	0.300	0.000	1.000
Chulalongkorn University, Thailand	0.100	0.300	0.000	1.000
University of Cape Town, South Africa	0.100	0.300	0.000	1.000
Top 10 authors by gender in the RePec ranking	0.332	0.471	0.000	1.000
Random names	0.332	0.471	0.000	1.000
Blind	0.003	0.058	0.000	1.000
Female	0.498	0.500	0.000	1.000
Publication quality: mid-tier publication	0.300	0.458	0.000	1.000
Publication quality: top-5 publication	0.300	0.458	0.000	1.000
Publication quality: Fake AI paper	0.100	0.300	0.000	1.000

Table 2A: Pairwise Correlations of Independent Variables

Variables	(1)	(2)	(3)	(4)
(1) Affiliation	1.000			
(2) Name of authors	-0.013	1.000		
(3) Gender	-0.015	0.024	1.000	
(4) Publication quality	-0.000	0.000	-0.000	1.000

Table 3A: Predicting AI's Recommendations for Review and Acceptance at a Top-5 Economics Journal: Ordinary Least Squares with Submission Fixed Effects

Variables	(1) Top-5 Desk Rejection Score	(2) Top 5 Acceptance Score
Affiliations (Reference: Retracted information on affiliation)		
Harvard	0.207***	0.190***
	(0.022)	(0.017)
MIT	0.286***	0.239***
	(0.023)	(0.017)
LSE	0.177***	0.148***
	(0.021)	(0.016)
Warwick	0.052**	0.048***
	(0.021)	(0.018)
NTU	0.033	0.033*
	(0.02)	(0.018)
Tokyo	0.008	0.023
	(0.022)	(0.017)
Malaya	-.039*	-0.061***
	(0.021)	(0.017)
Chulalongkorn	-.006	-0.021
	(0.02)	(0.017)
Cape Town	0.000	-.003
	(0.022)	(0.017)
Author's name (Reference: Bottom 10 authors by gender in the RePec ranking)		
Top 10 authors by gender in the RePec ranking	0.378***	0.300***
	(0.012)	(0.009)
Random names	-.031***	-0.018*
	(0.011)	(0.009)
Blind	-0.729***	-0.501***
	(0.178)	(0.117)
Author's gender (Reference: Male)		
Female	-0.099***	-0.058***
	(0.009)	(0.008)
Intercept	7.141***	7.390***
	(0.026)	(0.024)
Submission fixed effects	Yes	Yes
Observations	9030	9030
R^2	0.955	0.956

Notes: Bootstrap standard errors (1,000 replications) are in parentheses. *** p<.01, ** p<.05, * p<.1. Dependent variables are based on the following GPT prompts: Top 5 Desk Rejection Score (0 = "Definitely reject", …, 10 = "Definitely advance to peer review") and Top 5 Acceptance Score (0 = "Definitely reject", …, 10 = "Definitely recommend for publication"). For each paper, we asked GPT to perform three independent evaluations. Consequently, each observation represents an average AI evaluation score derived from these three individual data points.

Table 4A: Predicting Top 5 Recommend Accepting by Publication Quality: Ordinary Least Squares

Variables	(1) Low-ranked publication	(2) Mid-tier publication	(3) Top 5 publication	(4) Fake AI papers
Affiliations (Reference: Retracted information on affiliation)				
Harvard	0.310**	0.168***	0.098***	0.174***
	(0.121)	(0.033)	(0.034)	(0.040)
MIT	0.372***	0.214***	0.140***	0.215***
	(0.124)	(0.031)	(0.031)	(0.044)
LSE	0.307**	0.059*	0.077**	0.148***
	(0.121)	(0.033)	(0.031)	(0.042)
Warwick	0.133	0.028	-.002	0.004
	(0.122)	(0.034)	(0.033)	(0.041)
NTU	0.125	0.01	-0.028	0.011
	(0.118)	(0.034)	(0.032)	(0.042)
Tokyo	0.080	-0.009	0.004	0
	(0.119)	(0.035)	(0.034)	(0.044)
Malaya	0.038	-0.091**	-0.117***	-0.096**
	(0.119)	(0.036)	(0.034)	(0.042)
Chulalongkorn	0.047	-0.030	-0.070**	-0.048
	(0.122)	(0.035)	(0.034)	(0.046)
Cape Town	0.111	-0.051	-0.053	-0.048
	(0.122)	(0.034)	(0.034)	(0.043)
Author's name (Reference: Bottom 10 authors by gender in the RePec ranking)				
Top 10 authors by gender in the RePec ranking	0.320***	0.319***	0.228***	0.399***
	(0.067)	(0.017)	(0.017)	(0.021)
Random names	-.014	-0.021	-0.025	0.004
	(0.067)	(0.017)	(0.018)	(0.024)
Blind	-0.823	-0.493***	-0.278	-0.232**
	(0.651)	(0.165)	(0.240)	(0.112)
Author's gender (Reference: Male)				
Female	-0.069	-0.048***	-0.052***	-0.073***
	(0.055)	(0.015)	(0.015)	(0.018)
Intercept	4.935***	8.308***	8.574***	8.454***
	(0.099)	(0.028)	(0.027)	(0.035)
Observations	2709	2709	2709	903
R^2	0.021	0.188	0.117	0.382

Notes: Bootstrap standard errors (1,000 replications) are in parentheses. *** p<.01, ** p<.05, * p<.1. Dependent variables are based on the following GPT prompt: Top 5 recommend acceptance (0-10). For the prompt used to generate this outcome variable ('Top 5 Accept Rating Without Criteria'), refer to Appendix C. For each paper, we asked GPT to perform three independent evaluations. Consequently, each observation represents an average AI evaluation score derived from these three individual data points.

Table 5A: Predicting Top 5 Recommendation Categories With and Without Added Criteria: Ordered Logit

Variables	(1) Top 5 Recommendation Categories without added criteria	(2) Top 5 Recommendation Categories with added criteria
Paper's quality		
(Reference: Low-ranked journal)		
Mid-tier journals	7.775***	7.568***
	(2.784)	(0.371)
Top 5 journals	9.563***	9.024***
	(2.785)	(0.402)
Fake AI papers	7.627***	8.953***
	(2.787)	(0.396)
Affiliations		
(Reference: Retracted information on affiliation)		
Harvard	0.319***	0.677***
	(0.123)	(0.153)
MIT	0.393***	0.695***
	(0.124)	(0.156)
LSE	0.216*	0.537***
	(0.124)	(0.156)
Warwick	0.047	0.233
	(0.129)	(0.162)
NTU	-0.060	0.076
	(0.120)	(0.166)
Tokyo	0.030	0.255
	(0.127)	(0.161)
Malaya	-0.218*	-0.195
	(0.125)	(0.173)
Chulalongkorn	-0.015	-0.053
	(0.124)	(0.173)
Cape Town	-0.042	-.001
	(0.122)	(0.165)
Author's name		
(Reference: Bottom 10 authors by gender in the RePec ranking)		
Top 10 authors by gender in the RePec ranking	1.140***	1.341***
	(0.069)	(0.082)
Random names	-.002	0.005
	(0.069)	(0.097)
Blind	-0.477	-0.745
	(0.713)	(1.566)
Author's gender		
(Reference: Male)		
Female	-0.302***	-0.436***
	(0.055)	(0.071)
/cut1	0.624***	1.786***
	(0.106)	(0.142)
/cut2	10.761***	11.692***

	(2.785)	(0.416)
Observations	9030	9030
Pseudo R^2	0.420	0.559

Notes: Bootstrap standard errors (1,000 replications) are in parentheses. *** p<.01, ** p<.05, * p<.1. Responses in the dependent variables range from 1 “Reject/Reject with an option to resubmit”, 2 “Major revision/minor revision”, and 3 “Accept as is/Very minor revision”. The main difference between the two dependent variables is in the added detail prompt on additional criteria as understood by GPT that define the characteristics of a top-5 paper. For the prompt used to generate this outcome variable ('Top 5 Accept Rating Without Criteria'), refer to Appendix C.

Table 6A: Multiple Hypothesis Testing: standard and Young *p*-values

Variables	Top-5 Desk Rejection Score	Top 5 Acceptance Score	Log(predicted citations)	Research grant	Top conference	Research award	Receiving tenure	Nobel Prize potential
Affiliations (Reference: Retracted information on affiliation)								
Harvard	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	0.018	*0.018*	*0.004*	*0.003*	*0.018*	*0.018*	*0.018*	*0.018*
MIT	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	0.018	*0.018*	*0.016*	*0.003*	*0.001*	*0.003*	*0.017*	*0.014*
LSE	0.000	0.000	0.000	0.000	0.000	0.000	0.084	0.108
	0.024	*0.024*	*0.018*	*0.018*	*0.004*	*0.018*	*0.903*	*0.153*
Warwick	0.273	0.222	0.188	0.024	0.000	0.007	0.935	0.508
	0.999	*0.997*	*0.994*	0.554	*0.024*	*0.231*	*1.000*	*1.000*
NTU	0.493	0.403	0.558	0.345	0.006	0.774	0.790	0.182
	1.000	*1.000*	*1.000*	*1.000*	*0.218*	*1.000*	*1.000*	*0.993*
Tokyo	0.870	0.566	0.636	0.125	0.001	0.012	0.918	0.922
	1.000	*1.000*	*1.000*	*0.957*	*0.086*	*0.355*	*1.000*	*1.000*
Malaya	0.412	0.123	0.088	0.468	0.858	0.017	0.045	0.002
	1.000	*0.957*	*0.911*	*1.000*	*1.000*	*0.450*	*0.754*	*0.101*
Chulalongkorn	0.898	0.599	0.670	0.896	0.543	0.366	0.135	0.070
	1.000	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*0.969*	*0.857*
Cape Town	0.997	0.948	0.782	0.332	0.100	0.615	0.026	0.109
	1.000	*1.000*	*1.000*	*1.000*	*0.928*	*1.000*	*0.575*	*0.937*
Author's name (Reference: Bottom 10 people by gender in the RePec ranking)								
Top 10 authors by gender in the RePec ranking	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	0.000	*0.000*	*0.000*	*0.001*	*0.000*	*0.000*	*0.001*	*0.000*
Random names	0.231	0.416	0.032	0.193	0.997	1.000	1.000	0.969
	0.997	*1.000*	*0.636*	*0.994*	*1.000*	*1.000*	*1.000*	*1.000*
Blind	0.000	0.001	0.061	0.000	0.000	0.000	0.000	0.018
	0.024	*0.086*	*0.840*	*0.001*	*0.018*	*0.001*	*0.000*	*0.018*
Author's gender (Reference: Male)								

Female	0.000	0.001	0.000	0.248	0.000	0.061	0.234	0.000
	0.018	*0.065*	*0.018*	*0.999*	*0.020*	*0.840*	*0.997*	*0.018*
Paper's quality (Reference: Low-ranked journal)								
Mid-tier journals	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	0.000	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*
Top 5 journals	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	0.000	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*
Fake AI papers	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	0.000	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*

Notes: Both standard and adjusted p-values are derived using the *wyoung* command in STATA. The adjusted Young p-values are in italics. All independent variables were included in the family of tests for FWER adjustment to control for Type I error across the set of hypotheses.

Appendix B: Prompts for generating Fake AI papers

Prompt: “Generate a paper that could be published in a top-5 economics journal.”
Model: GPT-o1
Response format: Text
Reasoning Effort: Medium

Appendix C: Prompts for Submission Evaluation

Top-five desk rejection Score

"In your capacity as an editorial board/reviewer for one of the most prestigious and highly selective top-five economics journals (such as Econometrica, Journal of Political Economy, or The Quarterly Journal of Economics), please determine whether you would allow this submission to proceed past the desk review stage (0 = you would definitely reject at desk, 10 = you would definitely advance to peer review).",

Top-five acceptance Score

"In your capacity as a reviewer for one of the most prestigious and highly selective top-five economics journals (such as Econometrica, Journal of Political Economy, or The Quarterly Journal of Economics), please determine whether you would recommend this submission for publication (0 = you would definitely reject, 10 = you would definitely recommend for publication).",

Top-five review recommendation score

"In your capacity as a reviewer for one of the most prestigious and highly selective top-five economics journals (such as Econometrica, Journal of Political Economy, or The Quarterly Journal of Economics), please determine whether you would recommend this submission for publication using the following 7-point scale: 1 = Definite Reject: Fatal flaws in theory/methodology, insufficient contribution, or serious validity concerns that make the paper unsuitable for the journal, 2 = Reject with Option to Resubmit: Significant issues with theory, methodology, or contribution, but potentially salvageable with major revisions and fresh review, 3 = Major Revision: Substantial changes needed to theory, empirics, or exposition, but the core contribution is promising enough to warrant another round, 4 = Minor Revision: Generally strong paper with few small changes needed in exposition, robustness checks, or literature discussion, 5 = Very Minor Revision: Excellent contribution needing only technical corrections or minor clarifications, 6 = Accept As Is: Exceptional contribution ready for immediate publication",

Top-five review recommendation score (with added criteria)

"In your capacity as a reviewer for one of the most prestigious and highly selective top-5 economics journals (such as Econometrica, Journal of Political Economy, or The Quarterly Journal of Economics), please determine whether you would recommend this submission for publication using the following 7-point scale: 1 = Definite Reject: Fatal flaws in theory/methodology, insufficient contribution, or serious validity concerns that make the paper unsuitable for the journal, 2 = Reject with Option to Resubmit: Significant issues with theory, methodology, or

contribution, but potentially salvageable with major revisions and fresh review, 3 = Major Revision: Substantial changes needed to theory, empirics, or exposition, but the core contribution is promising enough to warrant another round, 4 = Minor Revision:

Generally strong paper with few small changes needed in exposition, robustness checks, or literature discussion, 5 = Very Minor Revision: Excellent contribution needing only technical corrections or minor clarifications, 6 = Accept As Is: Exceptional contribution ready for immediate publication; Papers published in the Top 5 economics journals (American Economic Review, Quarterly Journal of Economics, Journal of Political Economy, Econometrica, and Review of Economic Studies) are often distinguished from those in other journals by several key factors:

Depth of Contribution Originality and Innovation: Top 5 papers typically address questions of broad, foundational importance or propose groundbreaking methodologies. They often set new standards in the field or open new research avenues.

Generalizability: Findings are relevant to a wide range of settings, not just niche contexts. Big Questions: These papers tackle issues with substantial implications for policy, theory, or practice.

Methodological Rigor: High Standards of Empirical Methods: Empirical papers in the Top five journals employ state-of-the-art econometric techniques and robust identification strategies (e.g., natural experiments, randomized controlled trials, and structural modeling).

Theoretical Sophistication: Theoretical contributions are mathematically rigorous and provide deep insights, often with broad applicability.

Thorough Robustness Checks: Authors typically provide extensive sensitivity analyses to demonstrate the robustness of their results.

Writing and Presentation Quality Clarity and Structure: The narrative is compelling and accessible, even to non-specialists in the subfield, while maintaining academic precision.

Polished Presentation: Papers are meticulously written, with clear figures, tables, and appendices. The results are easy to interpret and visually intuitive.

Tight Argumentation: Papers avoid unnecessary digressions, focusing directly on the key question and results.

Data Quality Novelty of Data: Top 5 papers often leverage unique or hard-to-access datasets that enable the study of questions previously out of reach.

Rigorous Cleaning and Documentation: The data handling and analysis process is highly transparent, with all steps carefully documented.

Relevance and Impact Policy Relevance: Many Top 5 papers have clear implications for public policy or major economic debates, making their findings influential beyond academia.

Cross-Disciplinary Interest: These papers often resonate with researchers in related disciplines, such as political science, sociology, or psychology, enhancing their visibility and citation potential.

Citations: Papers in Top 5 journals often become highly cited due to their broad applicability and significance.

Extensive Peer Review and Revisions Stringent Referee Process: Top 5 journals have rigorous review processes, often involving multiple rounds of detailed feedback and revisions. High Rejection Rates: Acceptance rates are extremely low (e.g., ~5%), ensuring only the most impactful papers are published.

Network Effects and Prestige Author Reputation: Papers by well-known authors or prestigious institutions are more likely to receive attention and scrutiny during the review process.

Citations of Existing Literature: Top 5 papers typically build upon or challenge widely recognized works, further cementing their place in prominent scholarly conversations.

Comparison with Other Journals Scope and Niche: Non-Top 5 journals may focus on narrower questions or less generalizable findings, which, while still valuable, may not have the same broad impact.

Data Availability: Some journals may accept papers using less novel or standard datasets, provided the analysis is sound.

Methodological Simplicity: Papers in lower-ranked journals may employ standard or less sophisticated methodologies, especially in empirical studies.

Less Competitive Review Process: Non-Top 5 journals generally have higher acceptance rates and shorter review timelines, making them accessible to a broader range of researchers."

Funding competitiveness score

"As a reviewer for a major research funding organization, please evaluate whether this research proposal would be competitive for major funding (0 = definitely not fundable, 10 = definitely fundable at the highest award level)."

Top conference acceptance score

"As a program committee member for prestigious economics conferences, please evaluate whether this work would be accepted for presentation (0 = definitely reject, 10 = definitely accept for prominent session)."

Predicted citations
"Based on the novelty, methodology, and potential influence of this research, please project the actual number of citations this paper will receive in the next 10 years (output should be a specific predicted citation count)"

Research award score
"As a committee member for major research awards, please evaluate whether this work could be competitive for prestigious recognition (0 = definitely not award-worthy, 10 = definitely award-worthy)."

Nobel potential score
"As a member of the Nobel Prize Committee for Economic Sciences at the Royal Swedish Academy of Sciences, please provide a realistic evaluation of whether this research publication could contribute to winning the Nobel Prize in Economics (0 = Shows no indication of Nobel Prize potential,10 = Shows definitive Nobel Prize potential)",

Tenure case strength score
"As a senior member of a research university's tenure and promotion committee, please evaluate whether this research portfolio would support a strong case for tenure, considering both the quantity and quality of contributions (0 = definitely deny tenure, 10 = exceptionally strong case for tenure)."

Appendix D: GPT-4o mini API Call Structure
*prompt = the prompts in Appendix C
*full_content = the abstract and full text from the research paper of interest

```
response = client.chat.completions.create(
           model="gpt-4o-mini",
           messages=[
               {"role": "system", "content": prompt},
               {"role": "user", "content": full_content}
           ],
           response_format={
               "type": "json_schema",
               "json_schema": {
                   "name": "submission_evaluation",
                   "strict": True,
                   "schema": {
                       "type": "object",
                       "properties": {
                           "rating": {
                               "type": "number",
```

```
                                "description": f"Rating for {property_name} ({'0 or 1' if
is_published_check else '0-10'})"
                            }
                        },
                        "required": ["rating"],
                        "additionalProperties": False
                    }
                }
            },
            temperature=1,
            max_tokens=2024,
            top_p=1
        )
```