

How Well Do LLMs Perform in Peer Review?

Evidence from Finance and Accounting

Hai Lu* Yichen Luo† Luping Yu‡

October 14, 2026

Abstract

Referees increasingly use large language models (LLMs), while finance and accounting journals disagree about whether to allow it. We ask which part of a referee’s job the models can do. GPT-4o, whose training data end before the 2024 conference season, referees the papers presented at the 2024 AFA and NBER finance meetings, and we compare its verdicts with publication in the top-three finance journals over the next two years. Papers the model scores one standard deviation higher are 44% more likely to be published, net of author reputation. The result is not driven by look-ahead bias, as the model does not recall the outcomes. The information comes from its judgements of correctness and exposition. Its judgements of originality and significance carry none, and a referee guide that stresses importance does not help. The split between correctness and contribution holds for accounting papers and for a newer model, GPT-5. LLMs can check whether research is right, but not whether it matters. The technical check can go to the model, while the judgement of contribution stays with people.

JEL Classification: A11, G00, C45

Keywords: Peer review; large language models; research evaluation; AI policy

*University of Toronto and Peking University; email: hai.lu@rotman.utoronto.ca

†University College London; email: yichen.luo.22@ucl.ac.uk

‡University of Macau; email: lupingyu@um.edu.mo

1 Introduction

Generative AI has reached both sides of peer review. It lowers the cost of producing manuscripts (Novy-Marx and Velikov, 2026). At *Organization Science*, submissions rose 42% after the release of ChatGPT, and the editors attribute nearly all of the increase to heavily AI-written manuscripts (Gartenberg et al., 2026). In economics, Paul Goldsmith-Pinkham finds that the share of papers with at least one passage that a detector classifies as AI-written rises from almost zero in 2022 to roughly a quarter of NBER working papers and nearly half of arXiv economics papers, with the acceleration beginning in 2025.¹ Referees use the same tools, and they are stretched thin. At ICLR 2026, an AI detector classified about 21% of the reviews as fully AI-generated and found signs of AI use in more than half (Naddaf, 2025).² More than 30% of the referee reports at *Organization Science* show some AI use, and the editors warn that AI, combined with the incentive to publish, pushes science toward more rather than better research (Gartenberg et al., 2026). If AI multiplies the number of papers and already writes many of the reviews, can it also help tell which papers are better?

Computer science has started to answer this question, because some of its leading conferences publish their reviews. Comments written by GPT-4 overlap with those of human reviewers about as much as two human reviewers overlap with each other (Liang et al., 2024). The Stanford Agentic Reviewer, calibrated on reviews from ICLR 2025, produces scores that agree with a human reviewer about as well as two human reviewers agree with each other (Jiang and Ng, 2025). The conferences disagree on whether reviewers may use these models at all, from a ban at CVPR to deployments at AAAI-26 and NeurIPS 2026 (Table 1). At AAAI-26, authors and committee members rated the large language model (LLM) reviews above human reviews at identifying technical errors (Biswas et al., 2026).

Business research has no comparable evidence, although its journals are already setting rules, and the rules disagree within each field. In finance, the *Journal of Finance* lets referees use AI to draft or edit their reports and to look up references, provided they fact-check the output and remain accountable for it.³ Elsevier, which publishes the *Journal of Financial Economics*, bars reviewers from uploading any part of a manuscript to an AI tool and allows the tools only to improve the language of a report or to search the literature, with disclosure.⁴ In accounting, *The Accounting Review* lets reviewers use AI to improve the language of their

¹See <https://paulgp.substack.com/p/how-do-we-know-if-writing-is-ai>.

²The International Conference on Learning Representations (ICLR) is a leading machine learning conference. Evaluating its 2026 submissions took 18,054 reviewers and 76,139 reviews, about five papers per reviewer in two weeks (Naddaf, 2025).

³See <https://afajof.org/submissions/>.

⁴See <https://www.elsevier.com/about/policies-and-standards/generative-ai-policies-for-journals>.

report, with disclosure, but not to read the manuscript.⁵ *Contemporary Accounting Research* prohibits reviewers from using the tools at all.⁶ These policies have been continually updated to keep pace with advances in AI technology. An experiment in which an AI agent wrote referee reports on three of the author’s own finance papers in under two minutes each has prompted a call for guardrails on referees’ and editors’ use of these tools (Subrahmanyam, 2026). Referee reports at finance and accounting journals are confidential, so no one outside an editorial office can say whether the verdict of an LLM on a finance paper tracks the judgement of the field, or which part of its verdict does.

We build a proxy benchmark from the one part of the review process that is public: the decision. We take the 363 papers from the 2024 AFA Annual Meeting and the six NBER finance program meetings of 2024 whose PDFs the conferences hosted; the 280 outside the AFA PhD student poster session form the main sample. We ask GPT-4o, whose training data end three months before the AFA meeting, to referee the conference version. We call our prompt the naive prompt: it is the three sentences a referee who has done no prompt engineering might write, asking for a report, a recommendation, scores for originality, significance, correctness, and exposition, and an overall score. We then observe which papers the *Journal of Finance*, the *Journal of Financial Economics*, and the *Review of Financial Studies* published within two years of the meeting. The model’s scores are ex ante judgements, and the question is whether they carry information about the ex post decisions of editors and referees. Two further tests follow: To ask whether the result is specific to finance, we repeat the design on the 41 accounting papers of the 2023 and 2024 Conferences on Financial Economics and Accounting, a joint accounting and finance conference that hosts the full text of its papers. We use only its accounting papers, not those of its finance sessions. To ask whether a newer model can judge what an older one cannot, we repeat it with GPT-5.

Our key message is that an LLM referee can tell whether a finance paper is right and clearly written, but not whether it matters. Five findings support it.

First, the model’s verdict predicts publication. A one-standard-deviation rise in the model’s score makes a paper 44% more likely to appear in a top-three finance journal, or 2.9 percentage points from a base rate of 6.5%. The estimate is net of the authors’ school ranking, publication record, and editorial experience. Nor does the model remember the outcomes: when we ask whether each paper has been published, it says it does not know for 90% of them even when given the authors, and the score keeps its predictive power once its answers are controlled for.

Second, this information lies in the model’s judgements of *correctness*, whether the analysis,

⁵See <https://aaahq.org/Research/Journals/The-Accounting-Review/Guide-for-Authors>.

⁶See <https://www.caaa.ca/car-editorial-policies>.

empirical design, and interpretation are convincing, and of *exposition*, whether the paper is clearly written. Its originality and significance scores carry none, and for top-three publication its overall score predicts less well than its correctness or exposition score alone because it mixes in judgements that carry no information. This is the opposite of what the profession says about its own decisions. In the guide by [Berk et al. \(2017\)](#), most submitted papers are correct and the decision often turns on importance, so the model is informative about the margin that most submissions clear and uninformative about the margin on which the decision turns.

Third, handing the model a referee guide does not help. Given the article by [Berk et al. \(2017\)](#) to read before the paper, the model’s originality and significance scores still carry no information, the information again lies in its correctness score, and its overall score no longer predicts publication. Telling the model that importance is what matters does not make its verdict more informative.

Fourth, the pattern is not specific to finance. On the 41 accounting papers, the model’s overall and correctness scores predict publication in the top-three accounting journals, *The Accounting Review*, the *Journal of Accounting Research*, and the *Journal of Accounting and Economics*, within two years, its exposition score does so at the 10% level, and its originality and significance scores do not.

Fifth, the limit is not specific to an older model. GPT-5, released in August 2025, shows the same split: its correctness score predicts publication, and its originality and significance scores do not. Its later knowledge cutoff opens the possibility of look-ahead bias, so we repeat the look-ahead probe with GPT-5 and drop the two papers whose journal it names and that were submitted or online before its cutoff. The problem we document is therefore one that a newer model inherits rather than solves.

What can finance and accounting researchers take from these results? We organize the answer around three questions.

The first question is whether a model can replace a human referee. The answer is no, and the reason is more specific than a general distrust of the models. On importance, the margin that editors describe as decisive ([Berk et al., 2017](#)), the model’s judgement carries no information about the decision. Human referees find this margin hard as well: their recommendations on the same paper disagree widely ([Welch, 2014](#)), and the classic papers they once rejected show how poorly importance is judged at the time ([Gans and Shepherd, 1994](#)). But it is the margin the field cannot delegate. The limit is not in the model’s reading, since the same reading yields an informative judgement of correctness, and more knowledge does not close it, since a newer model that has read more of the literature shows the same limit. Whether a question matters

is a judgement call about what a field run by people will value, and the models do not yet make that call as editors do. Computer scientists reach the same conclusion: an LLM finds about half of the errors planted in short papers but errs more often than not when asked which of two abstracts is better (Liu and Shah, 2023), and at ICLR 2026 its scores predict acceptance less well than human scores do (Baumann et al., 2026).

The second question is whether a model can check a paper before a person reads it. The answer is yes, on the margins of correctness and exposition. A referee drafting a report can ask the model whether the design, the evidence, and the writing hold up, and treat the answer as a first reading that a person then verifies. An editor facing a growing volume of submissions can use the same judgement to decide where a referee’s time is best spent. An author can run the check before submitting. The division resembles the one found in other knowledge work, where AI helps on tasks inside its competence and hurts outside it (Dell’Acqua et al., 2026), and in security analysis, where a machine’s reading of hard data and an analyst’s reading of soft information do best combined (Cao et al., 2024). It also speaks to the field’s worry that fragile and irreproducible results pass human review (Harvey, 2017; Hail et al., 2020). On that part of the job a model is a cheap second reader that does not tire, and work on replication packages shows how much such a reader can catch (Schwartz et al., 2026).

The third question is what the results imply for the rules that journals and conferences are writing. The useful question is not whether to admit the model, on which the rules in Table 1 differ, but what to ask of it: the technical check can go to the model, and the judgement of contribution should stay with people. Section 4 develops this for the venues in the table.

We contribute to three literatures. The first studies peer review and publication in economics, finance, and accounting: what editors and referees reward (Laband, 1990; Ellison, 2002b; Card and DellaVigna, 2020; Berk et al., 2017), how much referees agree (Welch, 2014), how authors’ standing and connections and the circulation of a paper before submission shape outcomes (Laband and Piette, 1994; Colussi, 2018; Brogaard et al., 2014; Huber et al., 2022; Brown, 2005), and what predicts the publication of finance working papers (Dai et al., 2023; Hou et al., 2025). Peer review and publication are integral to the process of intellectual exploration, shaping not only which research findings enter the scholarly record but also how ideas are evaluated, refined, and advanced. In accounting, Moizer (2009) and Bloomfield et al. (2018) examine the review process itself, and the editors of *Accounting Horizons* have asked what role people should keep as AI enters academic research (Graham and Tucker, 2026). We add a measure of content produced by a model that reads the paper, and show which of its judgements align with editors’ decisions.

The second literature evaluates LLMs as referees. In computer science, reviews written by LLMs are compared with the public reviews of human referees (Liang et al., 2024; Jiang and Ng, 2025; Baumann et al., 2026), a randomized trial at ICLR 2025 has tested LLM feedback to referees (Thakkar et al., 2026), and AAAI-26 has reported on its pilot of LLM reviews (Biswas et al., 2026). In economics, Pataranutaporn et al. (2026) show that LLM ratings of published papers track the prestige of the journals that published them, and that some models rate papers higher when they are attributed to prominent authors and institutions. Li (2026) uses a model’s reading of working papers to separate the roles of merit and connections in journal placement. In finance and accounting, Newton and Riddiough (2026) show that a rubric-guided model scores AFA papers with high reliability, Novy-Marx and Velikov (2026) shows that the models can produce finance papers at scale, and Ciconte et al. (2026) that AI-drafted accounting papers are polished but not ready for a top journal. We relate the scores of a model given the naive prompt to editorial outcomes and trace the information to its correctness and exposition scores.

The third literature uses LLMs as tools in finance and accounting research, from forecasting returns and corporate policies (Lopez-Lira and Tang, 2026; Chen et al., 2026; Jha et al., 2024) to reading disclosures and measuring constructs in text (de Kok, 2025). It studies the effects of the models on information processing in markets (Bertomeu et al., 2025), with surveys by Eisfeldt and Schubert (2025) and Cao et al. (2026), and it has developed the tests for look-ahead bias that we adopt (Glasserman and Lin, 2024; Lopez-Lira et al., 2025; Sarkar and Vafa, 2024; Gao et al., 2025) and documented the bias in finance and accounting (Levy, 2026). Our results add a task, the evaluation of research, on which the division of labour between model and person can be measured.

Section 2 describes the data and design, Section 3 reports the results, Section 4 draws out the implications for the rules on AI in peer review, and Section 5 concludes. To keep the paper concise and increase its readability, we move the detailed discussions of the sample, the referee protocol, the tracking of publication outcomes, and additional tests to the Internet Appendix.

2 Data and Design

The design has three parts: a set of papers whose versions are fixed in time, an LLM referee that could not have seen what happened to them, and the editorial decisions that followed. It is the validation against realized outcomes, with an explicit test for leakage, that Ludwig et al. (2026) recommend before an LLM’s output is used in empirical work. The Internet Appendix gives the details, and Table IA.9 defines all variables.

2.1 Papers

The main sample consists of every paper on the programs of the 2024 AFA Annual Meeting (January 2024) and the six 2024 NBER finance program meetings for which the conference hosts the PDF: 363 papers after duplicates and non-paper files are removed. The Western Finance Association (WFA), European Finance Association (EFA), and China International Conference in Finance (CICF) do not post the full text of their 2024 papers. The main analysis uses the 280 papers outside the AFA PhD student poster session, a separate track for doctoral students. We use conference programs rather than a working-paper repository because a program timestamps a paper: the file on the program is the version that existed at the meeting, whereas a repository such as SSRN serves only the latest version (Dai et al., 2023). The model reads the full conference version, including tables, references, and appendices. Internet Appendix IA.1 describes the construction of the sample and of the text.

As a robustness check with evidence from another discipline, we apply the same design to the 41 accounting papers presented at the 2023 and 2024 Conferences on Financial Economics and Accounting (Internet Appendix IA.1), a joint accounting and finance conference that also hosts the full text of its papers. We use only its accounting papers, not those of its finance sessions. The accounting sample is small because it comes from a single conference: most accounting conferences, the high-quality Financial Accounting and Reporting Section (FARS) Midyear Meeting of the American Accounting Association among them, either do not post the PDFs of the papers they present or restrict them to participants, so the version presented at the meeting cannot be obtained.

2.2 The LLM Referee

The referee is `gpt-4o-2024-11-20`, whose stated knowledge cutoff is 1 October 2023, three months before the AFA meeting. Each paper is evaluated in a fresh session, with no other paper in context.

The prompt, which we call the naive prompt, imitates a referee who does not engage in prompt engineering. In three sentences it asks the model to review the paper as a referee for a finance journal, to write a report and recommend acceptance, revision, or rejection, and to score the paper from 0 to 10 on four dimensions and from 0 to 100 overall. The four dimensions, their one-sentence definitions, and the two scales follow Newton and Riddiough (2026):

- *Originality*: does the paper present a genuinely new idea, angle, dataset, design, or synthesis relative to established academic knowledge?
- *Significance*: how important is the question, and how large is the contribution if the results

are correct?

- *Correctness*: are the analysis, empirical design, theoretical logic, and interpretation scientifically convincing?
- *Exposition*: how clear, well organized, and well communicated is the paper?

The prompt contains no rubric, no example, and no instruction on how to weigh the dimensions (Internet Appendix [IA.2](#) reproduces it). For the accounting papers it is addressed to an accounting journal. Because a chat interface gives a different answer each time, we evaluate each paper four times in independent sessions and use the mean of the four draws, standardized across papers, so that a regression coefficient measures the effect of one standard deviation and does not depend on the units of the scale.

2.3 Publication Outcomes

The outcome is the editorial decision of the top-three journals of the field. $Top3Fin_i$ equals one if finance paper i is online at the *Journal of Finance*, the *Journal of Financial Economics*, or the *Review of Financial Studies* within two years of its meeting, the main outcome of [Dai et al. \(2023\)](#). $Top3Acc_i$ equals one if accounting paper i is online at *The Accounting Review*, the *Journal of Accounting Research*, or the *Journal of Accounting and Economics* within two years of its meeting. A fixed window gives every paper the same time to be accepted, and a paper published in any other journal has an outcome of zero. We match papers to journal articles by title and authors and check every uncertain match by hand, following [Hou et al. \(2025\)](#), so that papers whose titles changed before publication are found. The window had closed for every paper when we collected the outcomes. Of the 276 finance papers of the estimation sample, 18 (6.5%) are in a top-three finance journal; of the 41 accounting papers, 5 (12.2%) are in a top-three accounting journal. Internet Appendix [IA.3](#) describes the matching and reports the broader journal sets of [Dai et al. \(2023\)](#).

2.4 Specification

Following Equation (1) of [Dai et al. \(2023\)](#), we estimate

$$\Pr(Top3Fin_i = 1) = \Lambda(\alpha \cdot Score_i + \beta' \mathbf{X}_i + \mu_c), \quad (1)$$

where Λ is the logistic function, $Score_i$ is one of the standardized LLM scores, and μ_c distinguishes AFA from NBER papers. The controls $\mathbf{X}_i$ are the log length of the paper, the number of authors, and three measures of the authors' standing: whether any author is at one of the top-twenty finance schools, the prior JF, JFE, and RFS record of the most published author,

from OpenAlex (Internet Appendix [IA.4](#)), and $Editorial_i$, one if any author served as an editor or associate editor of a top-three journal of the field in the ten years up to the conference, 2014–2024 ([Brogaard et al., 2014](#)). For the accounting papers the outcome is $Top3Acc_i$, the record and editorial variables refer to the three accounting journals, the school indicator is the same as for the finance papers, and the fixed effect distinguishes the 2023 and 2024 meetings. We report robust z -statistics; the text reports the average marginal effect of the overall score. The estimation sample of the main specification has 276 papers: of the 280, we drop the paper none of whose authors could be matched to OpenAlex, the paper for which the model names the journal of an article submitted before its cutoff (Internet Appendix [IA.5](#)), and the two papers whose journal GPT-5 names and that were submitted or online before its cutoff ([Section 3.5](#)).

[Table 2](#) reports summary statistics. Two features stand out. The model assigns relatively high and uniform overall scores, averaging 81.9 for the finance papers (Panel A), with a standard deviation of only 2.4. In contrast, the dimension scores exhibit greater relative dispersion, as reflected in their higher coefficients of variation. For example, the coefficient of variation is 2.9% ($2.4/81.9$) for the overall score, compared with 4.3% ($0.36/8.40$) for the correctness score. The accounting papers (Panel B) exhibit similar patterns. And the authors are strong, as expected at these meetings: 56% of the finance papers have an author at a top-twenty school, 82% have an author with a prior top-three article, and 28% have an author who has served on the editorial board of a top-three journal.

3 Results

We proceed in seven steps. We first show that the model’s verdict predicts publication and locate the information in its judgements of correctness and exposition ([Sections 3.1](#) and [3.2](#)). We then show that a referee guide does not help ([Section 3.3](#)), that the pattern holds for accounting papers ([Section 3.4](#)), and that a stronger model reproduces it ([Section 3.5](#)). The last two steps ask where the information comes from: not from the authors’ standing or editorial connections ([Section 3.6](#)), and not from memory of the outcomes ([Section 3.7](#)).

3.1 The LLM’s Verdict Predicts Publication

Column (1) of [Table 3](#) relates the overall score to publication in a top-three finance journal. The coefficient is 0.495 ($z = 2.39$). Papers scored one standard deviation higher are 44% more likely to be published in the JF, JFE, or RFS within two years: 2.9 percentage points on a base rate of

6.5% in the estimation sample.⁷ The estimate is net of author reputation. Neither the authors' school, nor their prior top-three record, nor their editorial experience predicts publication in this sample. These papers have already passed a program committee's selection, which correlates with reputation. What remains to be explained is the editorial decision among papers of similar standing, and the model's reading of the paper carries information about that decision that the authors' names do not. Among the controls, only the number of authors predicts publication (Internet Appendix [IA.6](#)).

The result is not specific to the top-three journals. Adding the *Journal of Financial and Quantitative Analysis* and the *Review of Finance* gives a coefficient of 0.451 on the overall score ($z = 2.39$), and adding the second tier of finance journals gives 0.561 ($z = 2.92$) ([Table IA.6](#)).

3.2 Right and Clear, Not Important

Which judgement carries the information? Columns (2) to (5) of [Table 3](#) enter the four dimension scores one at a time. Correctness (0.691, $z = 2.00$) and exposition (0.667, $z = 2.68$) predict publication. Originality (-0.069 , $z = -0.33$) and significance (-0.141 , $z = -0.56$) carry no information: their coefficients are small, wrongly signed, and far from significant. The split is the same for the broader journal sets, where exposition again predicts on its own, correctness does so at the 10% level, and originality and significance again carry no information ([Table IA.6](#)).

The judgements of correctness and of contribution are separate in the model's scores. Across the finance papers, the correctness score is almost uncorrelated with the originality score (0.05) and the significance score (-0.08), and the same holds for the accounting papers (-0.13 and 0.01 ; [Table IA.8](#) in the Internet Appendix). A paper that the model rates as more correct is not rated as more original or more significant: the model makes two separate judgements, one of whether a paper is right and one of whether it matters. Its overall score is correlated with all four dimension scores (between 0.48 and 0.62 for the finance papers) and so combines the two.

For top-three publication, the model's overall score therefore dilutes its best signals. Correctness alone (pseudo- R^2 of 0.092) and exposition alone (0.096) explain more of the variation in publication than the overall score (0.072), which mixes these judgements with judgements of originality and significance that carry none.

The split runs against how the profession describes its own decisions. [Berk et al. \(2017\)](#) advise referees that most submitted papers are correct, so that the decision often turns on

⁷The 2.9 percentage points are the average marginal effect of the standardized overall score in column (1): for each of the 276 papers we compute the change in the predicted probability of publication per unit of the standardized score, with the controls at their observed values, and average over papers; because the score is standardized, a unit is one standard deviation. The delta-method standard error of the average marginal effect is 1.3 percentage points ($z = 2.16$). The base rate is the share of the 276 papers in a top-three journal, 18 papers or 6.5% ([Table IA.2](#)). The 44% is the ratio of the two, 2.90 over 6.52.

the breadth of the question and the size of the advance over the literature, and that assessing importance is the hardest part of the job. The model’s information lies in the dimensions that, by this account, most submissions already clear, and it has none in the dimensions on which the decision often turns. The pattern is consistent with what each judgement requires. Whether a design identifies what it claims, whether the evidence supports the interpretation, and whether the argument is clearly made can be checked against the paper itself. Whether a question matters, and whether the answer is new, call for a judgement about where the paper stands in a literature that keeps moving and about what the field values, and the model does not yet make this judgement as editors do. In the terms of [Ellison \(2002a\)](#), who separates the importance of a paper’s main ideas from its other qualities such as rigour and polish, the model judges the second and not the first.

3.3 A Referee Guide Does Not Help

A referee who uses a model can give it the advice that human referees follow, and a referee who wants the model to judge contribution might expect this to help. We test this on the same papers, keeping the model, the four draws, and the regression sample the same. The prompt places the full text of [Berk et al. \(2017\)](#), the advice summarized in [Section 3.2](#), before the naive prompt.

The guide does not help ([Table 4](#)). Compared with the naive prompt, it gives neither the originality nor the significance score any predictive power for top-three publication, and the information again lies in correctness (0.758, $z = 2.16$). Its overall score no longer predicts publication (0.203, $z = 0.80$), nor does its exposition score (0.007, $z = 0.03$).

The reading we favour is that the guide tells the model what to look for without giving it what it lacks: knowing that importance is what matters does not supply the judgement needed to assess importance.

3.4 Accounting Papers

Is the pattern specific to finance? Accounting provides a useful comparison because the two disciplines share confidential peer-review processes, selective publication standards, and a strong emphasis on research contributions. However, accounting research often requires specialized knowledge of reporting standards, auditing practices, and institutional settings. These differences may affect how LLMs assess research quality and the extent to which their evaluations align with editorial decisions. In addition, leading accounting journals have adopted particularly restrictive policies on referees’ use of LLMs ([Table 1](#)), potentially limiting their adoption in the

peer-review process. These considerations make accounting an interesting discipline for assessing the generalizability of our findings. We repeat the test of [Table 3](#) on the 41 accounting papers described in [Section 2](#), five of which are in a top-three accounting journal within two years. The model does not know these outcomes, so no accounting paper is dropped (columns (3)–(4) of [Table IA.4](#)).

[Table 5](#) reports the estimates. The overall score predicts publication (1.558, $z = 2.23$), and the information again lies in correctness (1.506, $z = 2.09$) and exposition (1.271, $z = 1.94$) rather than in originality (0.904, $z = 1.62$) or significance (-0.237 , $z = -0.37$). The sample is small and the coefficients are correspondingly large and imprecise, so we read the table for its pattern rather than its magnitudes. In accounting as in finance, the model’s verdict carries information about the editorial decision.

3.5 A Stronger Model

GPT-4o was released in 2024, and a natural question is whether a newer model closes the gap on originality and significance: a model that has read more of the literature might judge novelty and importance where its predecessor cannot. We repeat the design with GPT-5 (gpt-5-2025-08-07), which replaced GPT-4o as the default model of ChatGPT in August 2025, giving it the naive prompt. Its knowledge cutoff, 30 September 2024, falls after some of the publications we study, so we repeat with GPT-5 the look-ahead probe, which asks the model for the publication status of each paper ([Section 3.7](#)). Given the title and authors, it names the correct journal of four finance papers. Two of them are known to have been submitted or online before its cutoff, and we drop these two from the sample of [Table 3](#) and of this section.

GPT-5 grades more strictly than GPT-4o: it gives every paper a lower correctness score, by 1.74 points on average for the finance papers, and on average lower originality and significance scores, while its overall score for the finance papers is unchanged ([Table IA.7](#) in the Internet Appendix). Because the scores are standardized, a stricter scale does not change the estimates.

GPT-5 reproduces the pattern ([Table 6](#)). Its information again lies in correctness: the correctness score predicts top-three publication (0.677, $z = 2.31$), with a pseudo- R^2 of 0.086 against 0.080 for the overall score. Its overall score predicts publication at the 10% level (0.806, $z = 1.88$), where GPT-4o’s does at the 5% level on the same papers (0.495 in [Table 3](#)). The exposition result is also robust to GPT-5 (0.490, $z = 1.71$). Neither originality (-0.148 , $z = -0.66$) nor significance (0.389, $z = 1.33$) predicts publication.

The stronger model therefore does not change the answer. It is about as informative as its predecessor about whether a paper is right, and no more informative about whether it

matters, although it has read more of the literature. Knowing more does not by itself supply the judgement that the decision on contribution requires.

3.6 Author and Editor Information

The conference PDF shows who wrote the paper, and the model reads the names, affiliations, and acknowledgements. Author standing predicts publication in the universe of finance working papers (Dai et al., 2023), and a connection to an editor raises the chance of publication at the editor’s journal (Brogaard et al., 2014). If the model’s score reflected these facts rather than the content of the paper, our conclusion would change: the model would be a reader of title pages, not of papers.

One result already speaks against this. The authors’ school, prior top-three record, and editorial experience are controlled for in every specification, and none of them predicts publication in the conference sample (Section 3.1), so there is little for the score to proxy for. For comparison, human referees respond strongly to standing: in a field experiment with a finance manuscript, the share of reviewers recommending rejection fell from 65% when only an unknown author was shown to 23% when only a Nobel laureate was shown (Huber et al., 2022). Interacting the score with indicators for an author with a prior top-three article, an editor among the authors, or an all-early-stage team, following Table 11 of Dai et al. (2023), gives no interaction that is significant at the 10% level (Table IA.3 in the Internet Appendix).

Editorial connections are the sharper test, because an editor’s name on a title page is a fact the model can read and a signal that an editor’s decision might follow (Laband and Piette, 1994; Brogaard et al., 2014). $Editorial_i$ (Section 2) equals one for 28% of the finance papers and 24% of the accounting papers (Table 2), and it is insignificant in every column of Tables 3 to 6, including those for the referee guide, the accounting papers, and GPT-5. On the evidence of the author characteristics we measure, the score’s information is not explained by who wrote the paper.

3.7 Not Memory

Another alternative is that the model remembers the outcomes (look-ahead bias). Every publication in the sample appeared after the model’s cutoff; the earliest went online on 28 October 2023. A signal could still have leaked through an author’s CV listing a revise-and-resubmit. We therefore ask the model, in a fresh session, for the publication status of every paper. Given the title, it answers that it does not know for 96% of the papers; given the title and authors, for 90% (Table IA.4). It never reports a revise-and-resubmit, and its occasional claims that a

paper is published are guesses, made more often for papers that remain unpublished than for those that are published. Adding an indicator for a claimed publication and its interaction with the score (Gao et al., 2025) leaves the coefficient on the score at 0.491 and 0.466 ($z = 2.29$ and 2.18) for the two information sets (Internet Appendix IA.5). For the accounting papers the model answers that it does not know for every paper given the title and authors, and for all but one given the title alone (columns (3) and (4) of Table IA.4).

4 What the Results Imply for the Rules on AI in Peer Review

Journals and conferences are writing rules for referees’ use of LLMs without evidence on what the models can judge. Table 1 collects the rules in force on 30 September 2026 at the leading computer science conferences, at the top finance and accounting journals and their publishers, and at two leading economics journals. A survey of 802 journals finds that most high-impact journals now have some rule of this kind (Wang and Gong, 2026). Two patterns stand out in the table.

First, the fields disagree about *whether*. In computer science, CVPR 2025, NeurIPS 2025, and the ACL Rolling Review ban or restrict reviewers’ use of the models, ICML 2026 lets authors choose the policy for their paper, ICLR 2026 permits use with disclosure, and AAAI-26 and NeurIPS 2026 supply LLM output to their own review process. In finance, accounting, and economics, the rules run from permission under conditions of accountability, confidentiality, or disclosure at the *Journal of Finance*, the *Review of Financial Studies*, the *Review of Finance*, the *American Economic Review*, and *Econometrica*, through bans on uploading a manuscript to an AI tool at the *Journal of Financial Economics*, the *Journal of Accounting Research*, and *The Accounting Review*, or to a public tool at the *Review of Accounting Studies* and *Management Science*, to an outright ban at *Contemporary Accounting Research*.

Second, where the rules permit, the permitted use is clerical. A referee may draft, edit, organize, and look up references; the assessment stays with the referee. The *Review of Finance* says the models must not substitute for the referee’s own evaluation, and *Econometrica* that the referee must form an independent evaluation. *Management Science* says that all conclusions must be the reviewer’s own, and *The Accounting Review* states that the critical assessment required for peer review is outside the scope of generative AI. The premise behind the permissive and the restrictive rules is the same: the model’s verdict on a paper is not to be trusted. The rules differ only in whether the drafting help is worth the risk to confidentiality.

Our evidence refines that premise, and the refinement matters for how the rules should be written. The critical assessment of a paper has two parts. For the first part, whether the analysis

is sound, the model’s judgements of whether the analysis is sound and whether the paper is clearly written carry information about the editorial decision for GPT-4o, and the results are robust to the accounting papers and to GPT-5. For the second part, whether the question matters and the answer is new, the model’s judgement carries no information, and asking it to weigh that part, as the referee guide does, does not help. A rule that treats the model’s verdict as one thing therefore makes one of two mistakes. A ban discards a check that is informative on the margin of correctness. A deployment that passes on the model’s overall score, whether to triage submissions or to seed a review, admits a judgement that mixes an informative part with an uninformative one, and for top-three publication in finance the mix predicts worse than the correctness score alone (Section 3.2). Our findings suggest an important distinction between the tasks that LLMs can usefully support, such as assessing technical correctness and clarity, and those that require greater human judgement, such as evaluating a paper’s intellectual contribution.

Some venues already separate the two tasks, which is itself evidence that the division is workable. AAAI-26’s LLM reviews carried no scores or recommendations, and the authors and committee members who read them rated them above human reviews at identifying technical errors, the margin on which our evidence says the model is informative (Biswas et al., 2026). *Econometrica* runs its conditionally accepted papers through an AI tool whose stated purpose is to identify technical and expository issues, not to assess the contribution of a paper, while requiring referees to form their own evaluation.⁸ Bhargava et al. (2026) propose that journals place the models in a structured, contestable front-line review under human oversight rather than leave referees to use them privately. Gans (2026) shows that how an editor shares an AI assessment with referees determines whether it adds information or duplicates the model’s blind spots. Our results say which assessment is worth sharing.

Two consequences follow for the venues in Table 1. For a referee working under a rule like that of the *Journal of Finance*, the useful request is a technical check of the design, the evidence, and the writing, and the guide experiment shows that asking the model to apply the field’s standard of contribution does not help. And the upload bans bind: the check requires the model to read the full paper, which a referee at the *Journal of Financial Economics*, the *Journal of Accounting Research*, *The Accounting Review*, or *Management Science* may not provide to a public tool. Under these rules the check can be run by the journal itself, under its own confidentiality terms, as the computer science conferences do, or by a referee through a tool that guarantees confidentiality, as the *Review of Financial Studies* rule allows. It can also be

⁸See <https://www.econometricsociety.org/publications/econometrica/editorial-procedures-and-policies> and <https://www.econometricsociety.org/publications/econometrica/guidelines-referees>.

run by the authors before submission, where no confidentiality question arises. Authors are the one party that no rule restricts, and the technical check is the one they can use. Bans on referees are in any case hard to enforce: in a randomized comparison at ICML 2026, a ban and a limited-use policy had near-zero effects on review scores and decisions, and 22.5% of the reviewers under the ban reported using a model anyway (Kim et al., 2026).

Two cautions apply. The first is that the anticipated use of LLM evaluations may induce authors to optimize their manuscripts for model-generated scores rather than for research quality. Firms revise their disclosures once they know that machines read them (Cao et al., 2023), and authors have already hidden instructions to AI reviewers in the text of preprints (Gibney, 2025). A journal that runs the check must therefore treat its output as a reading to be verified rather than as a verdict, and keep the prompt and the model under its own control. The second is that our evidence has limits that the rules should respect. It comes from a naive prompt and two models, it compares the model’s verdict with the decision rather than with the reports that editors read, and it concerns papers that had already passed a conference’s selection. It does not show that a model can replace the technical part of a referee report, only that its judgement on that part is informative. A journal that runs the check will want to measure, on its own submissions and reports, how much of the information survives. Only the journal can do this, because it alone holds the reports.

5 Conclusion

Which part of a referee’s job can a large language model do? We answer with the one public record of refereeing in finance: the publication decision. Given three sentences of instruction, GPT-4o’s verdict on the papers of the 2024 AFA and NBER meetings predicts which of them the top-three finance journals published within two years, net of the authors’ reputation and with no memory of the outcomes. A one-standard-deviation rise in its score raises the chance of publication by 44%. The prediction rests on two of the four judgements we ask for, correctness and exposition. The other two, originality and significance, add nothing, and a referee guide that tells the model to weigh importance does not help. The accounting papers and GPT-5 give the same split between correctness and contribution.

The results answer three questions that finance and accounting researchers are asking. Can a model replace a referee? No: under the naive prompt its judgements of originality and significance carry no information on the margin of contribution, which is the margin on which editors say their decisions turn. Can it assist human reviewers? Yes: its judgements of whether a paper is right and whether it is clearly written carry information for GPT-4o, and the results

are robust to the accounting papers and to GPT-5. This is useful to referees drafting a report, to editors facing a growing volume of submissions, and to authors before they submit. What should the rules say? The technical check can go to the model, and the judgement of contribution should stay with people. Judgements of contribution remain inherently contextual and are not readily captured by model-generated scores, so at this stage our evidence suggests using LLMs as complementary tools in peer review while preserving human judgement as central to evaluating contributions.

Unlike some leading computer science conferences, finance and accounting journals keep their referee reports confidential. An LLM referee for our fields can therefore be neither calibrated on the judgements of our referees nor evaluated against them, and we can compare its verdicts only with publication decisions, a coarser signal that arrives years later. A release of anonymized referee reports and decisions by finance and accounting journals would allow researchers to compare a model’s reports with those that editors read, to calibrate referees for our fields, and to measure which parts of human judgement the models can and cannot yet reproduce. A future step is to move from the naive prompt studied here to the rubric-guided design of [Newton and Riddiough \(2026\)](#), and to measure whether the input of human judgement makes LLM scores of originality and significance predict better.

References

- Bajo, E., M. Barbi, and D. Hillier (2020). Where should I publish to get promoted? A finance journal ranking based on business school promotions. *Journal of Banking & Finance* 114, 105780.
- Baumann, J., J. Pei, S. Koyejo, and D. Hovy (2026). Stop Automating Peer Review Without Rigorous Evaluation. In *Proceedings of the 43rd International Conference on Machine Learning, Position Paper Track*. arXiv:2605.03202.
- Berk, J. B., C. R. Harvey, and D. Hirshleifer (2017). How to Write an Effective Referee Report and Improve the Scientific Review Process. *Journal of Economic Perspectives* 31(1), 231–244.
- Bertomeu, J., Y. Lin, Y. Liu, and Z. Ni (2025). The Impact of Generative AI on Information Processing: Evidence from the Ban of ChatGPT in Italy. *Journal of Accounting and Economics* 80(1), 101782.
- Bhargava, H. K., S. H. Bana, Z. Zhang, L. Brandimarte, V. Choudhary, J. F. Li, P. Loupos, and D. Zantedeschi (2026). Fighting Fire with Fire: Infusing Artificial Intelligence into Peer Review to Sustain Quality Scholarship. *Management Science* 72(9), 7381–7393.
- Bini, R., L. W. Cong, Y. Huang, and X. Jin (2026). Large Language Models and Financial Decision Making. NBER Working Paper.
- Biswas, J., S. Schoepp, G. Vasan, A. Opiari, A. Zhang, Z. Hu, S. Joseph, M. Lease, J. J. Li, P. Stone, K. L. Wagstaff, M. E. Taylor, and O. C. Jenkins (2026). AI-Assisted Peer Review at Scale: The AAAI-26 AI Review Pilot. Working paper, arXiv:2604.13940, 15 April 2026.
- Bloomfield, R., K. Rennekamp, and B. Steenhoven (2018). No System Is Perfect: Understanding How Registration-Based Editorial Processes Affect Reproducibility and Investment in Research Quality. *Journal of Accounting Research* 56(2), 313–362.
- Brogaard, J., J. Engelberg, and C. A. Parsons (2014). Networks and Productivity: Causal Evidence from Editor Rotations. *Journal of Financial Economics* 111(1), 251–270.
- Brown, L. D. (2005). The Importance of Circulating and Presenting Manuscripts: Evidence from the Accounting Literature. *The Accounting Review* 80(1), 55–83.
- Cao, S., W. Jiang, J. Wang, and B. Yang (2024). From Man vs. Machine to Man + Machine: The Art and AI of Stock Analyses. *Journal of Financial Economics* 160, 103910.

- Cao, S., W. Jiang, B. Yang, and A. L. Zhang (2023). How to Talk When a Machine Is Listening: Corporate Disclosure in the Age of AI. *Review of Financial Studies* 36(9), 3603–3642.
- Cao, S. S., W. X. Chen, G. Ma, and S. Srinivasan (2026). Generative AI in Capital Markets: Information Production, Dissemination, and Processing. *Journal of Accounting Research* 64(3), 1427–1450.
- Card, D. and S. DellaVigna (2020). What Do Editors Maximize? Evidence from Four Economics Journals. *Review of Economics and Statistics* 102(1), 195–217.
- Chen, Y., B. T. Kelly, and D. Xiu (2026). Expected Returns and Large Language Models. Working paper, SSRN 4416687, February 2026 version.
- Ciconte, W. A., A. M. Rozario, and O. Urcan (2026). Using AI to Identify Exogenous Shocks and Conduct Archival Accounting Research. *Review of Accounting Studies* 31(3), 1638–1676.
- Colussi, T. (2018). Social Ties in Academia: A Friend Is a Treasure. *Review of Economics and Statistics* 100(1), 45–50.
- Currie, R. R. and G. S. Pandher (2020). Finance journal rankings: Active scholar assessment revisited. *Journal of Banking & Finance* 111, 105717.
- Dai, R., L. Donohue, Q. F. S. Drechsler, and W. Jiang (2023). Dissemination, Publication, and Impact of Finance Research: When Novelty Meets Conventionality. *Review of Finance* 27(1), 79–141.
- de Kok, T. (2025). ChatGPT for Textual Analysis? How to Use Generative LLMs in Accounting Research. *Management Science* 71(9), 7888–7906.
- Dell’Acqua, F., E. McFowland, E. Mollick, H. Lifshitz-Assaf, K. C. Kellogg, S. Rajendran, L. Kraymer, F. Candelon, and K. R. Lakhani (2026). Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of Artificial Intelligence on Knowledge Worker Productivity and Quality. *Organization Science* 37(2), 403–423.
- Eisfeldt, A. L. and G. Schubert (2025). Generative AI and Finance. *Annual Review of Financial Economics* 17(1), 363–393.
- Ellison, G. (2002a). Evolving Standards for Academic Publishing: A q-r Theory. *Journal of Political Economy* 110(5), 994–1034.
- Ellison, G. (2002b). The Slowdown of the Economics Publishing Process. *Journal of Political Economy* 110(5), 947–993.

- Gans, J. S. (2026). Designing AI-Augmented Peer Review. NBER Working Paper 35688, August 2026.
- Gans, J. S. and G. B. Shepherd (1994). How Are the Mighty Fallen: Rejected Classic Articles by Leading Economists. *Journal of Economic Perspectives* 8(1), 165–179.
- Gao, Z., W. Jiang, and Y. Yan (2025). Detecting Lookahead Bias in LLM Forecasts. Working paper, arXiv:2512.23847.
- Gartenberg, C., S. Hasan, A. Murray, and L. Pierce (2026). More Versus Better: Artificial Intelligence, Incentives, and the Emerging Crisis in Peer Review. *Organization Science* 37(3), 795–812.
- Gibney, E. (2025). Scientists Hide Messages in Papers to Game AI Peer Review. *Nature* 643(8073), 887–888.
- Glasserman, P. and C. Lin (2024). Assessing Look-Ahead Bias in Stock Return Predictions Generated by GPT Sentiment Analysis. *Journal of Financial Data Science* 6(1), 25–42.
- Graham, G. M. and J. W. Tucker (2026). The Use of Artificial Intelligence in Academic Research. *Accounting Horizons* 40(1), 1–10.
- Hail, L., M. Lang, and C. Leuz (2020). Reproducibility in Accounting Research: Views of the Research Community. *Journal of Accounting Research* 58(2), 519–543.
- Hansen, S., M. McMahon, and A. Prat (2018). Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach. *Quarterly Journal of Economics* 133(2), 801–870.
- Harvey, C. R. (2017). Presidential Address: The Scientific Outlook in Financial Economics. *Journal of Finance* 72(4), 1399–1440.
- Hou, W., E. Smajlbegovic, and D. Urban (2025). Ranking finance conferences: An update. *Journal of Empirical Finance* 84, 101652.
- Huber, J., S. Inoua, R. Kerschbamer, C. König-Kersting, S. Palan, and V. L. Smith (2022). Nobel and Novice: Author Prominence Affects Peer Review. *Proceedings of the National Academy of Sciences* 119(41), e2205779119.
- Jha, M., J. Qian, M. Weber, and B. Yang (2024). ChatGPT and Corporate Policies. NBER Working Paper 32161, February 2024.

- Jiang, Y. and A. Ng (2025). Stanford Agentic Reviewer. <https://paperreview.ai/>. Stanford ML Group. Technical overview at <https://paperreview.ai/tech-overview>.
- Kim, S. S. Y., W. H. Deng, J. W. Vaughan, B. Su, W. Su, A. Agarwal, S. Li, M. Jaggi, D. G. Goldstein, N. B. Shah, and M. Dudík (2026). Use and Effects of LLMs in Peer Review: A Randomized Experiment and Survey at ICML 2026. Working paper, arXiv:2609.19420, September 2026.
- Laband, D. N. (1990). Is There Value-Added from the Review Process in Economics? Preliminary Evidence from Authors. *Quarterly Journal of Economics* 105(2), 341–352.
- Laband, D. N. and M. J. Piette (1994). Favoritism versus Search for Good Papers: Empirical Evidence Regarding the Behavior of Journal Editors. *Journal of Political Economy* 102(1), 194–203.
- Levy, B. (2026). Caution Ahead: Numerical Reasoning and Look-Ahead Bias in AI Models. *Journal of Accounting Research* 64(3), 1139–1188.
- Li, N. (2026). Merit or Networks? What Decides Where Research Is Published. Working paper, arXiv:2606.03763, June 2026.
- Liang, W., Y. Zhang, H. Cao, B. Wang, D. Y. Ding, X. Yang, K. Vodrahalli, S. He, D. S. Smith, Y. Yin, D. A. McFarland, and J. Zou (2024). Can Large Language Models Provide Useful Feedback on Research Papers? A Large-Scale Empirical Analysis. *NEJM AI* 1(8), AIoa2400196.
- Liu, R. and N. B. Shah (2023). ReviewerGPT? An Exploratory Study on Using Large Language Models for Paper Reviewing. Working paper, arXiv:2306.00622.
- Lopez-Lira, A. and Y. Tang (2026). Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models. *Journal of Financial Economics* 184, 104335.
- Lopez-Lira, A., Y. Tang, and M. Zhu (2025). The Memorization Problem: Can We Trust LLMs’ Economic Forecasts? Working paper, SSRN 5217505.
- Ludwig, J., S. Mullainathan, and A. Rambachan (2026). Large Language Models: An Applied Econometric Framework. *Annual Review of Economics* 18(1), 283–316.
- Moizer, P. (2009). Publishing in Accounting Journals: A Fair Game? *Accounting, Organizations and Society* 34(2), 285–304.

- Naddaf, M. (2025). Major AI Conference Flooded with Peer Reviews Written Fully by AI. *Nature* 648, 256–257.
- Newton, J. C. S. and S. J. Riddiough (2026). AI Referees in Finance. Working paper, SSRN 6446183, 18 March 2026.
- Novy-Marx, R. and M. Velikov (2026). Artificial Intelligence–Powered (Finance) Scholarship. *Journal of Economic Literature* 64(1), 5–37.
- Ouyang, S., H. Yun, and X. Zheng (2025). How Do Large Language Models Trade? Evidence from Simulated Markets. Working paper, arXiv.
- Pataranutaporn, P., N. Powdthavee, C. Achiwaranguprok, and P. Maes (2026). Large Language Models and Human Judgement in Scientific Peer Review: Evidence from over 1000 Published Economics Papers. *Economica*. Forthcoming.
- Priem, J., H. Piwowar, and R. Orr (2022). OpenAlex: A Fully-Open Index of Scholarly Works, Authors, Venues, Institutions, and Concepts. arXiv:2205.01833.
- Sarkar, S. K. and K. Vafa (2024). Lookahead Bias in Pretrained Language Models. Working paper, SSRN 4754678, October 2024 version.
- Schwartz, M., I. Andrews, and J. M. Shapiro (2026). An LLM Workflow That Reproduces, Improves, and Extends Published Economics Research. NBER Working Paper 35782, September 2026.
- Subrahmanyam, A. (2026). The Scholarly Review Process in Finance: Experiments in Artificial Intelligence. Working paper, SSRN 6451699, March 2026.
- Thakkar, N., M. Yuksekgonul, J. Silberg, A. Garg, N. Peng, F. Sha, R. Yu, C. Vondrick, and J. Zou (2026). A Large-Scale Randomized Study of Large Language Model Feedback in Peer Review. *Nature Machine Intelligence* 8(3), 326–336.
- Wang, Z. and M. Gong (2026). A Cross-Disciplinary Analysis of AI Policies in Academic Peer Review. *Learned Publishing* 39(1), e2035.
- Welch, I. (2014). Referee Recommendations. *Review of Financial Studies* 27(9), 2773–2804.

Table 1: Rules for Large Language Models in Peer Review

This table summarizes the rules for reviewers’ use of large language models at leading computer science conferences (Panel A), at finance and accounting journals and their publishers (Panel B), and at economics and management journals (Panel C), from the venues’ own policy pages accessed 30 September 2026. Year is the conference year or, for journals, the year of the policy page. A publisher named in parentheses means that the journal states no rule of its own and the publisher’s rule applies. “Bans” and “Restricts” mark rules that forbid or limit the use of LLMs by reviewers; “Permits” marks rules that allow it under conditions; “Choice” marks a rule under which authors choose the policy for their paper and reviewers are assigned one; “Deploys” marks venues that themselves supply LLM output to their review process.

Venue	Year	Rule for reviewers	Stance
<i>Panel A: Computer science</i>			
CVPR	2025	LLMs may not be used to write reviews or meta-reviews at any step; a violating reviewer’s own papers may be desk rejected	Bans
NeurIPS	2025	Submissions may not be shared with any LLM; LLMs may help a reviewer understand concepts and improve the writing	Restricts
ACL Rolling Review	2026	Generative assistance may not produce the first draft of a review; paraphrasing is allowed, uploads to tools that do not preserve privacy are not	Restricts
ICML	2026	Authors state whether their paper requires a no-LLM policy or allows a permissive one, and each reviewer is assigned one policy; the permissive policy lets an LLM explain the paper and polish the review but not judge it	Choice
ICLR	2026	Reviewers may use LLMs with disclosure and are responsible for the output; reviews with false claims or hallucinated references breach the code of ethics, and the reviewer’s own papers may be desk rejected; all reviews were screened with two LLM detectors and the flags sent to area chairs	Permits
AAAI	2026	The conference adds LLM-generated reviews to the initial review stage alongside human reviews, without numerical scores or ratings and not for automated accept or reject decisions; no human reviewer is replaced	Deploys
NeurIPS	2026	A voluntary randomized experiment gives reviewers an LLM assistant in OpenReview that helps them analyse a submission but does not write the review; outside the experiment no LLM use is sanctioned	Deploys
<i>Panel B: Finance and accounting</i>			
<i>Journal of Finance</i>	2026	AI may help draft or edit referee reports and look up prior references; referees must fact-check its output	Permits
<i>Rev. Fin. Studies</i>	2026	Manuscripts may not be fed into any tool that does not guarantee strict confidentiality, including an opt-out from model training; the referee remains responsible for the report	Permits
<i>Review of Finance</i>	2026	LLMs may be used in a supportive way but not as a substitute for the referee’s own evaluation, interpretation, or recommendations	Permits
<i>J. Financial Economics</i> (Elsevier)	2026	Reviewers should not upload a manuscript or any part of it into an AI tool; private tools may support language and literature search, with disclosure	Restricts
<i>Accounting Review</i>	2026	Reviewers should not upload any part of a manuscript into an AI tool; AI may improve the language of the report, with disclosure; the critical assessment required for peer review is outside the scope of generative AI	Restricts
<i>J. Accounting Research</i> (Wiley)	2026	Manuscripts may not be uploaded to AI tools, in full or in part; AI drafting assistance must be disclosed	Restricts
<i>Rev. Acc. Studies</i> (Springer Nature)	2026	Manuscript content may not be uploaded to public AI tools; generating a review, producing a recommendation, or delegating critique to AI is not permitted	Restricts

Continued on next page

Table 1 continued

Venue	Year	Rule for reviewers	Stance
<i>Contemp. Acc. Research</i>	2026	Reviewers are prohibited from using generative AI tools in the review process, including to improve the language of a completed report	Bans
<i>Panel C: Economics and management</i>			
<i>Amer. Economic Review</i>	2026	Any AI use must be disclosed in the reviewer’s cover letter and limited to tools that guarantee the manuscript will not be used for training	Permits
<i>Econometrica</i>	2026	Referees are expected to read the submission themselves and form an independent evaluation; AI may improve the clarity, organization, and language of the report; conditionally accepted papers are checked with an AI tool for technical and expository issues, not for contribution	Permits
<i>Management Science</i>	2026	No part of a manuscript may be uploaded to a generative AI tool that could compromise confidentiality; all conclusions must be the reviewer’s own	Restricts

Sources, in table order (all pages accessed 30 September 2026):

CVPR, 2025 Reviewer Guidelines, <https://cvpr.thecvf.com/Conferences/2025/ReviewerGuidelines>
 NeurIPS, 2025 Policy on the Use of Large Language Models, <https://neurips.cc/Conferences/2025/LLM>
 ACL Rolling Review, Reviewer Guidelines, <https://aclrollingreview.org/reviewerguidelines>
 ICML, 2026 Policy for LLM Use in Reviewing, <https://icml.cc/Conferences/2026/LLM-Policy>
 ICLR, ICLR 2026 Response to LLM-Generated Papers and Reviews, blog post of 19 November 2025, <https://blog.iclr.cc/2025/11/19/iclr-2026-response-to-llm-generated-papers-and-reviews/>
 ICLR, A Retrospective on the ICLR 2026 Review Process, blog post of 31 March 2026, <https://blog.iclr.cc/2026/03/31/a-retrospective-on-the-iclr-2026-review-process/>
 AAAI, AAAI Launches AI-Powered Peer Review Assessment System, 16 May 2025, <https://aaai.org/aaai-launches-ai-powered-peer-review-assessment-system/>
 NeurIPS, 2026 AI-Assisted Reviewing Experiment, <https://neurips.cc/Conferences/2026/ai-reviewing-experiment>
 American Finance Association, Journal of Finance submissions page, <https://afajof.org/submissions/>
 Society for Financial Studies, Review of Financial Studies: Submit a Report, <https://sfs.org/review-of-financial-studies/submit-a-report/>
 Review of Finance, Policy Update Regarding the Use of Large Language Models, 23 March 2026, <https://revfin.org/policy-update-regarding-the-use-of-large-language-models-llms/>
 Elsevier, Generative AI Policies for Journals, updated June 2026, <https://www.elsevier.com/about/policies-and-standards/generative-ai-policies-for-journals>
 American Accounting Association, The Accounting Review: Guide for Authors, <https://aaahq.org/Research/Journals/The-Accounting-Review/Guide-for-Authors>
 Wiley, Using AI Tools in Your Research, <https://www.wiley.com/en-us/publish/article/ai-guidelines/>
 Springer Nature, Editorial Policies: Using AI in Peer Review, <https://www.springernature.com/gp/policies/editorial-policies/using-ai-in-peer-review>
 Canadian Academic Accounting Association, Contemporary Accounting Research Editorial Policies, <https://www.caaa.ca/car-editorial-policies>
 American Economic Association, American Economic Review Editorial Policy, <https://www.aeaweb.org/journals/aer/editorial-policy>
 Econometric Society, Econometrica Reviewer Guidelines, <https://www.econometricsociety.org/publications/econometrica/guidelines-referees>
 Econometric Society, Econometrica Editorial Procedures and Policies, <https://www.econometricsociety.org/publications/econometrica/editorial-procedures-and-policies>
 INFORMS, Management Science Submission Guidelines, <https://pubsonline.informs.org/page/mnsc/submission-guidelines>

Table 2: Summary Statistics

This table presents summary statistics for the 276 papers presented at the 2024 AFA Annual Meeting and the six 2024 NBER finance meetings outside the AFA PhD student poster session that enter the regressions (Panel A; [Section 2.4](#)) and for the 41 accounting papers presented at the 2023 and 2024 Conferences on Financial Economics and Accounting (Panel B; Internet Appendix [IA.1](#)). Scores are the means over four independent draws of GPT-4o (`gpt-4o-2024-11-20`) under the naive prompt reproduced in Internet Appendix [IA.2](#), on a 0–100 scale for the overall score and a 0–10 scale for the four dimensions. Publication outcomes are measured within two years of each paper’s meeting. In Panel B, *Log Prior Top_i* counts the authors’ articles in the three accounting journals of *Top3Acc_i* before 2024, *Editorial_i* refers to the same journals, and *Top20School_i* uses the same list of twenty finance schools as Panel A. All variables are defined in [Table IA.9](#).

<i>Panel A: Finance</i>						
	N	Mean	SD	P10	P50	P90
<i>Overall_i</i>	276	81.924	2.402	78.75	82.25	85.00
<i>Originality_i</i>	276	8.308	0.310	8.00	8.25	8.75
<i>Significance_i</i>	276	8.118	0.370	7.50	8.00	8.50
<i>Correctness_i</i>	276	8.397	0.363	8.00	8.50	8.75
<i>Exposition_i</i>	276	7.207	0.410	6.75	7.25	7.75
<i>Top3Fin_i</i>	276	0.065	0.247	0.00	0.00	0.00
<i>Top5Fin_i</i>	276	0.072	0.260	0.00	0.00	0.00
<i>Top10Fin_i</i>	276	0.080	0.271	0.00	0.00	0.00
<i>Log Length_i</i>	276	10.396	0.351	9.96	10.42	10.82
<i>N Authors_i</i>	276	2.931	0.964	2.00	3.00	4.00
<i>NBER_i</i>	276	0.167	0.373	0.00	0.00	1.00
<i>Top20School_i</i>	276	0.558	0.498	0.00	1.00	1.00
<i>Established_i</i>	276	0.822	0.383	0.00	1.00	1.00
<i>Log Prior Top_i</i>	276	1.730	1.102	0.00	1.95	3.04
<i>Early Stage_i</i>	276	0.109	0.312	0.00	0.00	1.00
<i>Editorial_i</i>	276	0.275	0.448	0.00	0.00	1.00

<i>Panel B: Accounting</i>						
	N	Mean	SD	P10	P50	P90
<i>Overall_i</i>	41	82.537	2.252	79.50	82.75	85.00
<i>Originality_i</i>	41	8.183	0.262	8.00	8.25	8.50
<i>Significance_i</i>	41	8.061	0.402	7.50	8.00	8.50
<i>Correctness_i</i>	41	8.506	0.401	8.00	8.50	9.00
<i>Exposition_i</i>	41	7.232	0.360	6.75	7.25	7.75
<i>Top3Acc_i</i>	41	0.122	0.331	0.00	0.00	1.00
<i>Log Length_i</i>	41	10.334	0.212	10.10	10.31	10.69
<i>N Authors_i</i>	41	2.732	1.096	1.00	3.00	4.00
<i>Top20School_i</i>	41	0.341	0.480	0.00	0.00	1.00
<i>Log Prior Top_i</i>	41	1.631	0.972	0.00	1.79	2.64
<i>Editorial_i</i>	41	0.244	0.435	0.00	0.00	1.00

Table 3: The LLM Referee’s Scores and Publication in the Top Three Finance Journals

This table presents logit estimates of equation (1). The sample is the 276 papers presented at the 2024 AFA Annual Meeting and the six 2024 NBER finance meetings outside the AFA PhD student poster session that enter the regressions: of the 280 papers outside the poster session, we drop the paper none of whose authors could be matched to OpenAlex, one paper for which GPT-4o names the journal of an article submitted before its knowledge cutoff (Internet Appendix IA.5), and two papers whose journal GPT-5 names and that were submitted or online before its cutoff (Section 3.5). The outcome is $Top3Fin_i$, one if the paper is online at the *Journal of Finance*, the *Journal of Financial Economics*, or the *Review of Financial Studies* within two years of its meeting. Scores are the means over four independent draws of GPT-4o (gpt-4o-2024-11-20) under the naive prompt reproduced in Internet Appendix IA.2, standardized across papers; each column enters one score, and a coefficient is the change in log odds per standard deviation of the score. All specifications include the log length of the paper text, the number of authors, $Top20School_i$, $Log Prior Top_i$, $Editorial_i$, and a conference (AFA versus NBER) fixed effect. All variables are defined in Table IA.9. Robust z-statistics in parentheses. ***, **, and * denote significance at the 1%, 5%, and 10% levels.

	$Top3Fin_i$				
	(1)	(2)	(3)	(4)	(5)
$Overall_i$	0.495** (2.39)				
$Originality_i$		-0.069 (-0.33)			
$Significance_i$			-0.141 (-0.56)		
$Correctness_i$				0.691** (2.00)	
$Exposition_i$					0.667*** (2.68)
$Log Length_i$	0.017 (0.03)	0.061 (0.11)	0.116 (0.20)	-0.050 (-0.08)	-0.328 (-0.58)
$N Authors_i$	0.597** (2.45)	0.567** (2.31)	0.573** (2.32)	0.622** (2.48)	0.623** (2.53)
$Top20School_i$	0.171 (0.34)	0.261 (0.51)	0.257 (0.53)	0.231 (0.48)	0.133 (0.27)
$Log Prior Top_i$	0.108 (0.39)	0.086 (0.31)	0.097 (0.35)	0.130 (0.50)	0.113 (0.37)
$Editorial_i$	-0.063 (-0.11)	0.034 (0.06)	0.039 (0.07)	-0.159 (-0.28)	-0.026 (-0.05)
Conference FE	Y	Y	Y	Y	Y
N Obs	276	276	276	276	276
Pseudo R ²	0.072	0.048	0.050	0.092	0.096

Table 4: The Referee Guide and Publication in the Top Three Finance Journals

This table presents logit estimates of equation (1) with scores obtained under a prompt that places the full text of the refereeing advice of Berk et al. (2017) before the naive prompt (Section 3.3; both prompts are reproduced in Internet Appendix IA.2). The sample is the 276 papers presented at the 2024 AFA Annual Meeting and the six 2024 NBER finance meetings outside the AFA PhD student poster session that enter the regressions (Section 2.4). The outcome is $Top3Fin_i$, one if the paper is online at the *Journal of Finance*, the *Journal of Financial Economics*, or the *Review of Financial Studies* within two years of its meeting. Scores are the means over four independent draws of GPT-4o (gpt-4o-2024-11-20) under this guided prompt, standardized across papers; each column enters one score, and a coefficient is the change in log odds per standard deviation of the score. All specifications include the log length of the paper text, the number of authors, $Top20School_i$, $Log Prior Top_i$, $Editorial_i$, and a conference (AFA versus NBER) fixed effect. All variables are defined in Table IA.9. Robust z-statistics in parentheses. ***, **, and * denote significance at the 1%, 5%, and 10% levels.

	Referee guide attached				
	(1)	(2)	(3)	(4)	(5)
$Overall_i$	0.203 (0.80)				
$Originality_i$		-0.189 (-0.76)			
$Significance_i$			-0.228 (-0.97)		
$Correctness_i$				0.758** (2.16)	
$Exposition_i$					0.007 (0.03)
$Log Length_i$	0.030 (0.05)	0.103 (0.18)	0.041 (0.08)	-0.007 (-0.01)	0.058 (0.10)
$N Authors_i$	0.562** (2.35)	0.579** (2.30)	0.562** (2.27)	0.638*** (2.81)	0.566** (2.31)
$Top20School_i$	0.221 (0.45)	0.260 (0.52)	0.239 (0.49)	0.189 (0.39)	0.241 (0.49)
$Log Prior Top_i$	0.083 (0.29)	0.096 (0.35)	0.115 (0.42)	0.068 (0.23)	0.091 (0.32)
$Editorial_i$	0.056 (0.10)	0.013 (0.02)	-0.024 (-0.04)	-0.014 (-0.02)	0.035 (0.06)
Conference FE	Y	Y	Y	Y	Y
N Obs	276	276	276	276	276
Pseudo R ²	0.051	0.052	0.053	0.105	0.048

Table 5: Accounting Papers: The LLM Referee’s Scores and Publication in the Top Three Accounting Journals

This table presents logit estimates of equation (1) for the 41 accounting papers presented at the 2023 and 2024 Conferences on Financial Economics and Accounting (Internet Appendix IA.1). The outcome is $Top3Acc_i$, one if the paper is online at *The Accounting Review*, the *Journal of Accounting Research*, or the *Journal of Accounting and Economics* within two years of its meeting. Scores are the means over four independent draws of GPT-4o (gpt-4o-2024-11-20) under the naive prompt reproduced in Internet Appendix IA.2, addressed to an accounting journal, standardized across papers; each column enters one score, and a coefficient is the change in log odds per standard deviation of the score. All specifications include the log length of the paper text, the number of authors, $Top20School_i$ (the same list of twenty finance schools as for the finance papers), $Log Prior Top_i$, which counts the authors’ articles published before 2024 in the three accounting journals, $Editorial_i$ for the same journals, and a fixed effect for the meeting (2023 versus 2024). All variables are defined in Table IA.9. Robust z-statistics in parentheses. ***, **, and * denote significance at the 1%, 5%, and 10% levels.

	$Top3Acc_i$				
	(1)	(2)	(3)	(4)	(5)
$Overall_i$	1.558** (2.23)				
$Originality_i$		0.904 (1.62)			
$Significance_i$			-0.237 (-0.37)		
$Correctness_i$				1.506** (2.09)	
$Exposition_i$					1.271* (1.94)
$Log Length_i$	-2.189 (-0.80)	0.955 (0.42)	0.773 (0.35)	-2.256 (-0.91)	1.035 (0.36)
$N Authors_i$	-0.864 (-1.36)	-0.947 (-1.18)	-0.915* (-1.79)	-0.814 (-1.38)	-1.051 (-1.58)
$Top20School_i$	1.637 (1.17)	1.505 (1.26)	1.674 (1.34)	1.070 (0.80)	1.693 (1.34)
$Log Prior Top_i$	-0.837 (-0.63)	-0.143 (-0.15)	-0.011 (-0.01)	-0.447 (-0.47)	0.015 (0.01)
$Editorial_i$	-0.191 (-0.07)	-0.667 (-0.40)	-0.328 (-0.18)	-0.356 (-0.16)	-0.664 (-0.41)
Conference FE	Y	Y	Y	Y	Y
N Obs	41	41	41	41	41
Pseudo R ²	0.280	0.207	0.163	0.248	0.301

Table 6: GPT-5's Scores and Publication in the Top Three Finance Journals

This table presents logit estimates of equation (1) with the scores of GPT-5 (Section 3.5). The sample is the 276 papers presented at the 2024 AFA Annual Meeting and the six 2024 NBER finance meetings outside the AFA PhD student poster session that enter the regressions (Section 2.4). The outcome is $Top3Fin_i$, one if the paper is online at the *Journal of Finance*, the *Journal of Financial Economics*, or the *Review of Financial Studies* within two years of its meeting. Scores are the means over four independent draws of GPT-5 (gpt-5-2025-08-07) under the naive prompt reproduced in Internet Appendix IA.2, standardized across papers; each column enters one score, and a coefficient is the change in log odds per standard deviation of the score. All specifications include the log length of the paper text, the number of authors, $Top20School_i$, $Log Prior Top_i$, $Editorial_i$, and a conference (AFA versus NBER) fixed effect. All variables are defined in Table IA.9. Robust z -statistics in parentheses. ***, **, and * denote significance at the 1%, 5%, and 10% levels.

	$Top3Fin_i$				
	(1)	(2)	(3)	(4)	(5)
$Overall_i$	0.806* (1.88)				
$Originality_i$		-0.148 (-0.66)			
$Significance_i$			0.389 (1.33)		
$Correctness_i$				0.677** (2.31)	
$Exposition_i$					0.490* (1.71)
$Log Length_i$	-0.803 (-1.11)	0.256 (0.42)	-0.296 (-0.46)	-0.671 (-0.94)	-0.263 (-0.44)
$N Authors_i$	0.604** (2.39)	0.556** (2.31)	0.581** (2.36)	0.611** (2.46)	0.522** (2.12)
$Top20School_i$	-0.019 (-0.03)	0.270 (0.53)	0.111 (0.21)	0.023 (0.04)	0.103 (0.20)
$Log Prior Top_i$	0.136 (0.46)	0.066 (0.23)	0.115 (0.41)	0.139 (0.47)	0.096 (0.33)
$Editorial_i$	-0.125 (-0.21)	0.093 (0.16)	-0.075 (-0.12)	-0.171 (-0.28)	-0.008 (-0.01)
Conference FE	Y	Y	Y	Y	Y
N Obs	276	276	276	276	276
Pseudo R^2	0.080	0.050	0.060	0.086	0.067

Internet Appendix

for “How Well Do LLMs Perform in Peer Review? Evidence from Finance and Accounting”

IA.1 Sample and Text

The finance sample consists of the papers presented at the 2024 AFA Annual Meeting (San Antonio, 5–7 January 2024) and at the six 2024 NBER finance meetings: the Asset Pricing and Corporate Finance program meetings in spring and fall, and the two Summer Institute meetings. We do not substitute repository versions for the conferences whose programs do not host the full text ([Section 2](#)).

We crawl the two programs and download every PDF the conference hosts. The AFA program lists 334 papers (243 in regular sessions and 91 in the PhD poster session) and hosts a PDF for 333 of them; the NBER programs list 54 papers and host 45. A paper presented at both conferences is kept once, giving 372 unique papers with a conference-hosted PDF. We drop 9 files that are not papers (poster images or two-page abstracts of fewer than five pages), leaving 363 unique papers. [Table IA.1](#) summarizes the construction. Papers are identified by a hash of the normalized title; when a paper appears both in a regular session and in the poster session we keep the regular-session file.

Following [Dai et al. \(2023\)](#), we convert each PDF, finance and accounting, to text by extracting its text layer, and we use optical character recognition (OCR) only when a file has no text layer, which is the case for one paper. Following [Hansen et al. \(2018\)](#), we remove running page headers, page footers, and page numbers; we retain footnotes, tables, references, and appendices because a referee reads them and because the longest paper (82,456 tokens) fits within the model’s 128,000-token context window. The removal is restricted to the first and last three lines of each page and to lines that recur on at least 40% of pages, so that numeric lines in the body, such as equation numbers and table cells, are never touched; the first occurrence of a recurring line is kept so that repeated table headers survive. The cleaned texts have a median length of 32,559 tokens and retain 99.8–99.9% of the characters of the raw text layer.

Table IA.1: Sample Construction

This table reports the number of papers on the programs of the 2024 AFA Annual Meeting and the six 2024 NBER finance meetings (the Asset Pricing and Corporate Finance program meetings in spring and fall and the two Summer Institute meetings), the number for which the conference hosts the submitted PDF, and the number of usable files. A file is usable if it has at least five pages; the excluded files are poster images and two-page abstracts. Papers presented at more than one meeting are counted once in the last row; a paper appearing both in a regular session and in the poster session is assigned to the regular session.

	Papers on program	Conference- hosted PDF	Usable (≥ 5 pages)
AFA 2024 Annual Meeting, regular sessions	243	242	240
AFA 2024 Annual Meeting, PhD poster session	91	91	83
NBER 2024 finance meetings (six meetings)	54	45	45
Unique papers	383	372	363

Accounting papers. The Conference on Financial Economics and Accounting hosts the full text of the papers on its program for its 2023 edition (Rutgers Business School, 3–4 November 2023) and its 2024 edition (University at Buffalo, 24–26 October 2024). We take the 24 papers in the accounting sessions of 2024 and the 21 papers in the accounting and joint sessions of 2023; four of the 2023 files are one-page abstracts, which leaves 41 papers. Their text is extracted and cleaned in the same way, and the prompt is the naive prompt with “an accounting journal” in place of “a finance journal”.

IA.2 The Referee Protocol and Prompts

The knowledge cutoff of `gpt-4o-2024-11-20` (Section 2.2) is also six months before the earliest NBER meeting in the sample; Section IA.5 tests whether the model knows the outcomes. Scoring each paper in a fresh session makes the scores absolute judgements rather than within-session rankings (Newton and Riddiough, 2026). The prompt asks for the three-way recommendation that finance journals ask referees to make (Berk et al., 2017). The response is constrained through the structured-output feature of the application programming interface (API) to a fixed schema with the fields `report` (string), `recommendation` (one of `Accept`, `Revise and resubmit`, `Reject`), `originality`, `correctness`, `significance`, `exposition` (integers in $[0, 10]$), and `overall` (integer in $[0, 100]$); the schema fixes the shape of the answer and is not visible to the model as an instruction.

The sampling temperature is left at the model’s default, because a referee who uses a chat interface does not set it. Each paper is scored four times, because repeated evaluation at the default temperature is the standard way to measure the reliability of LLM judgements (Newton and Riddiough, 2026; Ouyang et al., 2025; Bini et al., 2026). All 1,452 requests were submitted

through the batch API; 1,451 returned a parseable response.⁹

The prompt below is sent as a single user message, followed by the full text of the paper. The definitions of the four dimensions and the two scales are taken from Section 2.2 of [Newton and Riddiough \(2026\)](#); the statements there are rephrased as questions.

Please review this paper as a referee for a finance journal. Write a referee report and tell me whether it should be accepted, revised, or rejected. Also rate the paper from 0 to 10 on each of the following, and give an overall score from 0 to 100:

- Originality: does it present a genuinely new idea, angle, dataset, design, or synthesis relative to established academic knowledge?
- Correctness: are the analysis, empirical design, theoretical logic, and interpretation scientifically convincing?
- Significance: how important is the question, and how large is the contribution if the results are correct?
- Exposition: how clear, well organized, and well communicated is the paper?

{paper_text}

The guidance prompt of [Section 3.3](#) changes only the text before the rating instructions. It places the following sentence and the full text of [Berk et al. \(2017\)](#) before the naive prompt:

Before reviewing the paper below, please read the following article on how to referee research papers in finance and economics.

{guide_text}

The look-ahead probe of [Section IA.5](#) is sent in a fresh session at temperature 0 with the same model. The response is constrained to the fields **status** (one of **Published**, **Accepted**, **forthcoming**, **Revise and resubmit**, **Submitted**, **under review**, **Not submitted or not published**, **Do not know**), **journal**, **year**, and **confidence** (integer in $[0, 100]$). The line in brackets is included only in columns (2) and (4) of [Table IA.4](#).

What is the publication status of the following paper? Tell me whether it has been published in an academic journal, accepted for publication, given a revise-and-resubmit, or submitted to a journal, and if so which journal and in which year. If you do not know, say so.

Title: {title}

[Authors: {authors}]

⁹In one response the model placed an incomplete report in the refusal field of the API; that paper’s scores are the mean of three draws.

IA.3 Publication Outcomes

A calendar cutoff would give papers presented in November 2024 ten months less than papers presented in January, which is why we measure publication within a fixed window of two years after each paper’s meeting (Section 2).

We follow the two-stage procedure of Hou et al. (2025), who track the publication status of 6,805 finance conference papers. In the automated stage we search for each paper’s title in Crossref, OpenAlex, Semantic Scholar, and Google Scholar and match a candidate journal article when the normalized-title similarity is at least 0.85 and at least one author surname is shared; matches with title similarity between 0.60 and 0.85 are flagged for manual review. Records whose venue is a preprint server, a working-paper series, or a web domain are never treated as journal articles, and the date of a matched article is taken from Crossref because aggregators that merge a working paper with its published version sometimes report the working paper’s date. In the manual stage, which Hou et al. (2025) describe as necessary because “changes in paper titles” are “the most frequent obstacle when tracking papers,” we search Crossref by the authors’ surnames for every paper still unmatched, compare the abstracts of candidate articles whose author lists coincide with the program entry, and read the acknowledgement footnotes of the published versions. This stage recovers nine published papers whose titles changed before publication, eight of which state their conference title in a footnote and eight of which thank the 2024 AFA meeting. Forty-two grey-zone cases were decided by hand. Hou et al. (2025) report that title-based matching alone misses about two thirds of eventual publications over a 6–10 year horizon; over our shorter horizon it misses 11%.

The outcome variables follow the finance journal groups of Dai et al. (2023); Bajo et al. (2020) obtain the same top group from promotion data and Currie and Pandher (2020) the same second tier from scholar assessments. Besides $Top3Fin_i$, their main outcome, $Top5Fin_i$ adds the *Journal of Financial and Quantitative Analysis* and the *Review of Finance*, and $Top10Fin_i$ further adds the *Journal of Corporate Finance*, the *Journal of Banking & Finance*, *Financial Management*, the *Journal of Financial Markets*, and the *Journal of Financial Intermediation* (Dai et al., 2023, Internet Appendix). Because the sample is drawn from finance conferences, we use only these finance outcomes. Table IA.2 reports the distribution for the 276 papers of Table 3: within two years of the meeting, 18 (6.5%) are in a top-three finance journal, 20 (7.2%) in a top-five finance journal, and 22 (8.0%) in a top-ten finance journal. Among all 363 papers, including the poster session, the counts are 21 (5.8%), 24 (6.6%), and 27 (7.4%). The window is short by the standards of Hou et al. (2025), who find that many conference papers take more than five years to be accepted, so the unpublished group contains papers that will eventually

be published, which biases our estimates toward zero.

Table IA.2: Publication Outcomes

This table reports the publication status, two years after their meeting, of the 276 papers presented at the 2024 AFA Annual Meeting and the six 2024 NBER finance meetings outside the AFA PhD student poster session that enter the regressions (Panel A) and of the 41 accounting papers presented at the 2023 and 2024 Conferences on Financial Economics and Accounting (Panel B). The outcomes were collected on 18 September 2026 for the finance papers and on 5 October 2026 for the accounting papers. A paper is counted as published in a journal when a matching article is online within two years of its meeting; matching follows the two-stage procedure of [Hou et al. \(2025\)](#) described in Section [IA.3](#). The finance outcomes follow [Dai et al. \(2023\)](#): the top three finance journals (the *Journal of Finance*, the *Journal of Financial Economics*, and the *Review of Financial Studies*), the top five (adding the *Journal of Financial and Quantitative Analysis* and the *Review of Finance*), and the top ten (adding the *Journal of Corporate Finance*, the *Journal of Banking & Finance*, *Financial Management*, the *Journal of Financial Markets*, and the *Journal of Financial Intermediation*). The top three accounting journals are *The Accounting Review*, the *Journal of Accounting Research*, and the *Journal of Accounting and Economics*; three accounting papers are published in journals outside these three.

<i>Panel A: Finance</i>		
	Papers	Share (%)
$Top3Fin_i = 1$ (JF, JFE, RFS)	18	6.5
$Top5Fin_i = 1$ (top three plus JFQA, RoF)	20	7.2
$Top10Fin_i = 1$ (top five plus JCF, JBF, FM, JFM, JFI)	22	8.0
Journal outside the top ten finance journals	5	1.8
Not published	249	90.2
Papers	276	100.0

<i>Panel B: Accounting</i>		
	Papers	Share (%)
$Top3Acc_i = 1$ (TAR, JAR, JAE)	5	12.2
Journal outside the top three accounting journals	3	7.3
Not published	33	80.5
Papers	41	100.0

IA.4 Author Reputation

This section describes the reputation measures that enter the controls of equation (1) and the interaction specification of [Table IA.3](#), which follows Table 11 of [Dai et al. \(2023\)](#) and asks whether the score is more informative for some author groups than for others.

[Dai et al. \(2023\)](#) measure author standing by a top-twenty school indicator and a coauthor-network centrality computed over the whole SSRN finance network. We keep their school indicator: $Top20School_i$ equals one if any author’s affiliation on the conference program is one of the twenty schools of the UT Dallas research ranking computed on publications in JF, JFE, and RFS, using the worldwide ranking for 2019–2023, the last five-year window that ends before the 2024 meetings. The centrality is not reproducible on a conference sample, and we replace it with the publication record of the authors, in the spirit of the prolific-author measure of [Card](#)

and DellaVigna (2020), and with their editorial experience ($Editorial_i$). For each author we obtain the OpenAlex profile (Priem et al., 2022), taken from the authorship list of the paper’s own OpenAlex record where one exists (75% of authors) and otherwise from an author search by name disambiguated by the program affiliation; 3% of authors, mostly PhD students and authors with very common names, cannot be matched. $Log\ Prior\ Top_i$ is the log of one plus the largest number of JF, JFE, or RFS articles published before 2024 by any author of the paper. It is computed on records dated before 2024, so it is the reputation an editor would have observed at the time of the conference. We restrict works to the economics and business fields because OpenAlex merges some scholars with namesakes in other fields; a profile with no economics or business publication is treated as unmatched.

The sort variables of the interaction specification follow Table 11 of Dai et al. (2023), adapted to our data: their sorts are the school indicator, an indicator for a top-5% value of their centrality measure, and an early-stage indicator. Ours are $Established_i$, one if any author has a prior JF, JFE, or RFS article, the indicator form of $Log\ Prior\ Top_i$ and the prolific-author margin of Card and DellaVigna (2020); $Editorial_i$ (Section 2), which marks the authors closest to the editorial process (Brogaard et al., 2014); and $Early\ Stage_i$, one if every matched author’s first publication in OpenAlex is within the ten years before the conference, the “early stage researcher” definition of Dai et al. (2023). A sort variable enters the specification once, so when it is itself a control it replaces it.

Results. Table IA.3 reports the interactions. Of the 276 papers of Table 3, 11% are written entirely by early-stage researchers; the shares with an established author and with an editor among the authors are in Table 2.

Column (1) sorts on $Established_i$. For papers whose authors have no prior top-three article the score has a coefficient of 0.201 ($z = 1.13$), and the interaction is positive (0.337, $z = 1.11$), an implied coefficient of about 0.54 for papers with an established author; the difference is not significant. Column (2) sorts on $Editorial_i$. For papers without an editor or associate editor among the authors the score has a coefficient of 0.693 ($z = 2.97$), and the interaction is -0.599 ($z = -1.34$), an implied coefficient of about 0.09 for papers by editors. Column (3) sorts on $Early\ Stage_i$. The score is a strong predictor for teams with at least one experienced author (0.516, $z = 2.33$) and weaker for all-early-stage teams; the interaction is -0.350 ($z = -1.13$). With 30 early-stage papers the estimate is imprecise, but it points in the direction of the model being least informative where a referee’s judgement would be most valuable, for authors without a track record. The same specification on the 236 papers whose authors’ OpenAlex profiles all come from the paper’s own OpenAlex record, excluding the 40 papers with at least one profile

matched by a name search, gives a coefficient of 0.467 ($z = 2.28$) on the score for $Top3Fin_i$ (untabulated).

Table IA.3: The Predictability of Publication by Author Standing

This table presents logit estimates of equation (1) in which the standardized overall score is interacted with a sort variable, following Table 11 of Dai et al. (2023). Column (1) sorts on $Established_i$, one if any author has an article published before 2024 in the *Journal of Finance*, the *Journal of Financial Economics*, or the *Review of Financial Studies*; column (2) on $Editorial_i$, one if any author served as an editor or associate editor of these journals in 2014–2024; and column (3) on $Early Stage_i$, one if every matched author’s first publication is within the ten years before the conference (Section IA.4). The sample is the 276 papers presented at the 2024 AFA Annual Meeting and the six 2024 NBER finance meetings outside the AFA PhD student poster session that enter the regressions (Section 2.4). The outcome is $Top3Fin_i$, one if the paper is online at the *Journal of Finance*, the *Journal of Financial Economics*, or the *Review of Financial Studies* within two years of its meeting. The overall score is the mean over four independent draws of GPT-4o (gpt-4o-2024-11-20) under the naive prompt reproduced in Section IA.2, standardized across papers. The coefficient on $Overall_i$ is the effect for papers with the sort variable equal to zero; the interaction is the difference for papers with the sort variable equal to one. All specifications include the log length of the paper text, the number of authors, $Top20School_i$, $Log Prior Top_i$, $Editorial_i$ (which in column (2) enters only as the sort variable), and a conference (AFA versus NBER) fixed effect. All variables are defined in Table IA.9. Robust z -statistics in parentheses. ***, **, and * denote significance at the 1%, 5%, and 10% levels.

Sort variable	$Top3Fin_i$		
	(1) $Established_i$	(2) $Editorial_i$	(3) $Early Stage_i$
$Overall_i$	0.201 (1.13)	0.693*** (2.97)	0.516** (2.33)
$Sort Variable_i$	1.062 (0.76)	0.133 (0.24)	0.015 (0.01)
$Overall_i \times Sort Variable_i$	0.337 (1.11)	-0.599 (-1.34)	-0.350 (-1.13)
$Log Length_i$	0.033 (0.06)	0.070 (0.12)	0.025 (0.04)
$N Authors_i$	0.573** (2.09)	0.601** (2.37)	0.594** (2.31)
$Top20School_i$	0.153 (0.31)	0.152 (0.30)	0.169 (0.33)
$Log Prior Top_i$	-0.107 (-0.26)	0.112 (0.41)	0.100 (0.29)
$Editorial_i$	0.005 (0.01)		-0.054 (-0.09)
Conference FE	Y	Y	Y
N Obs	276	276	276
Pseudo R ²	0.080	0.079	0.072

IA.5 Look-Ahead Bias

Because every publication outcome in the sample occurs after the model’s knowledge cutoff (Section 3.7), the outcome itself cannot be in the training data, the design that Lopez-Lira and Tang (2026) use to rule out look-ahead bias in return prediction; Sarkar and Vafa (2024)

and Levy (2026) document how much such bias can distort LLM-based results in finance and accounting. A signal of the outcome can be. An author’s website or CV may list a paper as a revise-and-resubmit or as forthcoming at a journal before the cutoff, and a model that has read it could recognize the paper and score it accordingly.

We test for this channel by asking the model directly, as Lopez-Lira et al. (2025) do for economic data the model may have memorized. In a fresh session we ask the scoring model for the publication status of each paper: whether it is published, accepted, under revise-and-resubmit, or submitted, at which journal, and in which year, with the option to answer that it does not know (Section IA.2). We vary the identifying information in the spirit of Glasserman and Lin (2024), who vary the firm identifiers in news text: we give the model the conference title alone, and then the title and the authors, the pair through which a CV entry would link the paper to a journal. We set the temperature to 0, under which a memorized answer is most likely to surface. We ask about the unpublished papers as well as the published ones, because the unpublished papers show how often the model claims a publication that did not happen. For each information set we define $Claim_i$ as one if the model answers that the paper is published or accepted. Following Gao et al. (2025), we add $Claim_i$ and its interaction with the score to (1), one information set at a time. If the score predicts publication because the model knows the outcome, its coefficient should fall once $Claim_i$ is controlled for, and the interaction should be positive.

Section 3.7 reports the model’s main answers (columns (1)–(2) of Table IA.4). In addition, with the title and authors the model calls 10.1% of the papers published. These claims are guesses. The model makes them for 10.4% of the papers that remain unpublished at the end of the outcome window and for 7.4% of those that are published, and the claimed years are 2023 or earlier, whereas the earliest publication in the sample appears in late October 2023.

Of the 27 papers published in a journal within the outcome window, the model names the correct journal for one given the title and for another given the title and authors. We read the submission and acceptance dates from the published articles: both were submitted to the journal after the cutoff, so no editorial decision on them could have been public. Asked about every conference paper, the model also names the journal of one paper that was submitted in May 2023 and accepted in February 2024, so a revise-and-resubmit could have been public; we exclude it from every regression. For the 41 accounting papers (columns (3)–(4) of Table IA.4), the model names the journal of none of the eight papers published within the window, so no accounting paper is excluded.

The model’s claims carry no information about publication (Table IA.5). With the claim

and its interaction added, the coefficient on the overall score for $Top3Fin_i$ stays close to the 0.495 of Table 3 for both information sets (Section 3.7). Neither the claim nor the interaction is significant in any column; for the title and authors the interaction is 0.291 ($z = 0.69$). For $Top5Fin_i$ and $Top10Fin_i$ the coefficient on the overall score is between 0.435 and 0.578 and significant at the 5% level in every column. On the 248 papers the model does not call published given the title and authors, the coefficient on the overall score is 0.438 ($z = 2.24$) for $Top3Fin_i$ and 0.522 ($z = 2.74$) for $Top10Fin_i$ (untabulated). The score is somewhat higher for papers the model calls published (0.41 standard deviations, $t = 2.03$, untabulated), but this component does not predict which papers are published.

Table IA.4: What Does the Model Know About Publication Outcomes?

This table reports what GPT-4o (gpt-4o-2024-11-20), the model whose scores the paper studies, answers when asked in a fresh session at temperature 0 for the publication status of each paper (Section IA.5); the question is reproduced in Section IA.2. The model receives the conference title (columns 1 and 3) or the title and authors (columns 2 and 4), and gives one answer per paper. Columns (1)–(2) report the 276 papers presented at the 2024 AFA Annual Meeting and the six 2024 NBER finance meetings outside the AFA PhD student poster session that enter the regressions, and columns (3)–(4) the 41 accounting papers presented at the 2023 and 2024 Conferences on Financial Economics and Accounting. The top rows report the distribution of answers. The bottom rows count, among the papers published in a journal within two years of their meeting, the papers for which the model names the correct journal, and among these the papers submitted to that journal, or online, before the model’s knowledge cutoff of 1 October 2023, as dated in the published article. A repeat of the finance title-and-authors probe with identical requests gives the same answer for 97.5% of the papers.

	Finance		Accounting	
	Title (1)	Title and authors (2)	Title (3)	Title and authors (4)
<i>Answer (share of the papers)</i>				
Papers	276	276	41	41
Do not know	0.964	0.899	0.976	1.000
Published or accepted	0.036	0.101	0.024	0.000
Revise and resubmit or under review	0.000	0.000	0.000	0.000
<i>Papers published in a journal within the outcome window</i>				
Papers	27	27	8	8
Journal named correctly	1	1	0	0
of which submitted or online before the cutoff	0	0	0	0

Table IA.5: LLM Scores and Publication, Controlling for What the Model Claims to Know

This table presents logit estimates of equation (1) with $Claim_i$ and its interaction with the overall score added (Gao et al., 2025). $Claim_i$ equals one if GPT-4o, asked in a fresh session at temperature 0 for the publication status of the paper, answers that it is published or accepted when given the conference title (odd columns) or the title and authors (even columns), as in Table IA.4. Columns (1)–(2) report $Top3Fin_i$ (online at the *Journal of Finance*, the *Journal of Financial Economics*, or the *Review of Financial Studies*), columns (3)–(4) $Top5Fin_i$ (the top three finance journals plus the *Journal of Financial and Quantitative Analysis* and the *Review of Finance*), and columns (5)–(6) $Top10Fin_i$ (the top five plus the *Journal of Corporate Finance*, the *Journal of Banking & Finance*, *Financial Management*, the *Journal of Financial Markets*, and the *Journal of Financial Intermediation*), all measured within two years of each paper’s meeting. The sample is the 276 papers presented at the 2024 AFA Annual Meeting and the six 2024 NBER finance meetings outside the AFA PhD student poster session that enter the regressions (Section 2.4). The overall score is the mean over four independent draws of GPT-4o (gpt-4o-2024-11-20) under the naive prompt reproduced in Section IA.2, standardized across papers. All specifications include the log length of the paper text, the number of authors, $Top20School_i$, $Log Prior Top_i$, $Editorial_i$, and a conference (AFA versus NBER) fixed effect. All variables are defined in Table IA.9. Robust z -statistics in parentheses. ***, **, and * denote significance at the 1%, 5%, and 10% levels.

	<i>Top3Fin_i</i>		<i>Top5Fin_i</i>		<i>Top10Fin_i</i>	
	(1)	(2)	(3)	(4)	(5)	(6)
	Title	Title and authors	Title	Title and authors	Title	Title and authors
<i>Overall_i</i>	0.491** (2.29)	0.466** (2.18)	0.456** (2.34)	0.435** (2.25)	0.578*** (2.88)	0.568*** (2.80)
<i>Claim_i</i>	-0.108 (-0.13)	-0.313 (-0.43)	-0.448 (-0.53)	-0.505 (-0.73)	-0.469 (-0.55)	-0.525 (-0.77)
<i>Overall_i × Claim_i</i>	0.113 (0.20)	0.291 (0.69)	0.125 (0.23)	0.288 (0.71)	-0.072 (-0.13)	0.151 (0.38)
<i>Log Length_i</i>	0.019 (0.03)	0.025 (0.04)	0.143 (0.26)	0.163 (0.30)	-0.094 (-0.17)	-0.064 (-0.11)
<i>N Authors_i</i>	0.595** (2.44)	0.593** (2.35)	0.681*** (2.88)	0.656*** (2.77)	0.681*** (2.89)	0.642*** (2.73)
<i>Top20School_i</i>	0.171 (0.34)	0.177 (0.35)	-0.207 (-0.41)	-0.198 (-0.39)	-0.392 (-0.82)	-0.385 (-0.81)
<i>Log Prior Top_i</i>	0.110 (0.39)	0.117 (0.42)	0.093 (0.34)	0.112 (0.42)	0.035 (0.13)	0.063 (0.24)
<i>Editorial_i</i>	-0.060 (-0.10)	-0.070 (-0.12)	0.396 (0.70)	0.379 (0.66)	0.310 (0.56)	0.295 (0.53)
Conference FE	Y	Y	Y	Y	Y	Y
N Obs	276	276	276	276	276	276
Pseudo R ²	0.072	0.073	0.087	0.088	0.093	0.093

IA.6 Additional Results

Specification details. Dai et al. (2023) estimate the publication outcome by logit with year and field fixed effects and cluster standard errors by year; our sample is a single conference year, so the fixed effect distinguishes AFA from NBER papers and we report heteroskedasticity-robust standard errors. The length and author-count controls correspond to the log page count and coauthor count of Dai et al. (2023), and the prior top-three record and $Editorial_i$ take the place

of their coauthor-network centrality. The dimension scores enter one at a time because the overall score is, by the design of the protocol, a summary of the four dimensions.

Controls. Among the controls in [Table 3](#), only the number of authors is significant throughout: each additional author raises the log odds of top-three publication by about 0.60 ($z = 2.45$ in column 1); the other controls are insignificant ([Section 3.1](#)). The coauthor effect contrasts with the negative coauthor coefficient that [Dai et al. \(2023\)](#) find in the universe of SSRN postings; the difference is consistent with conference selection having already removed the papers for which team size proxies for lower quality.

Broader journal sets. [Table IA.6](#) reports the estimates for $Top5Fin_i$ and $Top10Fin_i$, which [Sections 3.1](#) and [3.2](#) summarize. Exposition predicts both outcomes (0.636, $z = 2.66$, and 0.633, $z = 2.76$), correctness does so at the 10% level (0.546, $z = 1.74$, and 0.551, $z = 1.88$), and originality and significance do not.

Table IA.6: The LLM Referee’s Scores and Publication in the Broader Finance Journal Sets

This table presents logit estimates of equation (1) for the two broader finance outcomes of Dai et al. (2023), measured within two years of each paper’s meeting: $Top5Fin_i$ (the top three finance journals plus the *Journal of Financial and Quantitative Analysis* and the *Review of Finance*) in columns (1)–(5), and $Top10Fin_i$ (the top five plus the *Journal of Corporate Finance*, the *Journal of Banking & Finance*, *Financial Management*, the *Journal of Financial Markets*, and the *Journal of Financial Intermediation*) in columns (6)–(10). The sample is the 276 papers presented at the 2024 AFA Annual Meeting and the six 2024 NBER finance meetings outside the AFA PhD student poster session that enter the regressions (Section 2.4). Scores are the means over four independent draws of GPT-4o (gpt-4o-2024-11-20) under the naive prompt reproduced in Section IA.2, standardized across papers; each column enters one score, and a coefficient is the change in log odds per standard deviation of the score. All specifications include the log length of the paper text, the number of authors, $Top20School_i$, $Log Prior Top_i$, $Editorial_i$, and a conference (AFA versus NBER) fixed effect. All variables are defined in Table IA.9. Robust z-statistics in parentheses. ***, **, and * denote significance at the 1%, 5%, and 10% levels.

	$Top5Fin_i$					$Top10Fin_i$				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
$Overall_i$	0.451** (2.39)					0.561*** (2.92)				
$Originality_i$		-0.115 (-0.56)					-0.115 (-0.62)			
$Significance_i$			-0.008 (-0.03)					0.019 (0.08)		
$Correctness_i$				0.546* (1.74)					0.551* (1.88)	
$Exposition_i$					0.636*** (2.66)					0.633*** (2.76)
$Log Length_i$	0.145 (0.27)	0.180 (0.35)	0.178 (0.33)	0.102 (0.18)	-0.179 (-0.32)	-0.083 (-0.15)	-0.038 (-0.07)	-0.051 (-0.09)	-0.137 (-0.25)	-0.403 (-0.72)
$N Authors_i$	0.664*** (2.90)	0.625*** (2.72)	0.626*** (2.74)	0.668*** (2.89)	0.689*** (3.01)	0.650*** (2.89)	0.595*** (2.65)	0.595*** (2.67)	0.642*** (2.86)	0.664*** (3.02)
$Top20School_i$	-0.203 (-0.40)	-0.105 (-0.21)	-0.134 (-0.27)	-0.164 (-0.33)	-0.221 (-0.46)	-0.385 (-0.81)	-0.278 (-0.59)	-0.310 (-0.67)	-0.343 (-0.73)	-0.402 (-0.89)
$Log Prior Top_i$	0.100 (0.38)	0.078 (0.29)	0.088 (0.33)	0.114 (0.45)	0.100 (0.35)	0.051 (0.20)	0.024 (0.09)	0.032 (0.12)	0.058 (0.23)	0.036 (0.13)
$Editorial_i$	0.372 (0.66)	0.460 (0.83)	0.460 (0.83)	0.325 (0.56)	0.399 (0.71)	0.278 (0.50)	0.394 (0.73)	0.394 (0.73)	0.260 (0.47)	0.325 (0.59)
Conference FE	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
N Obs	276	276	276	276	276	276	276	276	276	276
Pseudo R ²	0.086	0.067	0.066	0.095	0.110	0.091	0.060	0.059	0.090	0.103

GPT-4o and GPT-5 scores. Table IA.7 compares the scores that the two models give the same papers (Section 3.5). GPT-5 scores correctness lower for every paper, by 1.74 points for the finance papers and 2.28 points for the accounting papers, and scores originality and significance lower on average; its overall score is unchanged for the finance papers and 3.46 points lower for the accounting papers.

Table IA.7: GPT-4o and GPT-5 Scores of the Same Papers

This table compares the scores that GPT-4o (gpt-4o-2024-11-20) and GPT-5 (gpt-5-2025-08-07, minimal reasoning effort) give the same papers under the naive prompt. Panel A reports the 276 papers presented at the 2024 AFA Annual Meeting and the six 2024 NBER finance meetings outside the AFA PhD student poster session that enter the regressions, and Panel B the 41 accounting papers presented at the 2023 and 2024 Conferences on Financial Economics and Accounting. Each paper’s score is the mean over its four draws. The difference is the mean of GPT-5’s score minus GPT-4o’s score, with the paired t -statistic in parentheses; ***, **, and * denote significance at the 1%, 5%, and 10% levels. Higher, Lower, and Same report the percentage of papers for which GPT-5’s score is above, below, or equal to GPT-4o’s.

	Mean		Difference		GPT-5 score (% of papers)		
	GPT-4o	GPT-5	GPT-5 – 4o	t	Higher	Lower	Same
<i>Panel A: Finance (276 papers)</i>							
<i>Overall_i</i>	81.92	81.93	0.01	(0.03)	57.6	40.6	1.8
<i>Originality_i</i>	8.31	8.02	-0.29***	(-7.38)	26.4	58.3	15.2
<i>Significance_i</i>	8.12	7.89	-0.22***	(-5.84)	29.7	50.0	20.3
<i>Correctness_i</i>	8.40	6.66	-1.74***	(-43.77)	0.0	100.0	0.0
<i>Exposition_i</i>	7.21	7.18	-0.02	(-0.67)	40.2	39.5	20.3
<i>Panel B: Accounting (41 papers)</i>							
<i>Overall_i</i>	82.54	79.08	-3.46***	(-5.96)	19.5	78.0	2.4
<i>Originality_i</i>	8.18	7.64	-0.54***	(-7.09)	9.8	82.9	7.3
<i>Significance_i</i>	8.06	7.45	-0.61***	(-7.66)	9.8	87.8	2.4
<i>Correctness_i</i>	8.51	6.23	-2.28***	(-27.50)	0.0	100.0	0.0
<i>Exposition_i</i>	7.23	7.03	-0.20***	(-3.35)	12.2	56.1	31.7

Correlations of the scores. Table IA.8 reports the correlations across papers of the five scores that GPT-4o gives under the naive prompt, which Section 3.2 discusses.

Table IA.8: Correlations of the LLM Referee’s Scores

This table reports Pearson correlations across papers of the scores that GPT-4o (gpt-4o-2024-11-20) gives under the naive prompt. Panel A reports the 276 papers presented at the 2024 AFA Annual Meeting and the six 2024 NBER finance meetings outside the AFA PhD student poster session that enter the regressions, and Panel B the 41 accounting papers presented at the 2023 and 2024 Conferences on Financial Economics and Accounting. Each paper’s score is the mean over its four draws.

	(1)	(2)	(3)	(4)	(5)
<i>Panel A: Finance (276 papers)</i>					
(1) <i>Overall_i</i>	1				
(2) <i>Originality_i</i>	0.483	1			
(3) <i>Significance_i</i>	0.476	0.293	1		
(4) <i>Correctness_i</i>	0.564	0.053	-0.077	1	
(5) <i>Exposition_i</i>	0.618	0.254	0.229	0.406	1
<i>Panel B: Accounting (41 papers)</i>					
(1) <i>Overall_i</i>	1				
(2) <i>Originality_i</i>	0.525	1			
(3) <i>Significance_i</i>	0.456	0.454	1		
(4) <i>Correctness_i</i>	0.534	-0.130	0.007	1	
(5) <i>Exposition_i</i>	0.555	0.500	0.170	0.337	1

IA.7 Variable Definitions

Table IA.9: Variable Definitions

This table defines the variables used in the analyses. Scores are the mean over four independent draws of gpt-4o-2024-11-20, or of gpt-5-2025-08-07 in Table 6, at the default temperature; standardized versions have zero mean and unit variance across the papers of each regression. Publication outcomes are measured within two years of each paper’s meeting.

Variable	Definition
<i>Panel A: LLM referee scores</i>	
$Overall_i$	Overall score on a 0–100 scale, mean over draws.
$Originality_i$	Score on a 0–10 scale for whether the paper presents a genuinely new idea, angle, dataset, design, or synthesis relative to established academic knowledge (Newton and Riddiough, 2026).
$Significance_i$	Score on a 0–10 scale for how important the question is and how large the contribution is if the results are correct.
$Correctness_i$	Score on a 0–10 scale for whether the analysis, empirical design, theoretical logic, and interpretation are scientifically convincing.
$Exposition_i$	Score on a 0–10 scale for how clear, well organized, and well communicated the paper is.
<i>Panel B: Publication outcomes</i>	
$Top3Fin_i$	One if the paper is online at the <i>Journal of Finance</i> , the <i>Journal of Financial Economics</i> , or the <i>Review of Financial Studies</i> within two years of the paper’s meeting (Dai et al., 2023).
$Top5Fin_i$	One if $Top3Fin_i = 1$ or the paper is online at the <i>Journal of Financial and Quantitative Analysis</i> or the <i>Review of Finance</i> (Dai et al., 2023).
$Top10Fin_i$	One if $Top5Fin_i = 1$ or the paper is online at the <i>Journal of Corporate Finance</i> , the <i>Journal of Banking & Finance</i> , <i>Financial Management</i> , the <i>Journal of Financial Markets</i> , or the <i>Journal of Financial Intermediation</i> (Dai et al., 2023).
$Top3Acc_i$	One if the paper is online at <i>The Accounting Review</i> , the <i>Journal of Accounting Research</i> , or the <i>Journal of Accounting and Economics</i> within two years of the paper’s meeting.
<i>Panel C: Paper characteristics</i>	
$Log\ Length_i$	Natural logarithm of the number of tokens (o200k tokenizer) of the cleaned paper text.
$N\ Authors_i$	Number of authors listed on the conference program.
$NBER_i$	One if the paper was presented at a 2024 NBER finance meeting.
<i>Panel D: Author reputation (Section IA.4)</i>	
$Top20School_i$	One if any author’s affiliation on the conference program is one of the twenty schools of the UT Dallas Top 100 Business School Research Rankings restricted to the <i>Journal of Finance</i> , <i>Journal of Financial Economics</i> , and <i>Review of Financial Studies</i> , worldwide, 2019–2023 (Dai et al., 2023); the same list is used for the accounting papers.
$Established_i$	One if any author of the paper has an article in the <i>Journal of Finance</i> , <i>Journal of Financial Economics</i> , or <i>Review of Financial Studies</i> published before 2024 (Card and DellaVigna, 2020).
$Log\ Prior\ Top_i$	Natural logarithm of one plus the largest number of articles in the <i>Journal of Finance</i> , <i>Journal of Financial Economics</i> , or <i>Review of Financial Studies</i> published before 2024 by any author of the paper, from OpenAlex; for the accounting papers, articles in <i>The Accounting Review</i> , the <i>Journal of Accounting Research</i> , or the <i>Journal of Accounting and Economics</i> .
$Editorial_i$	One if any author of the paper is listed as an editor (editor, co-editor, senior or executive editor, editor-in-chief) or associate editor of the <i>Journal of Finance</i> , <i>Journal of Financial Economics</i> , or <i>Review of Financial Studies</i> in any year from 2014 to 2024, the ten years up to the conference, from the journals’ editorial-board pages and issue mastheads (Brogaard et al., 2014); for the accounting papers, of <i>The Accounting Review</i> , the <i>Journal of Accounting Research</i> , or the <i>Journal of Accounting and Economics</i> .
$Early\ Stage_i$	One if the first OpenAlex publication of every matched author of the paper is in 2014 or later, the ten years before the conference (Dai et al., 2023, Table 11).
<i>Panel E: Look-ahead probe (Section IA.5)</i>	
$Claim_i$	One if, asked in a fresh session at temperature 0 for the publication status of the paper given its conference title, or its title and authors, the model answers that the paper is published or accepted; one variable per information set.