Writing Matters: Generative AI as an Academic Impact Equalizer

Ron Kaniel[*] Huaxia Rui[†] Shujing Sun[‡] Pingle Wang[§]

June 2025

Abstract

We study the effect of generative AI, specifically ChatGPT, on academic writing quality and subsequent citations. Leveraging a novel dataset of over 120,000 papers from financial economics, management, and political science, we construct detailed writing quality measures using both traditional linguistic metrics and ChatGPT-based evaluations. We find that higher writing quality strongly predicts future citations. Following the release of ChatGPT in late 2022, writing quality improved significantly, particularly for authors from non-English-speaking institutions, those with non-English names, and those with fewer historical citations. These improvements translate into a narrowing of the citation count gap between these groups and their counterparts, and consequently to a more equal distribution of academic impact.

Keywords: Generative AI, ChatGPT, Academic Writing

[*]Ron Kaniel, University of Rochester, email: ron.kaniel@simon.rochester.edu
[†]Huaxia Rui, University of Rochester, email: huaxia.rui@simon.rochester.edu.
[‡]Shujing Sun, The University of Texas at Dallas, email: shujing.sun@utdallas.edu.
[§]Pingle Wang, The University of Texas at Dallas, email: pingle.wang@utdallas.edu.
We thank Shenxiong Li for his early contribution to the paper. All errors are our own.

"You can have brilliant ideas, but if you can't get them across, your ideas won't get you anywhere." – Lee Iacocca, former Ford president

1 Introduction

In academia, a successful paper rests on three foundational pillars: the novelty of ideas and results, the rigor of execution, and the clarity and engagement of exposition. While the first two dimensions are well recognized as central to scholarly impact, the third is often treated, at least by some, as secondary or cosmetic. Yet writing can play a critical role in shaping its influence. A well-executed idea may fail to resonate if poorly articulated, while a clearly written paper can entice and captivate readers, transcend disciplinary boundaries, and reach a broader audience. In an increasingly competitive academic environment, writing quality functions as a form of academic marketing, which is crucial in attracting attention, enhancing the visibility, and recognition of research (Warren et al., 2021).

The importance of exposition is particularly critical in social science disciplines such as financial economics, management, and political science, where complex ideas must be conveyed with precision and nuance. Yet conveying such ideas in an accessible and engaging way is no small feat, particularly in an uneven academic landscape. Scholars from English-speaking backgrounds often enjoy a distinct advantage in academic publishing (Flowerdew, 2008; Hyland, 2016). Their writing tends to be more polished, increasing the likelihood that their work captures attention and earns citations, a metric measuring academic impact. In contrast, non-native English speakers face additional linguistic hurdles that can obscure the value of their research, regardless of the underlying intellectual merit. Publishing in English, the de facto language of global scholarship, requires not only domain expertise but

also mastery of English-writing skills. Unfortunately for many non-native English speakers, the linguistic challenge becomes a decisive factor in whether their work is recognized or overlooked, regardless of the underlying research quality.

The emergence of generative artificial intelligence (AI), particularly models like ChatGPT, marks a potential turning point in how academic writing is produced and refined. Unlike traditional writing aids such as grammar checkers or style suggestions, large language models (LLMs) like ChatGPT offer context-aware feedback that can help structure arguments, refine tone, improve coherence, and elevate overall writing quality. Consequently, it seems they hold great potential to support democratized access to effective scholarly communication (Banerjee et al., 2023).

This raises several fundamental questions. First, can generative AI meaningfully enhance writing quality? Any individual with some experience using generative AI tools such as ChatGPT would likely conjecture that the answer is yes. Second, if so, how does improved writing translate into readership and impact? The answer here relies on the importance of writing quality in generating impact, independent of the novelty of the ideas and findings. A challenge in isolating this effect is a potential association between research quality and the exposition conveying the research. Third, and more critically, can generative AI help level the playing field in academic publishing, particularly for scholars from historically disadvantaged groups who face systemic barriers to visibility? After all, since academic impact is at least in part a tournament, and if AI tools improve writing across the board, they may benefit everyone equally and leave existing inequalities unchanged.

In this paper, we conclude that the answer to all three questions is affirmative. Generative AI not only improves writing quality, but that improvement is associated with increased scholarly impact, as measured by citations. Moreover, the benefits of AI are not evenly

distributed: they are more pronounced for scholars who face greater linguistic or structural barriers, suggesting that AI use can help level the playing field in academic publishing. We demonstrate these findings by compiling a novel dataset that links OpenAlex metadata with full-text SSRN documents, covering over 120,000 working and published papers across financial economics, management, and political science. Leveraging the late 2022 release of ChatGPT as a natural experiment, we examine its impact on writing quality and downstream citations. Through comprehensive mining of the texts from introduction sections, the most widely read and rhetorically significant part of academic articles, we construct comprehensive measures of writing quality that integrate conventional linguistic features with human-like evaluations generated by ChatGPT. Our findings show that ChatGPT meaningfully enhances writing quality, especially for authors from non-English-speaking backgrounds or with historically lower citations. As a result, these authors experience greater citation gains, and overall inequality in both writing quality and academic impact declines. Together, these results underscore the potential of generative AI to foster greater equity in scholarly recognition by supporting those previously disadvantaged by linguistic barriers.

Our analysis proceeds in the following steps. First, we document a strong positive correlation between writing quality and future citations. This finding holds across various dimensions of writing quality, including lexical diversity, semantic cohesion, and ChatGPT-generated holistic readability scores. Notably, the GPT-based assessment has the largest explanatory power for citations, highlighting the relevance of LLMs as proxies for human perception of writing quality.

Second, we evaluate how writing quality has changed over time, particularly following the release of ChatGPT. Controlling for field-by-time fixed effects, we find significant improvements in writing quality post-2022, after the release of ChatGPT. Importantly, we

disentangle the effect from Grammarly, a popular writing tool that has reached a steady 30 million daily active users since 2020. The findings show that gains in writing quality during the "Grammarly era" (2020-2022) are much smaller than those observed after the release of ChatGPT. Given the obvious advantage of generative AI tools over Grammarly in writing assistance, we take these results as evidence that ChatGPT has significantly elevated the quality of academic writing during our sample period.

Third, we leverage cross-author heterogeneity to understand who benefits the most. Authors affiliated with non-English institutions, those with non-English last names, and those with fewer historical citations see the largest improvements in writing quality. The gap in writing quality shrinks by almost 44% after ChatGPT was released. Importantly, these authors also experienced the most pronounced increases in citation counts after the introduction of ChatGPT, compared to others. Specifically, the citation difference between papers with and without an English-speaking affiliation reduces by 47.8% post-ChatGPT. The results suggest that generative AI tools help level the playing field, particularly for those historically disadvantaged by language barriers.

Finally, and perhaps most importantly, we find a reduction in disparities both in terms of writing quality and in terms of citations, as they become more evenly distributed post-ChatGPT, as measured by a decline in their Gini coefficients. This flattening of the distribution is driven primarily by gains at the lower end: previously under-cited papers now attract more attention, consistent with improved readability and accessibility. Mediation analysis confirms that enhanced writing quality is a key channel through which generative AI reduces citation inequality.

This paper contributes to several strands of literature. First, it adds to the growing body of research that links writing quality to scholarly impact. Prior studies have shown

that clearer exposition improves reader comprehension and increases citations (Warren et al., 2021; Plavén-Sigray et al., 2017). Compared to existing work, we construct comprehensive, multi-dimensional measures of writing quality, drawing from both the conventional linguistic literature and human-like evaluations generated by large language models. Applying these measures to a large corpus of over 120,000 social science papers, we provide robust, large-scale evidence that writing quality is a significant predictor of future citations. Moreover, and probably more important, a potential limitation of prior work is that the evidence is potentially conflated by endogeneity concerns related to the potential association between the quality of the research and the exposition. By utilizing the introduction of ChatGPT as a natural experiment that facilitates analysis that alleviates such concerns, we substantiate a causal link between writing quality and impact.

Second, we contribute to the literature on linguistic and institutional barriers in academic publishing. Scholars have long documented the disadvantages faced by non-native English speakers and researchers outside English-speaking institutions (Flowerdew, 2008; Hyland, 2016). These barriers manifest not only in publication rates but also in downstream visibility and long-term scholarly influence. Our findings suggest that generative AI, specifically ChatGPT, can partially mitigate these disparities. We show that writing quality improves more for authors from non-English-speaking institutions or with non-English names, and the higher writing quality further translates into increased citations. This heterogeneity in impact highlights the potential of generative AI to reduce structural inequality in academic recognition and amplify the voices of researchers from linguistically and institutionally marginalized backgrounds. Here again, the natural experiment nature of our work helps in establishing a clear link.

Third, we contribute to the nascent literature on the implications of large language

models for knowledge production. Recent work has begun to explore how LLMs affect coding productivity, education, and decision-making (Noy and Zhang, 2023), but little empirical evidence exists on their role in shaping academic writing and dissemination. Our study is novel in quantifying how generative AI changes scholarly communication, both by raising the overall quality of exposition and by redistributing visibility toward previously under-cited authors. In doing so, we offer a new perspective on how AI can impact the structure of academic influence itself.

The rest of the paper proceeds as follows. Section 2 describes the data and construction of writing quality measures. Section 3 presents the empirical findings, including the relationship between writing quality and citations, the differential impact of AI tools, and evidence on reduced inequality. Section 4 concludes.

2 Data and Variable Construction

Our analysis is based on data from OpenAlex, a comprehensive and open-source catalog of scholarly works, authors, institutions, and citation metrics. We focus on research papers in four related social science fields: economics, financial economics, management, and political science, based on the classification from OpenAlex. From OpenAlex, we extract authors' information (including names and institutional affiliations) and metadata on individual papers, such as written year and yearly citation counts.

We match papers from OpenAlex to corresponding entries on the Social Science Research Network (SSRN), which provides access to full-text PDFs. For each matched paper, we download the PDF and extract the introduction section. This allows us to construct a range of writing quality measures, including readability scores and metrics derived from

large language models such as ChatGPT. We focus on the introduction because it is the most important part of a scholarly article and shapes the reader's first impression. An introduction sets the stage for the research, motivates the research, outlines the problem and objectives, and often highlights the main findings and contributions. A well-written introduction showcases the paper to readers, convincing them of the importance of the question analyzed and the results; reeling readers in. It also shows the author's ability to deliver the core information from a paper, making it a strong indicator of overall writing quality and academic rigor of the work.

ChatGPT was publicly released on November 30, 2022. We introduce a dummy variable *PostGPT*, which equals one if the year is post-2022 and zero otherwise. Prior to ChatGPT's launch, Grammarly was the leading tool for refining writing. Although Grammarly debuted in 2009, it took a while for it to gain popularity. It achieved 30 million daily active users in 2020 and sustained this level thereafter. To isolate Grammarly's impact from that of ChatGPT, we define a dummy variable *Grammarly*, which equals one if the year is between 2020 and 2022, and zero otherwise. Our study period spans from 2010 to 2024, with 2024 selected as the end year due to it being the last data collection point. The starting year, 2010, is chosen to establish a ten-year benchmark from 2010 to 2019 for evaluating the influence of writing tools on writing quality and subsequent citations.

The resulting dataset enables us to study both the writing quality of research papers and their downstream academic impact on citations. Note that our dataset contains both working papers and papers that have been published. The inclusion of working papers is important, as papers in social science usually take many years to be published. The author information further enables us to leverage the difference-in-differences design and examine

https://www.grammarly.com/blog/company/grammarly-12-year-history/

heterogeneity across author characteristics and institutional backgrounds. Throughout the paper, we focus on cumulative three-year citation counts for articles that have received at least one citation. Our results are robust to alternative horizons in counting the cumulative citations from 1 year and up.

To evaluate the quality of academic writing in our sample, we construct several quantitative indices grounded in well-established linguistic research. These measures span three core dimensions of writing quality: readability, lexical diversity, and text cohesion. Below, we describe each measure in detail.

2.1 Readability

The first set of measures focuses on readability, or the clarity in academic writing. Specifically, we construct a composite readability score by averaging three classic indices: the Flesch-Kincaid Index (FKI), the Automated Readability Index (ARI), and the Gunning Fog Index (GFI). Each index captures a different aspect of textual complexity. The first measure is the Flesch-Kincaid Index (FKI), introduced by Flesch (1948) and Kincaid et al. (1975). The FKI assesses readability based on two main components: average sentence length (words per sentence) and average syllables per word. Higher FKI values indicate a higher U.S. school grade level needed to understand the text. The second measure is the Automated Readability Index (ARI) proposed by Smith and Senter (1967). The ARI evaluates text based on characters per word and words per sentence, using characters instead of syllables to estimate word complexity. Similar to FKI, higher ARI scores indicate a higher grade level required for comprehension. The third measure is the Gunning Fog Index (GFI) developed by Gunning (1952). GFI places emphasis on the proportion of complex words (i.e., words with three

When we compare citations across research papers, we use field-by-written year fixed effects for peer comparison.

or more syllables), excluding proper nouns, familiar jargon, and compound words. Given the slightly different emphasis by each measure, we construct a composite readability index by taking the average of the three scores to offer a comprehensive view of readability that focuses on sentence structure, word complexity, and length characteristics, and denote it as *ReadScore*.

The measure *ReadScore* reflects the complexity of writing. However, it's not always clear whether greater complexity translates to better writing quality. On one hand, academic writing requires a certain level of sophistication to effectively convey ideas, so a low *ReadScore* doesn't necessarily indicate better writing. For example, elementary school writing level yields a low *ReadScore*, yet clearly is insufficient for coherently conveying academic research. On the other hand, overly long and complex sentence structures can confuse readers and reduce clarity, ultimately lowering the quality of the writing. To address the limitations of this traditional metric, we introduce several additional measures.

2.2 Lexical Diversity

The second set of measures captures the richness of the writing. To measure lexical variety while minimizing sensitivity to text length, we use the Measure of Textual Lexical Diversity (MTLD) from McCarthy and Jarvis (2010). Unlike traditional type-token ratio (TTR) measures, MTLD calculates the number of word segments in which the TTR remains above a predefined threshold, accounting for both forward and backward traversals of the text. Higher MTLD values indicate richer and more varied vocabulary use. This makes MTLD especially relevant for evaluating the writing quality of scholarly articles, where texts tend to be long and structurally complex. In this context, higher MTLD values reflect a greater degree of lexical diversity, which is a key marker of strong academic writing.

2.3 Text Cohesion

The final set of measures captures the degree of connectedness in academic writing. Cohesion plays a vital role in understanding texts, especially for complex texts that present knowledge demands to the reader (McNamara and Kintsch, 1996; McNamara et al., 1996). As a result, we measure text cohesion as a key component in measuring the writing quality of academic articles. To assess textual cohesion, we use paragraph-level lexical and semantic overlap indices adapted from the TAACO tool developed by Crossley et al. (2016). We focus on two important indices to measure text cohesion: *Lexical Overlap* captures the lemma repetition across adjacent paragraphs and across spans of two paragraphs, representing how vocabulary is reused within the introduction. *Semantic Overlap* is based on the cosine similarity between adjacent paragraphs using latent semantic analysis (LSA), again at both one- and two-paragraph intervals. These measures reflect conceptual and thematic coherence. This dual approach allows us to capture both surface-level and deeper semantic connections across the text.

Lastly, we create a composite non-GPT-based measure by taking the simple average of the four measures described above, which captures the overall assessment of a paper's writing quality using traditional measures.

2.4 ChatGPT-based Readability

In addition to the established linguistic features discussed above, we incorporate a composite readability score generated by ChatGPT-4o. Specifically, we ask ChatGPT to evaluate the writing quality of the introduction section of an article across nine dimensions: readability, lexical diversity, sentence fluency, cohesion, coherence, syntactic complexity, semantic rich-

Lemmas are base forms of words. For example, the lemma of "running", "ran", and "runs" is "run".

ness, engagement, and persuasiveness, each rated on a scale from 1 to 10. We then take the average of these scores as our ChatGPT-based score. Unlike traditional metrics, which isolate specific dimensions such as sentence length, word complexity, or lexical diversity, the ChatGPT-based score provides a holistic assessment of the writing quality. As such, this composite score from the large language model serves as a valuable complement to traditional metrics and offers a perspective that more closely aligns with how human readers evaluate scholarly writing.

3 Empirical Results

This section presents our empirical findings in four steps, progressing towards the main thesis of our paper. We begin by establishing the relationship between writing quality and citations. We then show that writing quality improves after the introduction of ChatGPT, especially among disadvantaged authors: those affiliated with non-English-speaking institutions, non-English authors, and those with lower prior citation counts. Third, the improvement in writing quality translates into increased citations for these groups. Finally, we document a decline in inequality in both writing quality and citation outcomes, driven primarily by gains among less-cited papers.

3.1 Writing Quality Predicts Citations

We first examine the relationship between writing quality and a paper's academic impact. Table 1 presents regressions of log cumulative citations (measured three years post-publication) on a range of writing quality metrics. All specifications include field-by-written year fixed effects, so that we are comparing papers within their cohorts written around the same time.

In addition, all independent variables are standardized so that the magnitudes of coefficient estimates are comparable across different measures.

As shown in columns (1) to (4) of Table 1, papers with higher writing quality, proxied by the traditional measures, are associated with significantly higher future citations. Note that the coefficient estimate of *ReadScore* is an order of magnitude smaller than other measures. The measure *ReadScore* captures the number of years of education required to understand the content, but it is not clear whether this measure should be positively correlated with the impact of academic papers.

In column (5), we create a non-GPT-based writing quality measure by equal weighting all four traditional measures, and the result is robust. These findings suggest that writing quality plays an important role in how widely it is cited, which is consistent with a recent work that suggests writing clarity will influence citations (Warren et al., 2021).

More importantly, among writing quality metrics, GPT-readability exhibits the largest effect size: a one standard deviation increase is associated with a 15.5% increase in citations, as shown in column (6). In column (7), we include both GPT-based and non-GPT-based measures as independent variables, and the results suggest that the GPT-based readability measure is better able to explain cross-sectional variations in paper citations than the non-GPT-based measure.

3.2 The Impact of AI on Writing Quality

The rise of AI-powered writing tools, starting with the widespread adoption of Grammarly around 2020 and accelerating with the introduction of advanced models like ChatGPT in November 2022, has begun to reshape how people write. Grammarly made it easier to catch grammar issues and provided stylistic suggestions. But ChatGPT goes much further. It can

help organize ideas, tighten arguments, reorganize structures, and even perform context-aware rewriting, making it a powerful assistant not just for editing, but for thinking through and expressing complex ideas. Moreover, the introduction of ChatGPT was followed by a worldwide hype and an exponential adoption rate; within a couple of months, large parts of the academic community were using it.

We empirically test whether writing quality has changed following the introduction of ChatGPT in late 2022. Specifically, we regress each writing metric on two indicator variables: one for the Grammarly period (2020–2022) and one for the Post-ChatGPT period (2023 onward). We also control for field fixed effects. As shown in Table 2, the post-ChatGPT period is associated with substantial and statistically significant improvements in writing quality, except for *ReadScore*, which is marginally insignificant, as shown in column (1). In contrast, gains during the Grammarly period are smaller and less consistent, suggesting that recent advances in generative AI are more impactful than earlier tools.

Although the above findings capture before-and-after changes corresponding to the introduction of ChatGPT, they may not necessarily reflect the causal treatment effect of AI. A key challenge in this context is that the treatment is available to all, leaving no control group for a conventional difference-in-differences (DiD) framework. Nonetheless, we can leverage the heterogeneity in authors' English writing capabilities to identify the treatment effect. The rationale is the following: authors with greater exposure to English, proxied by the English names and affiliations with English-speaking institutions, tend to have stronger baseline English writing skills. As a result, they will rely less on ChatGPT for language assistance. Leveraging this variation, we compare changes in outcome variables between authors with high and low English exposure before and after the introduction of ChatGPT. In particular,

we adopt the following specification:

$$\begin{aligned} Y_{it} = \quad & \alpha_0 + \alpha_1 HighEngSkill_i + \alpha_2 PostGPT_t \\ & + \beta HighEngSkilli \times PostGPT_t + \alpha_3 X_{it} + \varepsilon_{it}, \end{aligned} \tag{1}$$

where we use authors' institutional affiliation, ethnicity, and prior academic influence to proxy for English skills. Specifically, we define the variable *EnglishInst* as equal to one if at least one author is affiliated with an English-speaking institution. The variable *EnglishName* equals one if at least one author has an English last name, using U.S. Census and Florida voter registration data. To capture prior academic standing, we use *Log(Max Team Cites)*, which is the natural logarithm of the cumulative citations of the most-cited author on the team, measured up to the year before the paper was written. We then estimate regressions of GPT-based readability scores on these variables and their interactions with both *Grammarly* and *PostGPT* indicators, including field-by-year fixed effects to control for temporal and disciplinary trends.

Table 3 presents the results. First, as expected, papers authored by teams with at least one English-speaking affiliation, at least one author with an English last name, or a more highly cited team member tend to exhibit higher baseline writing quality. Second, and more strikingly, authors from non-English backgrounds and teams with fewer prior citations show significantly greater improvements, compared to others, in writing quality following the introduction of ChatGPT, as indicated by the negative and highly significant interaction terms with the *PostGPT* dummy. Taking column (1) for example, the gap in writing quality between author teams with and without English-speaking affiliations shrinks by 44% (= 0.166/0.378). This pattern suggests that generative AI may help reduce longstanding linguistic and structural barriers in academic writing. Finally, consistent with earlier findings

in Table 2, we find no meaningful effect associated with the rollout of Grammarly.

3.3 The Impact of AI on Citations

Does better writing lead to greater impact? We turn to this question by looking at how citation patterns change across different groups of authors. As before, our outcome of interest is the log of cumulative citations a paper receives within three years of being written. We estimate a set of regressions using *EnglishInst*, *EnglishName*, and *Log(Max Team Cites)* as key variables, along with their interactions with the *Grammarly* and *PostGPT* indicators. All models include field-by-year fixed effects to account for differences across disciplines and over time.

The results are shown in Table 4. First, before the introduction of ChatGPT, authors from more advantaged backgrounds tend to receive significantly more citations. These gaps are sizable: papers with at least one author from an English-speaking institution get 37% more citations on average, and those with at least one author with an English last name get 24% more. Second, these citation gaps narrow significantly after ChatGPT becomes available. For example, the citation difference between papers with and without an English-speaking affiliation shrinks by 17.7%, which amounts to a 47.8% relative reduction. This pattern echoes with our earlier finding that writing quality improved more for historically disadvantaged groups, suggesting that the sudden availability of generative AI tools might have reduced the long-standing visibility gap through its excellent writing support.

3.4 Declines in Writing and Citation Inequality

We now turn to aggregate outcomes. Tables 5 and 6 document a sharp decline in inequality, both in writing quality and citations, following the introduction of ChatGPT. Table 5 shows

that the Gini coefficients of both GPT-readability and non-GPT-readability fall significantly after 2022, across all major fields. Taking column (1) of Panel A for example, the Gini Coefficient of GPT-readability reduces by 0.008, which is a 20.5% reduction relative to the pre-ChatGPT period. Similarly, in Panel B column (1), the Gini Coefficient of non-GPT-readability reduces by 0.023, which is a 20.2% reduction.

Table 6 presents similar patterns for citations. Across all fields, the Gini coefficient of cumulative citations drops by 0.09 (13.5% reduction), while the Gini of log citations declines by 0.04 (13% reduction). These are sizable reductions given the persistence of inequality in academic recognition. Specifically, the adjusted R^2 from Table 6 are large. This is because the Gini Coefficients are very persistent prior to the release of ChatGPT, and only sharply went down afterwards. In addition, the Grammarly period is not associated with similar declines, reinforcing the interpretation that it is the advent of generative AI that drives these shifts.

Moreover, Table 7 examines where the decline in citation inequality is most pronounced. We split the sample into three citation-based groups: low (1–10 citations), middle (11–20), and high (21+). Low citation group constitutes 81.9% of our sample, whereas middle and high citation groups constitute 9% and 9.1%. The results show that the reduction in inequality is concentrated at the bottom of the citation distribution. Among papers with 1–10 citations, the Gini coefficient of citation falls significantly post-ChatGPT. In contrast, changes among highly cited papers are small and not statistically significant.

Our result reveals a significant decline in the Gini coefficient for both writing quality and citations, suggesting a flattening of disparities across authors and paper impacts. However, it is unclear to what extent the reduction in inequality of citations is driven by improved writing quality enabled by AI assistance. To investigate this, we conduct a mediation analysis

(Imai et al., 2010; VanderWeele, 2015; Pearl and Mackenzie, 2018) with the Gini coefficient of citations as the outcome and the Gini coefficient of writing quality as the mediator (see Figure 1). Within the mediation framework, we conducted the following regression analysis:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 M_i + \beta_2 T_i + \beta_3 M_i \times T_i + \epsilon_i, \\ M_i &= \alpha_0 + \alpha_1 T_i + \zeta_i. \end{aligned} \tag{2}$$

In the above model, Y_i is the outcome variable (i.e., Gini Coef. of Citations), M_i is the mediator (i.e, Gini Coef. of Writing Quality), and T_i is the treatment (i.e., availability of ChatGPT). From the mediation model, we can calculate the *direct effect* as the average direct effect of the treatment on the outcome when the mediator is held at its value associated with being untreated, which is the difference $Y[1, M(0)] - Y[0, M(0)]$. The *indirect effect* is calculated as the average indirect effect of the treatment on the outcome through the mediator, which is the difference $Y[1, M(1)] - Y[1, M(0)]$. The results in Table 8 indicate that the majority of the reduction in citation inequality is mediated through writing quality, captured by the indirect effect. Moreover, the significant indirect effect alongside an insignificant direct effect suggests a case of full mediation. In other words, the equalizing impact of AI on citations is entirely through improved writing quality, rather than through other channels. This implies that AI tools are leveling the academic playing field not by enhancing visibility or research quality, but primarily by improving the writing quality, thereby increasing accessibility and impact across a broader spectrum of researchers.

3.5 Discussion

How might ChatGPT affect citations beyond just improving writing quality? One possibility is that it helps researchers improve the quality of their work, for example, by making it

easier to organize ideas or code more efficiently. Since better research often goes hand in hand with better writing, it's hard to fully separate the two. Another potential channel is that ChatGPT makes it easier to find and cite relevant literature, especially papers edited by ChatGPT, which could increase a paper's visibility in academic networks.

To tease apart these possibilities, we zoom in on a subsample of papers written by researchers from top global universities. The idea is that within this group, the overall research quality, such as the novelty of ideas and rigor of execution, is already very high. If ChatGPT were mainly boosting research quality, we would expect its impact to be fairly consistent across all authors in this group. If we still find differences, for example, based on whether authors have English or non-English last names, it's more likely that writing quality is the driving factor. Also, researchers at top universities tend to have strong reputations and wide-reaching networks, so their work likely gets attention regardless of writing assistance.

Using the U.S. News Global Universities ranking, we select papers that include at least one author from a top 50 institution. We replicate the analyses presented in Tables 3 and 4 using this subsample. The results, shown in Tables 9 and 10, reveal that within this subsample, the writing quality of authors with non-English last names continues to improve significantly relative to that of authors with English last names. More importantly, these authors also experience an increase in citations and academic impact. These findings provide suggestive evidence that generative AI helps level the playing field in academic writing by enabling high-quality research to overcome language barriers and achieve academic impact.

4 Conclusion

This study offers a fresh perspective on the importance of writing quality in shaping scholarly impact and how generative AI tools are transforming such relationship.

By combining large-scale textual data analysis with citation information, we show that enhanced writing quality is strongly associated with higher citations. Our approach adopts a holistic perspective on writing quality, integrating both traditional linguistic metrics and evaluations generated by large language models that closely align with human judgments. To the best of our knowledge, this paper is among the very first to quantify this relationship at scale across core social science fields.

Moreover, we show that generative AI, particularly ChatGPT, is not just improving writing quality but also leveling the academic impact landscape. The benefits are most pronounced for scholars who have faced linguistic barriers or who have lower visibility: those from non-English-speaking institutions, with non-English names, or with fewer prior citations. For these groups, improved writing quality leads to greater gains in citations. As a result, the introduction of ChatGPT is associated with a notable reduction in inequality for both scholarly exposition quality and recognition.

These findings speak to more than a technological shift—they signal a paradigm change in how academic knowledge is produced, refined, and circulated. Generative AI is not merely a writing aid; it has the potential to reduce frictions and biases due to paper exposition. By reducing linguistic and rhetorical frictions, tools like ChatGPT enhance how we communicate ideas and level the playing field by allowing papers to compete more on their intellectual merit. These changes have profound implications for academic gatekeeping, visibility, and inclusion.

Looking forward, this research raises important questions for future work. How will peer

review evolve in an era of AI-assisted manuscripts? What new norms or standards will govern disclosure, attribution, and authorship? Can generative AI continue to reduce long-standing disparities in academic participation, or will new inequalities emerge? As these tools become embedded in the academic workflow, we must understand not only how they change writing, but whose voices they elevate and whose they risk leaving behind.

Apart from exposition quality, there are of course other dimensions for which AI might have a significant impact on research, for example, increased productivity due to the ability of AI to help with writing code, or AI's ability to quickly summarize papers. Analysis of the impact on those dimensions is left for future work.

References

Banerjee, Onkur, Sean Martin, and David M. Schnyer, 2023, ChatGPT: Applications, opportunities, and threats, *Nature Human Behaviour* 7, 710–715.

Crossley, Scott A, Kristopher Kyle, and Danielle S McNamara, 2016, The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion, *Behavior research methods* 48, 1227–1237.

Flesch, Rudolph, 1948, A new readability yardstick., *Journal of applied psychology* 32, 221.

Flowerdew, John, 2008, Scholarly writers who use English as an additional language: What can Goffman's "Stigma" tell us?, *Journal of English for academic purposes* 7, 77–86.

Gunning, Robert, 1952, *The Technique of Clear Writing* (McGraw-Hill).

Hyland, Ken, 2016, *Academic publishing: Issues and challenges in the construction of knowledge* (Oxford University Press).

Imai, Kosuke, Luke Keele, and Dustin Tingley, 2010, A general approach to causal mediation analysis., *Psychological methods* 15, 309.

Kincaid, J Peter, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom, 1975, Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel, Institute for Simulation and Training, University of Central Florida.

McCarthy, Philip M, and Scott Jarvis, 2010, MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment, *Behavior research methods* 42, 381–392.

McNamara, Danielle S, Eileen Kintsch, Nancy Butler Songer, and Walter Kintsch, 1996, Are good texts always better? interactions of text coherence, background knowledge, and levels of understanding in learning from text, *Cognition and instruction* 14, 1–43.

McNamara, Danielle S, and Walter Kintsch, 1996, Learning from texts: Effects of prior knowledge and text coherence, *Discourse processes* 22, 247–288.

Noy, Shakked, and Whitney Zhang, 2023, Experimental evidence on the productivity effects of generative artificial intelligence, *Science* 381, 187–192.

Pearl, Judea, and Dana Mackenzie, 2018, *The book of why: the new science of cause and effect* (Basic books).

Plavén-Sigray, Pontus, Granville James Matheson, Björn Christian Schiffler, and William Hedley Thompson, 2017, The readability of scientific texts is decreasing over time, *Elife* 6, e27725.

Smith, Edgar A, and RJ Senter, 1967, *Automated readability index*, volume 66 (Aerospace Medical Research Laboratories, Aerospace Medical Division, Air Force Systems Command).

VanderWeele, Tyler, 2015, *Explanation in causal inference: methods for mediation and interaction* (Oxford University Press).

Warren, Nooshin L, Matthew Farmer, Tianyu Gu, and Caleb Warren, 2021, Marketing ideas: How to write research articles that readers understand and cite, *Journal of Marketing* 85, 42–57.

Table 1
Citations and Writing Quality

The table shows the relation between paper citations and writing quality. The dependent variable is the natural logarithm of cumulative citations up to three years. The independent variables are various writing quality measures and field-by-year fixed effects. *NonGPT-Readability* is the average readability score from non-GPT-based measures. All independent variables are standardized. Standard errors are clustered at the field and year levels.

	Log(Citation)						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
ReadScore	0.00689**						
	(2.27)						
Lexical Diversity		0.0634***					
		(13.91)					
Lexical Overlap			0.0571***				
			(13.08)				
Semantic Overlap				0.0926***			
				(15.80)			
NonGPT-Readability					0.0859***		0.0452***
					(16.78)		(10.96)
GPT-Readability						0.155***	0.144***
						(22.29)	(22.51)
Field-by-WrittenYear FE	Y	Y	Y	Y	Y	Y	Y
Adjusted R-squared	0.027	0.031	0.030	0.034	0.034	0.049	0.051
Observations	124,734	126,242	126,242	126,242	124,734	126,237	124,729

t statistics in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 2
Writing Quality and AI

The table shows the increase in writing quality after the introduction of ChatGPT. The dependent variables are various writing quality measures. The independent variables are two dummy: *Grammarly* is equal to one if the paper is written between 2020 and 2022, and zero otherwise; *PostGPT* is equal to one if the paper is written after 2022, and zero otherwise. All specifications include field fixed effects. Standard errors are clustered at the field and year levels.

	(1) ReadScore	(2) Lexical Diversity	(3) Lexical Overlap	(4) Semantic Overlap	(5) NonGPT-Readability	(6) GPT-Readability
Grammarly[20-22]	-0.0605**	2.134***	0.588***	0.00419	0.00543***	0.0646***
	(-2.33)	(12.75)	(3.61)	(0.91)	(2.71)	(3.53)
PostGPT	0.0739	4.493***	1.494***	0.0352***	0.0215***	0.236***
	(1.67)	(8.89)	(7.12)	(6.97)	(9.23)	(10.86)
Field-by-WrittenYear FE	Y	Y	Y	Y	Y	Y
Adjusted R-squared	0.008	0.016	0.003	0.017	0.005	0.017
Observations	125,278	126,807	126,807	126,807	125,278	126,802

t statistics in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 3
Heterogeneous Effect of AI on Writing Quality

The table shows the heterogeneous effect of AI on writing quality. The dependent variable is the GPT-based readability measure. We consider three heterogeneous groups based on the location of authors' affiliations, authors' ethnicity, and authors' past citations. The dummy variable *EnglishInst* is equal to one if at least one author is affiliated with an English-speaking institution. The dummy variable *EnglishName* is equal to one if at least one author has an English last name. The variable *Log(Max Team Cites)* is the natural logarithm of the cumulative citation of the highest cited author in the team, up to the year before the paper was written. We then interact the three variables with both *Grammarly* and *PostGPT* dummies. All specifications include field-by-year fixed effects. Standard errors are clustered at the field and year levels.

	GPT-Readability			
	(1)	(2)	(3)	(4)
EnglishInst	0.378***			0.298***
	(19.49)			(19.36)
EnglishInst × Grammarly[20-22]	0.0340			-0.00281
	(0.83)			(-0.09)
EnglishInst × PostGPT	-0.166***			-0.141***
	(-7.43)			(-6.84)
EnglishName		0.148***		0.0800***
		(12.77)		(8.50)
EnglishName × Grammarly[20-22]		0.0189		-0.00143
		(0.92)		(-0.09)
EnglishName × PostGPT		-0.0687***		-0.0409**
		(-3.80)		(-2.47)
Log(Max Team Cites)			0.0541***	0.0392***
			(16.99)	(15.35)
Log(Max Team Cites) × Grammarly[20-22]			0.00795	0.00609
			(1.30)	(1.27)
Log(Max Team Cites) × PostGPT			-0.0260***	-0.0196***
			(-5.36)	(-3.80)
Field-by-WrittenYear FE	Y	Y	Y	Y
Adjusted R-squared	0.087	0.031	0.083	0.121
Observations	126,802	126,802	126,802	126,802

t statistics in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 4
Heterogeneous Effect of AI on Citations

The table shows the heterogeneous effect of AI on citations. The dependent variable is the natural logarithm of cumulative citations up to three years. The independent variables include *EnglishInst*, *EnglishName*, and *Log(Max Team Cites)*, and their interactions with *Grammarly* and *PostGPT* dummies. All specifications include field-by-year fixed effects. Standard errors are clustered at the field and year levels.

	Log(Citation)			
	(1)	(2)	(3)	(4)
EnglishInst	0.369***			0.109***
	(26.55)			(10.03)
EnglishInst × Grammarly[20-22]	-0.0217			0.00857
	(-0.64)			(0.36)
EnglishInst × PostGPT	-0.177***			-0.0923***
	(-7.27)			(-3.34)
EnglishName		0.243***		0.0939***
		(16.59)		(7.79)
EnglishName × Grammarly[20-22]		-0.0335		-0.0131
		(-1.29)		(-0.61)
EnglishName × PostGPT		-0.105***		-0.0509***
		(-4.57)		(-3.23)
Log(Max Team Cites)			0.142***	0.135***
			(30.07)	(26.89)
Log(Max Team Cites) × Grammarly[20-22]			-0.0417***	-0.0428***
			(-5.81)	(-6.30)
Log(Max Team Cites) × PostGPT			-0.0755***	-0.0708***
			(-12.35)	(-11.32)
Field-by-WrittenYear FE	Y	Y	Y	Y
Adjusted R-squared	0.044	0.033	0.126	0.128
Observations	126,242	126,242	126,242	126,242

t statistics in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 5
Gini Coefficients of Writing Quality

The table shows the reduction in Gini Coefficients of GPT-based and non-GPT-based readability after the introduction of ChatGPT. For each field and each year, we compute the Gini Coefficients of readability. In column (1), we then regress it on *Grammarly* and *PostGPT* dummies. In columns (2) to (5), we estimate the regressions separately for each field.

Panel A: Gini Coef. of GPT-Readability

	Gini Coef. of GPT-Readability				
	(1)	(2)	(3)	(4)	(5)
Grammarly[20-22]	0.00125	0.000580	-0.000702	0.000642	0.00448
	(0.78)	(0.34)	(-0.49)	(0.14)	(1.52)
PostGPT	-0.00859***	-0.00606**	-0.00801***	-0.0133**	-0.00701***
	(-5.99)	(-2.52)	(-3.74)	(-3.01)	(-3.97)
Constant	0.0390***	0.0379***	0.0372***	0.0422***	0.0386***
	(66.21)	(39.04)	(77.91)	(28.06)	(34.22)
Sample	all fields	economics	financial_economics	management	political_science
Adjusted R-squared	0.352	0.288	0.654	0.372	0.411
Observations	60	15	15	15	15

Panel B: Gini Coef. of NonGPT-Readability

	Gini Coef. of NonGPT-Readability				
	(1)	(2)	(3)	(4)	(5)
Grammarly[20-22]	-0.00438	-0.00404	-0.00738**	-0.00307	-0.00303
	(-1.07)	(-1.75)	(-2.47)	(-0.58)	(-0.42)
PostGPT	-0.0227***	-0.0171***	-0.0198***	-0.0265***	-0.0272***
	(-6.89)	(-11.23)	(-3.29)	(-5.95)	(-9.30)
Constant	0.114***	0.109***	0.103***	0.115***	0.127***
	(71.64)	(197.56)	(98.55)	(59.68)	(83.41)
Sample	all fields	economics	financial_economics	management	political_science
Adjusted R-squared	0.331	0.860	0.685	0.641	0.638
Observations	60	15	15	15	15

t statistics in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 6
Gini Coefficients of Citations

The table shows the reduction in Gini Coefficients of citations after the introduction of ChatGPT. For each field and each year, we compute two measures of Gini Coefficients, based either on cumulative citations or the natural logarithm of cumulative citations up to three years. In column (1), we then regress it on *Grammarly* and *PostGPT* dummies. In columns (2) to (5), we estimate the regressions separately for each field.

Panel A: Gini Coefficients of Citation					
	Gini Coef. of Citation				
	(1)	(2)	(3)	(4)	(5)
Grammarly[20-22]	0.0240*	0.0461*	0.0293*	-0.00140	0.0219
	(1.94)	(1.95)	(1.89)	(-0.08)	(0.58)
PostGPT	-0.0911***	-0.0991***	-0.0936**	-0.124***	-0.0480
	(-5.78)	(-13.75)	(-2.60)	(-3.06)	(-1.60)
Constant	0.672***	0.661***	0.670***	0.700***	0.656***
	(173.84)	(148.34)	(99.63)	(91.49)	(162.82)
Sample	all fields	economics	financial_economics	management	political_science
Adjusted R-squared	0.541	0.792	0.630	0.644	0.175
Observations	60	15	15	15	15
Panel B: Gini Coefficients of Log(Citation)					
	Gini Coef. of Log(Citation)				
	(1)	(2)	(3)	(4)	(5)
Grammarly[20-22]	0.00475	0.00640	0.00580	0.00159	0.00521
	(1.52)	(1.77)	(1.30)	(0.28)	(0.58)
PostGPT	-0.0407***	-0.0368***	-0.0432**	-0.0526***	-0.0300***
	(-6.99)	(-9.51)	(-2.35)	(-3.58)	(-11.57)
Constant	0.312***	0.306***	0.311***	0.321***	0.309***
	(247.27)	(312.28)	(180.40)	(109.77)	(209.68)
Sample	all fields	economics	financial_economics	management	political_science
Adjusted R-squared	0.680	0.914	0.662	0.723	0.648
Observations	60	15	15	15	15

t statistics in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 7
Conditional Gini Coefficients of Citations

The table shows that the reduction in Gini Coefficients of citations comes from papers with fewer than 20 citations. We split the sample into three groups: papers with citations between 1 and 10 (81.9% of sample), between 11 and 20 (9%), and over 20 (9.1%). For each field, year, and citation group, we compute two measures of Gini Coefficients, based either on cumulative citations or the natural logarithm of cumulative citations up to three years. We then regress the Gini Coefficients on *Grammarly* and *PostGPT* dummies across all fields.

	$1 \leq Citation \leq 10$		$11 \leq Citation \leq 20$		$Citation > 20$	
	(1)	(2)	(3)	(4)	(5)	(6)
	Gini(Citation)	Gini(Log-Citation)	Gini(Citation)	Gini(Log-Citation)	Gini(Citation)	Gini(Log-Citation)
Grammarly[20-22]	0.00165	0.000154	-0.000124	-0.0000634	0.0497**	0.00526
	(0.96)	(0.16)	(-0.12)	(-0.19)	(2.07)	(1.58)
PostGPT	-0.0208***	-0.0154***	-0.00507*	-0.00159*	-0.0104	-0.00114
	(-3.40)	(-4.62)	(-1.86)	(-1.92)	(-0.32)	(-0.27)
Constant	0.407***	0.235***	0.110***	0.0374***	0.366***	0.0793***
	(456.60)	(521.83)	(165.63)	(164.88)	(54.31)	(79.40)
Adjusted R-squared	0.437	0.595	0.101	0.090	0.082	0.037
Observations	60	60	60	60	60	60

t statistics in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 8
Mediation Analysis

The table shows the mediation analysis. For each field and year, we compute two measures of Gini Coefficients, based either on cumulative citations or the natural logarithm of cumulative citations up to three years. In the mediation analysis, the outcome variable is the Gini Coefficients of citations and the mediator is the Gini Coefficients of GPT-Readability.

Outcome Variable	Mediator	Indirect Effect	Direct Effect	Proportion mediated
Gini Coef. of Citation	Gini Coef. of GPT-Readability	-0.0797***	-0.00989	0.890***
		(-3.09)	(-0.33)	(2.74)
Gini Coef. of Log(Citation)	Gini Coef. of GPT-Readability	-0.0313***	-0.00874	0.782***
		(-3.46)	(-1.14)	(4.10)
Gini Coef. of Log(Citation)	Gini Coef. of nonGPT-Readability	-0.0892***	-0.00534	0.943***
		(-4.55)	(-0.23)	(3.90)
Gini Coef. of Log(Citation)	Gini Coef. of nonGPT-Readability	-0.0349***	-0.00702	0.833***
		(-3.27)	(-0.88)	(4.27)

t statistics in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 9
Heterogeneous Effect of AI on Writing Quality - Top 50 Global Universities

The table shows the heterogeneous effect of AI on writing quality, conditional on a subsample where at least one author is from one of the top 50 global universities based on US News ranking. The dependent variable is the GPT-based readability measure. We consider three heterogeneous groups based on the location of authors' affiliations, authors' ethnicity, and authors' past citations. The dummy variable *EnglishInst* is equal to one if at least one author is affiliated with an English-speaking institution. The dummy variable *EnglishName* is equal to one if at least one author has an English last name. The variable *Log(Max Team Cites)* is the natural logarithm of the cumulative citation of the highest cited author in the team, up to the year before the paper was written. We then interact the three variables with both *Grammarly* and *PostGPT* dummies. All specifications include field-by-year fixed effects. Standard errors are clustered at the field and year levels.

	GPT-Readability			
	(1)	(2)	(3)	(4)
EnglishInst	0.140***			0.114***
	(7.11)			(5.89)
EnglishInst × Grammarly[20-22]	-0.0199			-0.0460
	(-0.60)			(-1.37)
EnglishInst × PostGPT	0.0585			0.0526
	(0.78)			(0.69)
EnglishName		0.0564***		0.0373***
		(6.76)		(4.88)
EnglishName × Grammarly[20-22]		0.0136		0.00212
		(0.95)		(0.14)
EnglishName × PostGPT		-0.0269*		-0.0288*
		(-1.84)		(-1.93)
Log(Max Team Cites)			0.0180***	0.0169***
			(11.03)	(10.68)
Log(Max Team Cites) × Grammarly[20-22]			0.00687*	0.00664
			(1.83)	(1.66)
Log(Max Team Cites) × PostGPT			-0.00396	-0.00431
			(-1.12)	(-1.11)
Field-by-WrittenYear FE	Y	Y	Y	Y
Adjusted R-squared	0.028	0.028	0.036	0.038
Observations	71,998	71,998	71,998	71,998

t statistics in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 10
Heterogeneous Effect of AI on Citations - Top 50 Global Universities

The table shows the heterogeneous effect of AI on citations, conditional on a subsample where at least one author is from one of the top 50 global universities based on US News ranking. The dependent variable is the natural logarithm of cumulative citations up to three years. The independent variables include *EnglishInst*, *EnglishName*, and *Log(Max Team Cites)*, and their interactions with *Grammarly* and *PostGPT* dummies. All specifications include field-by-year fixed effects. Standard errors are clustered at the field and year levels.

	Log(Citation)			
	(1)	(2)	(3)	(4)
EnglishInst	0.368***			0.173***
	(11.18)			(5.69)
EnglishInst × Grammarly[20-22]	-0.0571			-0.0604
	(-0.79)			(-0.84)
EnglishInst × PostGPT	-0.142**			-0.131*
	(-2.07)			(-1.83)
EnglishName		0.265***		0.123***
		(16.77)		(7.96)
EnglishName × Grammarly[20-22]		-0.0344		-0.00858
		(-1.01)		(-0.29)
EnglishName × PostGPT		-0.0935***		-0.0512**
		(-3.42)		(-2.14)
Log(Max Team Cites)			0.137***	0.134***
			(26.06)	(25.10)
Log(Max Team Cites) × Grammarly[20-22]			-0.0408***	-0.0412***
			(-5.06)	(-5.32)
Log(Max Team Cites) × PostGPT			-0.0639***	-0.0636***
			(-8.62)	(-8.58)
Field-by-WrittenYear FE	Y	Y	Y	Y
Adjusted R-squared	0.038	0.042	0.108	0.110
Observations	71,781	71,781	71,781	71,781

t statistics in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Figure 1
Mediation Framework
This figure illustrates the framework of mediation analysis conducted in Table 8 following Equation (2). In this framework, Gini Coef. of Citations, Y_i, is the outcome variable; Gini Coef. of Writing Quality, M_i, is the mediator; and $PostGPT$, T_i, is the treatment. From the mediation model, we can calculate the Direct Effect as the average direct effect of the treatment on the outcome when the mediator is held at its value associated with being untreated, which is the difference $Y[1, M(0)] - Y[0, M(0)]$. The Indirect Effect is calculated as the average indirect effect of the treatment on the outcome through the mediator, which is the difference $Y[1, M(1)] - Y[1, M(0)]$.

