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

DISCUSSION PAPER SERIES

IZA DP No. 18215

Bridging Language Barriers: The Impact of Large Language Models on Academic Writing

Burak Dalaman
Ali Furkan Kalay
Nathan Kettlewell

OCTOBER 2025

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 18215

Bridging Language Barriers: The Impact of Large Language Models on Academic Writing

Burak Dalaman
University of London

Ali Furkan Kalay
Macquarie University Centre for Health Economy

Nathan Kettlewell
University of Technology Sydney and IZA

OCTOBER 2025

ISSN: 2365-9793

IZA – Institute of Labor Economics

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

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

www.iza.org

ABSTRACT

Bridging Language Barriers: The Impact of Large Language Models on Academic Writing

Large language models (LLMs) have altered the nature of academic writing. While the influence of LLMs on academic writing is not uncontroversial, one promise for this technology is to bridge language barriers faced by nonnative English-speaking researchers. This study empirically demonstrates that LLMs have led to convergence in the lexical diversity of native and nonnative speakers, potentially helping to level the playing field. There has also been an increase in language complexity for nonnatives. We classify over one million authors as native or nonnative English speakers based on the etymological origins of their names and analyze over one million abstracts from arXiv.org, evaluating changes in lexical diversity and readability before and after ChatGPT's release in November 2022. The results demonstrate a sharp increase in writing sophistication among all researchers, with nonnative English speakers showing the greatest gains across all writing metrics. Our findings provide empirical evidence on the impact of LLMs in academic writing, supporting recent speculations about their potential to bridge language barriers.

JEL Classification: J24, I23

Keywords: language barrier, generative AI, academic equity, large language models, technology adoption, bayesian structural time series

Corresponding author:
Ali Furkan Kalay
Macquarie University
4–6 Eastern Road, NSW 2109
Australia

E-mail: alifurkan.kalay@mq.edu.au

1 Introduction

Since ChatGPT was launched on 30 November 2022, it and other generative AI large language models (LLMs) have had a profound effect on the nature of work. For academics, one of their primary uses is text editing and generation, which has spurned debate around the promise and pitfalls of using AI in academic writing (Hwang et al., 2023; Lingard et al., 2023). Those concerned worry that uncritical usage may, for example, lead to inaccurate reporting, less credibility and less lexical diversity across articles (Al-Zaabi et al., 2023). However, one hope for LLMs is their potential to improve equity by helping authors who struggle with writing, in particular nonnative English speakers (Berdejo-Espinola & Amano, 2023; Van Noorden & Perkel, 2023). 97% of academic research is published in English (Liu, 2017), which acts as a powerful barrier to productivity for nonnative speakers (Amano et al., 2016; Hanauer et al., 2019; Ramírez-Castañeda, 2020). Indeed, good language editing improves the perceived quality of academic research, increasing the likelihood that authors are accepted into conferences and see their research published in high-impact journals (Feld et al., 2024). This raises the question: how has the adoption of LLMs affected the writing styles of authors from native English-speaking and nonnative English-speaking backgrounds? Are we seeing "writing convergence"? Our paper is, to our knowledge, the first to address these questions.

The embrace of LLMs has been remarkable. After its launch, ChatGPT quickly became the fastest growing consumer app in history (Reuters, 2023). Today, it boasts over 180 million users (Duarte, 2024) whilst also competing against several other high-profile LLMs such as Google's Gemini and Anthropic's Claude. A growing body of empirical evidence demonstrates the widespread usage of LLMs on academic writing (Bisi et al., 2023; Geng et al., 2024; Kobak et al., 2024; Liang, Zhang, et al., 2024; Liang, Izzo, et al., 2024; Uribe & Maldupa, 2024). At the extreme end, AI has been used to co-author research articles (Stokel-Walker, 2023). More typically, it has served as a free tool to help authors summarize their research and edit text, offering an accessible alternative to commercialized products like Grammarly.

To further motivate our paper, following the approach in Kobak et al. (2024), we demonstrate the impact of LLMs using our own data (abstracts scraped from arXiv) in Figure 1. Looking at all abstracts, there is a clear trend break in excessive vocabulary counts (an indicator for potential LLM use) at the launch date of ChatGPT. This result corroborates similar evidence presented by Geng et al. (2024), Kobak et al. (2024), Liang, Zhang, et al. (2024) and Liang, Izzo, et al. (2024). Interestingly though, once we disaggregate authorship groups by English speaking background (details on how we achieve this are in Section 3), we see the effect is much stronger for articles authored by nonnative English-speaking researchers.

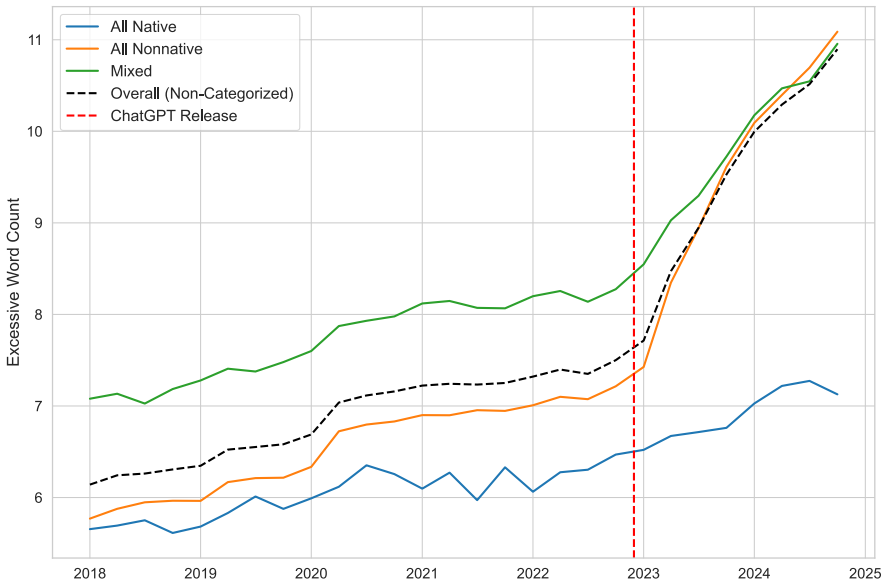

Figure 1: Quarterly Excessive Vocabulary Counts in arxiv.org Abstracts

Notes: This figure shows the quarterly average excessive word counts extracted from abstracts of articles published on arxiv.org from January 1, 2018, to November 14, 2024. "All Native" refers to articles written by authors who are all native English speakers; "All Nonnative" includes articles written solely by non-native English speakers; "Mixed" represents articles with a combination of both native and non-native authors. "Overall (Non-Categorized)" reflects data from all articles, regardless of authorship. The red dashed line indicates the release date of ChatGPT on 30 November 2022. These numbers are computed based on excessive word dictionary created by Kobak et al. (2024).

Deficiencies with LLMs are well known, such as their tendency to hallucinate and generate non-existing references, and reliance on formal rationality leading to biased outputs (Nishant et al., 2024). Nonetheless, the potential to enhance productivity is clear, though evidence on LLMs' impact (and generative AI more broadly) is only just emerging. In the medical field, for example, generative AI has been used for administrative tasks, while its potential for improving diagnosis and treatment remains an active area of investigation (Thirunavukarasu et al., 2023). Kreitmeir & Raschky (2024) find that Italy's ChatGPT ban affected the quality and quantity of software and code produced by GitHub users, although its effects were heterogeneous, demonstrating that AI needs to be mindfully applied to be productivity enhancing. For academics, LLMs can assist with many tasks, including programming and data analysis, teaching, administrative work and (the focus of this paper) academic writing. For nonnative English speakers, assistance with academic writing is likely to be particularly important. Berdejo-Espinola & Amano (2023) argue that AI has the potential to address the long-standing disadvantage faced by this group, who need to invest more energy into writing, are less successful in research reviews, and face severe financial costs associated with text editing in order to compete. Homogeny in academic writing should, in principle, lead to a more meritocratic evaluation of research, ultimately benefiting the users of research – society at large.

We provide evidence on the impact of LLMs on academic writing by comparing abstracts published on arXiv before and after the release of ChatGPT. Our analysis focuses on

two important dimensions of writing: lexical diversity (related to the frequency of unique words) and lexical complexity (readability). Our own analysis serves as an example of the potential of LLMs to shape academic research – we leverage an LLM in a novel way to classify authors as native or nonnative English speakers based on the etymological information embedded in their names. Manual verification of our approach finds that it is highly accurate. Articles are divided into three categories of authorship: all native English-speaking, a mixture of native and nonnative English-speaking, and all nonnative English-speaking. By using arXiv, we ensure that we capture the latest research, before it has been through peer-review and journal editing. arXiv is also desirable because it covers a wide range of academic fields. Our dataset includes 1,250,890 articles with 1,043,289 unique authors over the period from 1 January 2018 to 14 November 2024.

To accurately forecast how academic writing would have evolved after November 2022 had ChatGPT (and other LLMs) not emerged, we estimate a Bayesian structural time series (BSTS) model. This approach is advantageous as it allows for the decomposition of time series data into its components and provides robust estimates of how the gradual introduction of the intervention has impacted the writing trends.

Our main findings are as follows. First, we find that the introduction of ChatGPT is associated with an increase in lexical diversity and lexical complexity for all authorship groups, and this effect increases over time (mirroring results in Figure 1 on the detection of LLM usage). Second, for both measures, we find nonnative English-speaking researchers exhibit the most pronounced improvements in both metrics, with lexical diversity (Type-Token Ratio) among all three authorship groups converging by November 2024, and lexical complexity (Automated Readability Index) nearing similar levels. These trends suggest that LLMs are helping to bridge language barriers in academic writing. Lastly, we observe that the adoption of LLMs in academic writing is still ongoing. Our analysis reveals steadily increasing trends in the selected writing metrics, indicating its growing integration and influence. Notably, we find computer scientists adopt LLMs more rapidly than researchers in other fields.

Our paper is organized as follows. Section 2 describes the data. Section 3 details the selected metrics, explains how we classify authors' native language, and outlines our estimation strategy. Section 4 presents and discusses the results, including robustness checks. Section 5 considers the long-term impacts and broader implications of LLMs. Section 6 concludes the paper.

2 Data

We obtained articles published at arXiv.org starting from 2018. arXiv is an extensive open-access repository containing pre-print research articles (working papers) from multiple disciplines (physics, mathematics, computer science, quantitative biology, quantitative finance, statistics, electrical engineering and systems science, and economics). arXiv is well-suited to evaluating the impact of LLMs for several reasons. First, pre-print writing is generally less polished, making it more tractable for analyzing the extent of AI-assisted writing. Second, pre-prints capture the very latest research, while the writing in journal publications is difficult to date due to slow and variable review processes. Third, arXiv is one of the largest pre-print databases, containing articles across various fields, which allows for a broad evaluation of LLMs' impact on academic writing.

Our dataset covers 1,250,890 articles published between 1 January 2018 and 14 November 2024, with 1,043,289 unique authors. This window offers several years of pre-intervention data to capture underlying trends. For each article, we collect extensive metadata including author names, titles, abstracts, field categories and publication dates (initial submission). We obtained the data from kaggle.com, which hosts bulk downloads officially provided by arXiv.org (arXiv.org Submitters, 2024).

Our text analysis only considers the abstracts of articles.[1] We focus on abstracts since they are often the most refined and polished sections of an article and typically written in a more standardized and uniform format than the main text, both within and across disciplines.[2] In contrast, there is considerable heterogeneity in the structure and content of the main body of articles. For example, comparing a technical article (with equations, algorithms, etc.) to a text-based article is inherently difficult. Consequently, we focus on abstracts as they offer a consistent and practical basis for detecting potential LLM usage.

Additionally, and arguably of greater importance, a poorly written abstract can adversely affect a reader's impression of the quality of an article (Feld et al., 2024). A poorly written abstract may dissuade someone from reading the rest of an article, or prevent an editor from sending an article for peer review. Admission to academic conferences is often based solely on the submission of abstracts. Suffice to say, abstracts are a crucial component of any academic article, and a key component of an academic's production.

[1] arXiv allows authors to update pre-print articles; however, we only use abstracts at the time of first submission to the archive. To ensure the reliability of our writing metrics, we only include abstracts containing more than 25 words in our analysis (7,007 articles have less than 25 words in our sample).

[2] Bisi et al. (2023) found that AI usage in academic writing is more prevalent in abstracts than in the body text.

3 Methodology

3.1 Writing metrics

Our analysis first examines the lexical richness of academic abstracts using the Type-Token Ratio (TTR) (Chotlos, 1944; Templin, 1957), defined as the number of unique words (types) divided by the total number of words (tokens). A higher TTR indicates greater lexical diversity, and thus potentially reflects a higher quality of writing. While TTR is an extensively used measure of lexical diversity, it is not without limitations, particularly its sensitivity to text length (Tweedie & Baayen, 1998). One simple correction for this is to use the natural logarithm of the TTR, sometimes called Herdan's C (Herdan, 1964), which we do as a robustness exercise.

We evaluate the readability of the abstracts using the Automated Readability Index (ARI) (Smith & Senter, 1967), which primarily considers word length and sentence length.[3] The measure is designed to estimate the U.S grade level required to comprehend a piece of text. Higher ARI scores therefore typically imply more complex vocabulary and sentence structures, often indicative of advanced language proficiency. It is important to note, however, that a higher ARI score does not necessarily imply better writing quality. Depending on the context and intended audience, lower ARI scores may signal clarity and conciseness, qualities equally important in effective communication. Nevertheless, given that academic texts are frequently written with sophisticated language, higher ARI scores often suggest the writer's skills for employing more complex linguistic constructions.

Although measuring lexical diversity is relatively straightforward, assessing overall readability is inherently more complex. Various tools are available for evaluating reading quality, some of which are considerably more advanced than the ARI. For instance, commercial platforms like Grammarly and deep learning-based algorithms offer sophisticated features that may yield nuanced insights. However, these tools often rely on proprietary "black box" methodologies, which evolve continuously and lack transparency, potentially hindering replicability of our findings.

We present our findings using TTR and ARI to ensure transparency and replicability in our analysis. This choice allows us to maintain methodological clarity while avoiding the financial and computational burdens associated with more modern methods. Additionally, our scraped arXiv dataset, along with the computed metrics and author classifications, is included in the supplementary materials, allowing researchers to verify and build upon our findings.

[3]The ARI formula is $4.71(\text{characters/words}) + 0.5(\text{words/sentences}) - 21.43$. Some other popular readability indexes, such as the Gunning Fog Index (Gunning, 1952), emphasize syllables rather than word length. We present results using the Gunning Fog Index as a robustness exercise.

3.2 Classification

We employed OpenAI's `gpt-4o-2024-08-06` model to classify authors as either native or nonnative English speakers based on their names, leveraging the etymological information embedded in authors' names. Furthermore, many arXiv authors, particularly senior researchers, are well-known academics whose backgrounds and articles may be recognized by LLMs. Consequently, we anticipate a reasonable degree of accuracy in our author classification.[4] Appendix A elaborates on our classification procedures and prompting.

After classifying authors, each article was assigned to one of three language categories: *All Native* (all authors classified as native English speakers), *All Nonnative* (all authors classified as nonnative English speakers), and *Mixed* (a combination of native and nonnative authors).[5] This classification results in 817,580 articles authored exclusively by nonnative English speakers, 370,877 articles by mixed-language teams, and 62,433 articles authored exclusively by native English speakers. Although the sample distribution is unbalanced, with relatively fewer articles in the *All Native* category, the large overall sample size minimizes concerns about statistical power.

The use of LLMs makes it feasible to classify the language backgrounds of more than one million authors. However, we acknowledge that this may also result in some misclassification errors. To evaluate classification accuracy, we therefore performed a manual validation on a random subset of 100 authors classified by the gpt-4o model – 50 classified as native and 50 as nonnative. For these 100 authors, we manually searched for their details online and generated an alternative classification based on the country they obtained their bachelor's degree.[6]

Among the 50 nonnative authors, five were identified as misclassified. Similarly, for the 50 native authors, six misclassifications were observed. Our manual review suggests an approximate accuracy rate of 90%. However, we contend that the true accuracy is likely higher, as the country where an individual completes their undergraduate degree does not always correspond to their home country. Notably, it is more common for individuals with etymologically nonnative names to relocate to native English-speaking countries for higher education than the reverse (e.g., students from China pursuing undergraduate studies in

[4] See, for example, Huang et al. (2024) and Lamichhane (2023), who document the success of LLMs in more sophisticated classification tasks using models less advanced than those employed in this study.

[5] We perform our analysis at the article level rather than on an individual basis because articles are collaborative efforts. The writing quality of an individual author can vary significantly depending on their co-authors. For instance, native English speakers might be tasked with writing key sections such as the introduction and abstract, which are often polished collaboratively by the team. Categorizing articles instead of individuals helps control for these collaborative dynamics and isolates the impact of linguistic background on writing quality.

[6] For those in academia (the majority) this was usually readily available on personal or faculty websites. In some cases we found this information from other sources such as LinkedIn.

Western countries). For instance, some individuals in our sample with etymologically Asian names were classified as nonnative authors by the LLM but were considered native speakers in our manual checks due to their undergraduate studies in English-speaking countries. This indicates that some of these authors, despite being classified as native in our review, could be a nonnative English speaker.

Attenuation Bias in Disparity: Given the high accuracy of our classification approach, we do not expect misclassification to greatly affect our results. To the extent it does, if the misclassification errors are random they would tend to attenuate differences between our three author groups, and if LLMs have had a greater impact on the writing of nonnative than native authors (as we hypothesize), it would bias upwards the estimates for natives and downwards the estimates for nonnatives. Therefore, measurement error would result in an attenuation bias in the disparity between native and nonnative English-speaking researchers' writing metrics.

3.3 Descriptive statistics

Table 1: Number of Articles by Discipline

	Authorship Groups			
	Native	Nonnative	Mixed	All
Number of Articles				
Computer Science	23,084	430,235	205,842	659,161
Mathematics	24,952	222,088	58,895	305,935
Statistics	6,360	68,478	37,376	112,214
Economics	926	6,917	2,720	10,563
Electrical Engineering	1,633	65,036	26,516	93,185
Physics	7,142	80,442	52,483	140,067
Biology	1,711	12,542	9,631	23,884
Finance	828	8,345	2,885	12,058
Total	**62,433**	**817,580**	**370,877**	**1,250,890**

Notes: This table presents the total number of articles in various disciplines, categorized by authorship groups, based on articles published on arxiv.org in our sample, spanning from January 1, 2018, to November 14, 2024. The authorship categories are defined as follows: "Native" refers to articles authored entirely by native English speakers, "Nonnative" includes articles authored entirely by non-native English speakers, and "Mixed" indicates articles with a combination of native and non-native authors. The "All" column represents the total number of articles across all authorship categories for each discipline. The last row, "Total," provides the overall number of articles, which is not equivalent to the sum of all disciplines because a single article may be labeled with multiple disciplines.

Table 1 shows the number of articles by discipline for each authorship group. The single most represented discipline is computer science, comprising 53% of articles overall. Notably, this representation is higher in nonnative and mixed authorship articles (53% and 56% respectively) than in native authorship articles (37%). While our main analysis pools all articles together, to mitigate concerns that difference between language background are actually just differences between disciplines, we present results only including

Table 2: Lexical Characteristics of Abstracts with Language Categories

	Authorship Groups			
	Native	Nonnative	Mixed	All
Abstract Word Count				
Mean	142.640	154.230	168.775	157.964
25%	93	111	128	115
50%	137	153	165	156
75%	187	195	209	199
Std	63.094	58.761	57.320	59.035
Abstract Character Length				
Mean	965.722	1066.821	1163.952	1090.573
25%	628	767	887	794
50%	929	1063	1150	1084
75%	1269	1357	1446	1381
Std	428.788	406.535	388.224	405.742
Type-Token Ratio Index				
Mean	0.634	0.628	0.630	0.629
25%	0.573	0.574	0.580	0.576
50%	0.630	0.627	0.629	0.628
75%	0.690	0.680	0.679	0.680
Std	0.092	0.084	0.077	0.083
Automated Readability Index				
Mean	17.777	18.002	18.091	18.017
25%	15.400	15.800	16.000	15.800
50%	17.600	17.800	18.000	17.800
75%	19.900	19.900	20.000	20.000
Std	3.713	3.540	3.225	3.459
Herdan's C Index				
Mean	0.905	0.905	0.908	0.906
25%	0.891	0.892	0.896	0.893
50%	0.906	0.907	0.909	0.908
75%	0.921	0.921	0.922	0.921
Std	0.026	0.024	0.021	0.023
Gunning Fog Index				
Mean	16.303	16.150	16.294	16.201
25%	14.320	14.260	14.490	14.330
50%	16.130	16.020	16.200	16.050
75%	18.060	17.880	17.960	17.910
Std	3.047	2.886	2.756	2.857

Notes: This table presents the lexical characteristics of article abstracts, categorized by authorship groups, based on articles published on arxiv.org in our sample, spanning from January 1, 2018, to November 14, 2024. The authorship groups are defined as follows: "Native" refers to abstracts authored entirely by native English speakers, "Nonnative" includes abstracts authored entirely by non-native English speakers, "Mixed" indicates abstracts with contributions from both native and non-native authors, and "All" includes all abstracts irrespective of authorship groups. "Abstract Word Count" is the total number of words in the abstract, while "Abstract Character Length" reflects the overall length in terms of characters, including spaces. Type-Token Ratio (TTR) (Chotlos, 1944; Templin, 1957) measures lexical diversity. Automated Readability Index (ARI) (Smith & Senter, 1967) measures readability based on word and sentence lengths. Herdan's C (log-TTR) (Herdan, 1964) is a correction of TTR that uses the natural logarithm to reduce sensitivity to text length. Gunning Fog Index (Gunning, 1952) is a readability index that emphasizes the use of complex words.

articles in computer science as a robustness exercise.

Table 2 presents summary statistics for key features of abstracts, including readability and complexity indexes. Articles authored by nonnative and mixed authorship groups tend to be slightly longer on average, and with slightly less variance, compared to those written by natives.

3.4 Statistical Analysis

Our statistical methodology is driven by the question: How would the selected writing metrics have evolved had ChatGPT (and other LLMs that followed) not been released? Modeling this counterfactual scenario presents two primary challenges. First, the adoption of ChatGPT and other LLMs into workflows has been gradual, meaning the *treatment* and its effects develop progressively over time, rather than occurring as a sudden event. Second, there is no control group available for direct comparison, which complicates causal inference. These characteristics of the intervention lead us to the use of time series models to forecast the counterfactuals. However, creating accurate long-term forecasts introduces the additional risk of instability in the model, since more than two years have passed since ChatGPT was first released.

To address these concerns, we use a Bayesian structural time series (BSTS) framework, which is well-suited for modeling complex interventions and forecasting counterfactuals (Brodersen et al., 2015). BSTS decomposes the observed outcome into trend, seasonal, and other latent components, allowing for a nuanced representation of the underlying time series dynamics. Moreover, it accommodates gradual treatment effects by providing pointwise impact estimates. To validate our findings, we conduct a placebo test in subsequent sections.

We compute the weekly averages of the selected writing metrics. Formally, let y_t denote the observed outcome at time (week) t. The BSTS model is represented via two core equations. The *observation equation* relates the observed data y_t to a latent d-dimensional state vector α_t:

$$y_t = Z_t^\top \alpha_t + \varepsilon_t, \tag{1}$$

where ε_t is a scalar observation error with variance σ_t^2. The *state equation* describes how the latent state evolves over time:

$$\alpha_{t+1} = T_t \alpha_t + R_t \eta_t, \tag{2}$$

where η_t is a q-dimensional system error with a $q \times q$ state-diffusion matrix Q_t, and R_t is a $d \times q$ control matrix. By expressing the error term as $R_t \eta_t$, the model incorporates reduced-rank state components, which is particularly useful for modeling seasonal effects. In this study, we specify a local linear trend and weekly seasonality in our model (see Brodersen et al., 2015 for further details on latent variable evolution).

We estimate the model parameters and latent states using Bayesian methods, implemented via an `R` package (Brodersen et al., 2017). Specifically, we perform 5,000 iterations of Markov Chain Monte Carlo (MCMC) sampling using a Gibbs sampler to draw from the posterior distribution of the model parameters. After fitting the model to the pre-intervention period, we generate counterfactual forecasts, $\hat{y}_t$, representing how the outcome variable would have evolved had the intervention not occurred. The causal impact is then computed as the difference between the observed outcome and the counterfactual:

$$\widehat{\text{Impact}}_t = y_t - \hat{y}_t. \tag{3}$$

By quantifying the difference over the post-intervention period, we capture both the magnitude and trajectory of ChatGPT's impact on the selected writing metrics. To enhance the reliability of our analysis, we also compute the cumulative impact, as it provides a clearer picture of the overall effect.

The key identification assumption in our model is that, in the absence of the intervention (had ChatGPT not been released), the underlying time series dynamics—such as trends and seasonality—would have remained stable during the post-intervention period. We evaluate the plausibility of this assumption through a placebo test.

4 Results

In this section, we first descriptively investigate the changes in TTR and ARI metrics since 2018 for our three authorship groups: native; nonnative; and mixed-authored articles, and use the BSTS model to formally estimate the impact of LLMs on these metrics. We then conduct additional analyses to validate our findings.

4.1 Type-Token Ratio

Figure 2 depicts the quarterly average TTR for the three authorship groups. For all groups, there is a downward or flat trend prior to the release of ChatGPT, and a clear upward trend following the release of ChatGPT. Before ChatGPT, articles authored by native speakers displayed the highest lexical diversity, while those authored by nonnative speakers showed the lowest. Following the release of ChatGPT, nonnative authored articles experienced the most significant shift, and by November 2024, they had the highest lexical diversity among the three groups. This shift effectively closed the gap between nonnative speakers and the other groups as of November 2024.

Figure 3 presents the BSTS estimation results for the weekly average TTR of articles authored by native speakers. The first panel displays the observed data, model fitted values and forecasts, and the credible intervals for the estimated values. Estimates to the right of the dashed line are forecasts that we interpret as what would have happened had ChatGPT and other LLMs not been released. The second panel shows the estimated pointwise impact and credible intervals, while the bottom panel shows the estimated cumulative impact and credible interval. Despite the descriptive figures suggesting a small positive effect, neither the pointwise nor cumulative analyses provide strong evidence of effects being credibly different from zero.

Figure 4 presents the results for mixed-authored articles. The pointwise impact estimates indicate that, during the past five weeks of the sample (October-November 2024), the average TTR increased by 0.012 for mixed-authored articles, with the estimates generally credibly different from zero. This corresponds to an approximately 0.15 standard deviation increase.

Figure 5 shows the results for nonnative authored articles. The impact is larger and more precise compared to that observed for mixed-authored articles. The average impact over the last five weeks (October and November 2024) corresponds to an approximately 0.2 standard deviation increase.

To provide a closer examination of the most recent period in our sample, Table 3 presents the pointwise impacts for the last five weeks. Overall, the results imply that ChatGPT and other LLMs have significantly improved lexical richness in academic writing, with the most notable benefits observed in articles authored by nonnative speakers. In contrast, no impact was found for articles authored by native speakers.

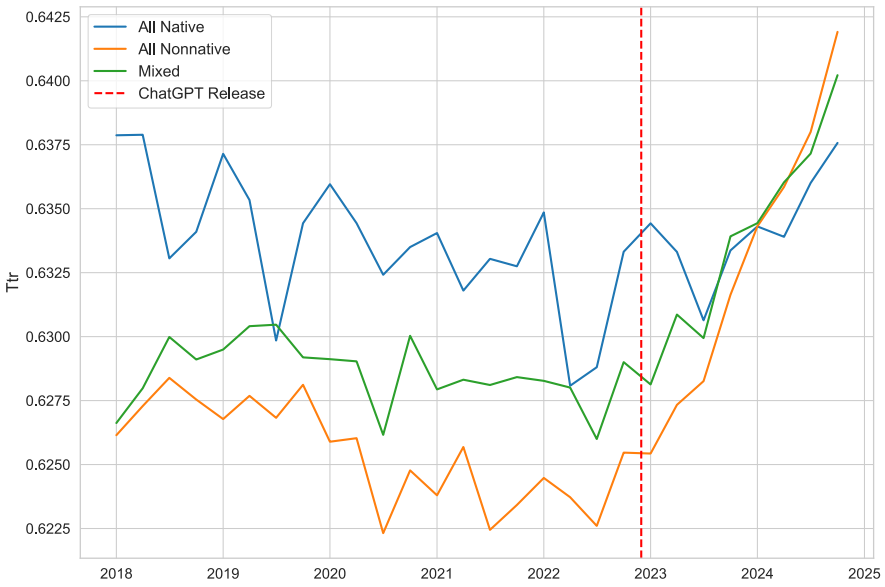

Figure 2: Quarterly Average Type-Token Ratio by Language Categorization

Notes: This figure displays the quarterly average Type-Token Ratio (TTR), as defined by Chotlos (1944) and Templin (1957), calculated from abstracts of articles published on arxiv.org from January 1, 2018, to November 14, 2024. TTR, which measures lexical diversity, is defined as the number of unique words (types) divided by the total number of words (tokens), with higher values indicating greater variety in vocabulary. The articles are categorized into three authorship groups: "All Native" refers to articles written entirely by native English-speaking authors; "All Nonnative" includes articles authored solely by non-native English speakers; "Mixed" represents articles written by a combination of native and non-native authors. The red dashed vertical line marks the release date of ChatGPT on 30 November 2022.

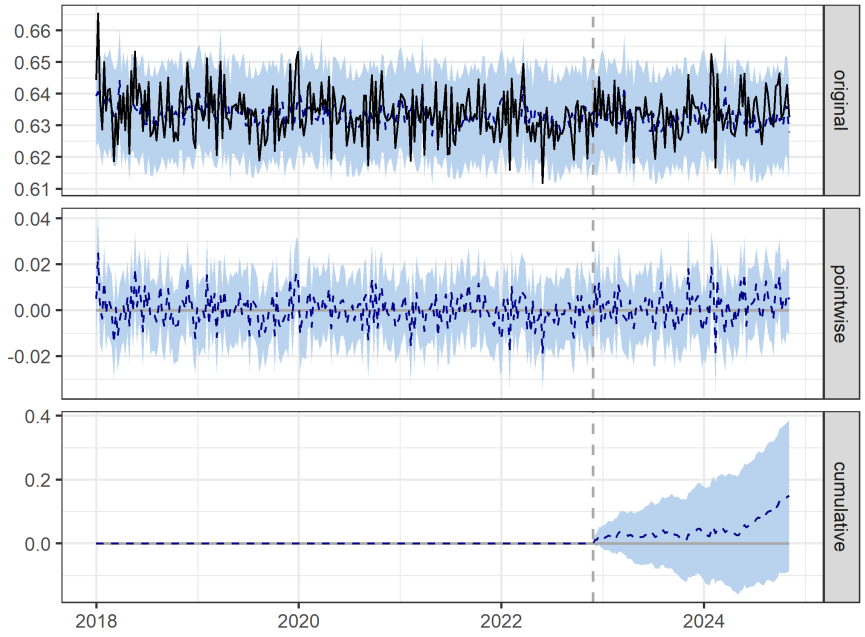

Figure 3: The Impact of ChatGPT on the Type-Token Ratio for Articles Authored by Native English Speakers

Notes: This figure, generated using the `CausalImpact` package (Brodersen et al., 2017), illustrates the effect of ChatGPT's release on the Type-Token Ratio (TTR) for articles authored exclusively by native English speakers, calculated from abstracts of articles published on arxiv.org from January 1, 2018, to November 14, 2024. TTR, which measures lexical diversity, is defined as the number of unique words (types) divided by the total number of words (tokens). The top panel shows the observed data (solid black line) and the counterfactual predictions (blue line), with the shaded region representing the 95% credible intervals. The dashed vertical line marks the release date of ChatGPT, 30 November 2022. The middle panel displays the pointwise impact, calculated as the difference between observed and predicted values. The bottom panel visualizes the cumulative impact over time.

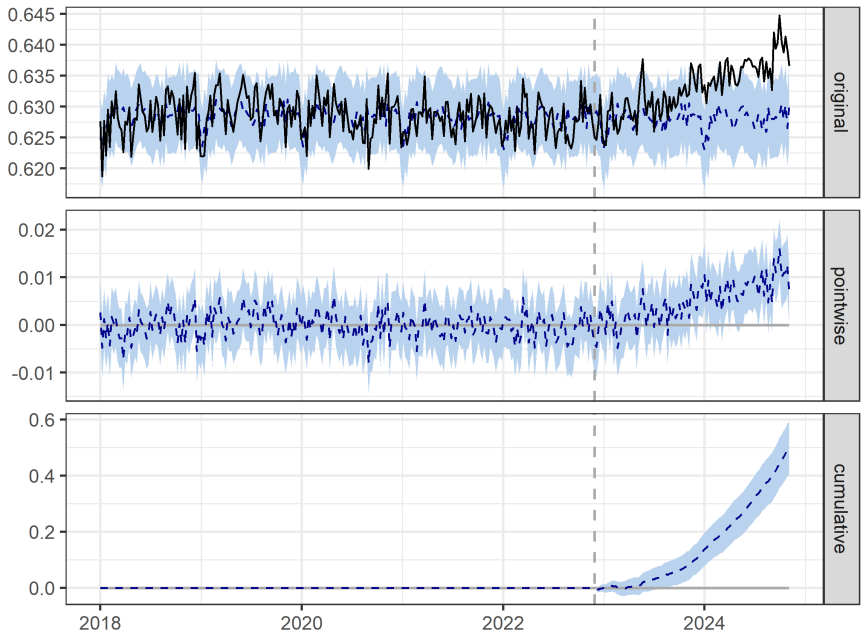

Figure 4: The Impact of ChatGPT on the Type-Token Ratio for Articles Authored by Mixed Native and Nonnative English Speakers

Notes: This figure illustrates the effect of ChatGPT's release on the Type-Token Ratio (TTR) for articles authored by native and nonnative English-speaking researchers. For interpretation, refer to the figure note provided with Figure 3.

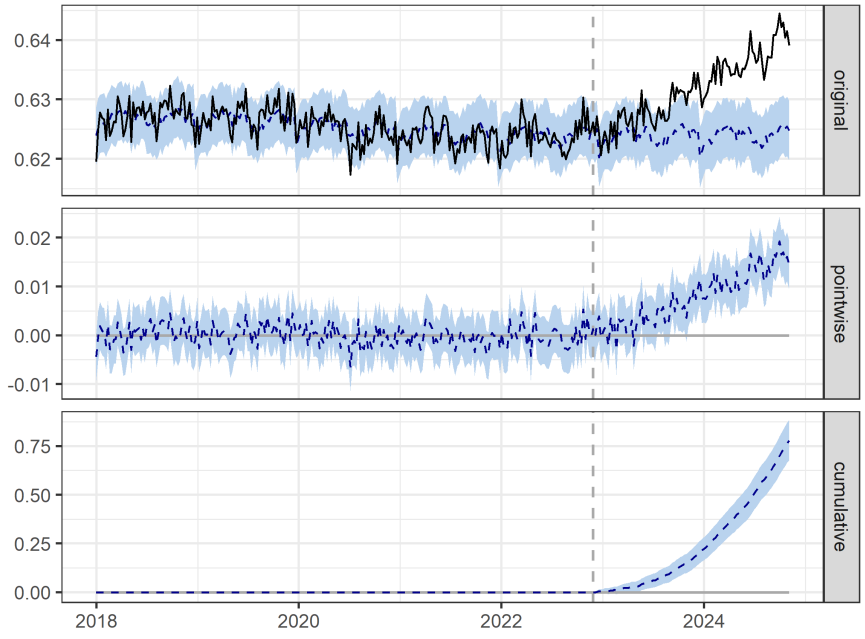

Figure 5: The Impact of ChatGPT on the Type-Token Ratio for Articles Authored by Nonnative English Speakers

Notes: This figure illustrates the effect of ChatGPT's release on the Type-Token Ratio (TTR) for articles authored exclusively by nonnative English-speaking researchers. For interpretation, refer to the figure note provided with Figure 3.

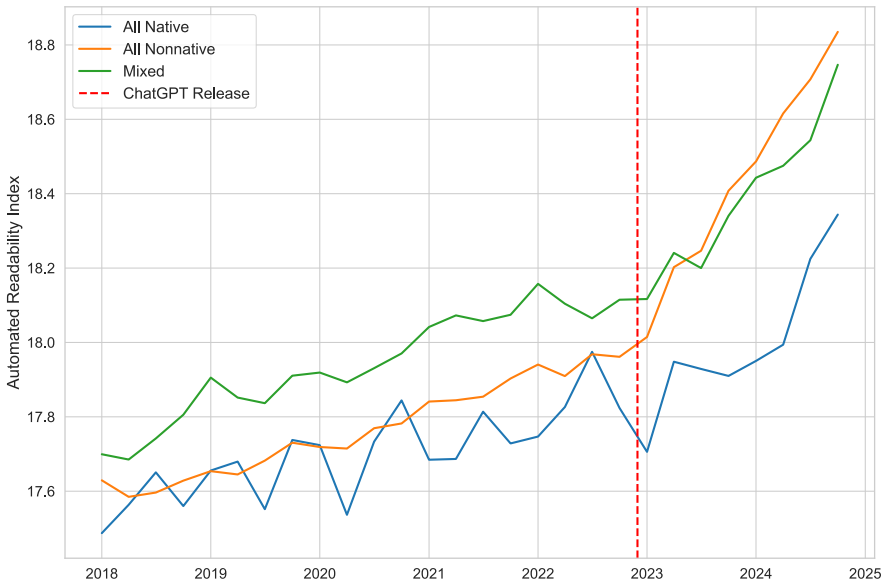

Figure 6: Quarterly Average Automated Readability Index by Language Categorization

Notes: This figure displays the quarterly average Automated Readability Index (ARI) calculated from abstracts of articles published on arxiv.org from January 1, 2018, to November 14, 2024. ARI, which primarily considers word length and sentence length (Smith & Senter, 1967). ARI measures the readability of text, with higher values indicating greater complexity and difficulty in reading. The articles are categorized into three authorship groups: "All Native" refers to articles written entirely by native English-speaking authors; "All Nonnative" includes articles authored solely by non-native English speakers; "Mixed" represents articles written by a combination of native and non-native authors. The red dashed vertical line marks the release date of ChatGPT on 30 November 2022.

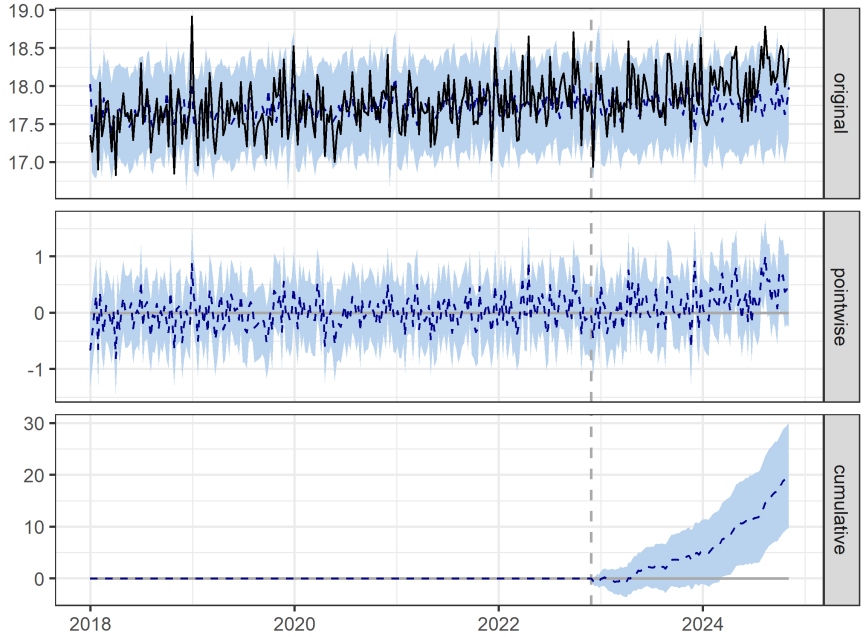

Figure 7: The Impact of ChatGPT on the Average Automated Readability Index for Articles Authored by Native English Speakers

Notes: This figure, generated using the CausalImpact package (Brodersen et al., 2017), illustrates the effect of ChatGPT's release on the Automated Readability Index (ARI) for articles authored exclusively by native English speakers, calculated from abstracts of articles published on arxiv.org from January 1, 2018, to November 14, 2024. ARI, which measures text readability and complexity (Smith & Senter, 1967). The top panel shows the observed data (solid black line) and the counterfactual predictions (blue line), with the shaded region representing the 95% credible intervals. The dashed vertical line marks the release date of ChatGPT, 30 November 2022. The middle panel displays the pointwise impact, calculated as the difference between observed and predicted values. The bottom panel visualizes the cumulative impact over time.

Table 3: Pointwise (Standardized) Estimated Impacts on the Type-Token Ratio (TTR) by Authorship Group for the Last Five Weeks

	Native		**Mixed**		**Nonnative**	
Date	Impact	Std. Impact	Impact	Std. Impact	Impact	Std. Impact
2024-10-06	0.0150	0.1803	0.0160	0.1930	0.0192	0.2318
2024-10-13	-0.0002	-0.0023	0.0099	0.1199	0.0166	0.2000
2024-10-20	0.0039	0.0474	0.0109	0.1311	0.0170	0.2050
2024-10-27	0.0012	0.0146	0.0111	0.1335	0.0153	0.1842
2024-11-03	0.0071	0.0850	0.0127	0.1526	0.0161	0.1937
Average	0.0054	0.0650	0.0121	0.1460	0.0168	0.2030

Notes: This table summarizes the estimated pointwise impacts on the Type-Token Ratio (TTR) over the last five weeks in our sample. The table categorizes the results by three authorship groups: Native, Mixed, and Nonnative. The "Impact" columns show the raw estimated effects of the intervention on TTR for each group, while "Std. Impact" columns report the standardized impacts, calculated by dividing the raw estimates by the standard deviation of TTR in the sample. The estimates correspond to the last five weeks of pointwise impacts displayed in Figures 3, 4, and 5, shown in the middle panels of these figures.

4.2 The Automated Readability Index

Figure 6 presents the quarterly averages of the ARI for the three authorship groups. Unlike TTR, ARI consistently trends upwards across all groups during the pre-ChatGPT period, with a noticeable increasing trend during the post-ChatGPT period, especially for nonnative authored articles. In the pre-ChatGPT period, mixed authored articles had the highest readability score, but by November 2024, nonnative authored articles are the highest.

Next, we estimate the effects for each authorship group using BSTS. Figure 7 presents the results for native English speakers. While the pointwise impact estimates are imprecise, the cumulative impact suggests a steady rise in the ARI over time, with the credible interval differing from zero. The pointwise impact over the last five weeks of our sample period is estimated to be approximately 0.56, representing a roughly 0.16 standard deviation increase, or half a U.S. grade level based on the measure underlying the index.

Figure 8 shows a clearer increase in the ARI for mixed-authored articles. The estimated pointwise impact during the same period is approximately 0.63, representing roughly a 0.18 standard deviation increase. The steady rise in the ARI suggests that the introduction of ChatGPT may have reinforced an existing trend toward increased writing complexity.

Figure 9 shows an even more dramatic increase in the ARI for nonnative English speakers. During the same period, the estimated pointwise impact is approximately 0.88, almost a full school grade, representing roughly a 0.25 standard deviation increase. Both the pointwise and cumulative impacts for nonnative authors are nearly double those observed

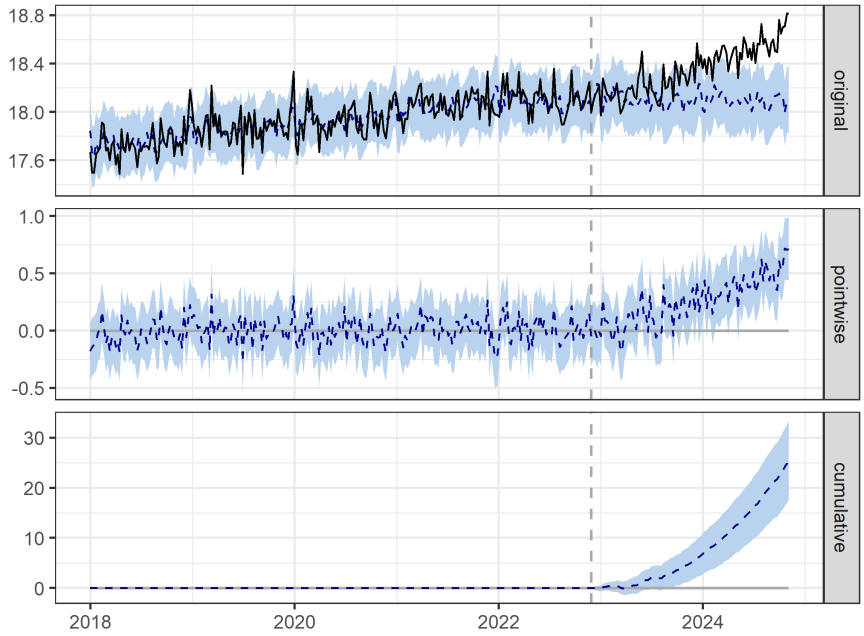

Figure 8: The Impact of ChatGPT on the Average Automated Readability Index for Articles Authored by Mixed Native and Nonnative English Speakers

Notes: This figure illustrates the effect of ChatGPT's release on the Automated Readability Index (ARI) for articles authored by native and nonnative English-speaking researchers. For interpretation, refer to the figure note provided with Figure 7.

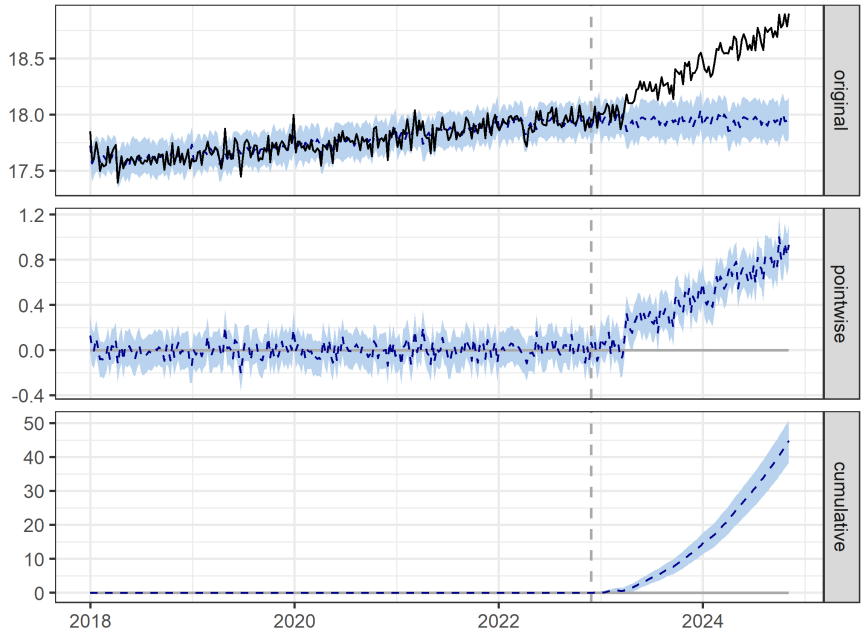

Figure 9: The Impact of ChatGPT on the Average Automated Readability Index for Articles Authored by Nonnative English Speakers

Notes: This figure illustrates the effect of ChatGPT's release on the Automated Readability Index (ARI) for articles authored exclusively by nonnative English-speaking researchers. For interpretation, refer to the figure note provided with Figure 7.

for native English speaker-authored articles.

To provide a closer examination of the most recent period in our sample, Table 4 presents the pointwise impacts for the last five weeks. Overall, while the ARI continues to increase across all authorship groups, the impact is more pronounced among exclusively nonnative authored articles. This contrasts with the results for lexical diversity, as our findings suggest a potential widening of the gap in writing complexity between native and nonnative speakers. Post-ChatGPT, the complexity of nonnative-authored articles increased further, whereas the increase among native authored articles was comparatively smaller.

Table 4: Pointwise (Standardized) Estimated Impacts on the Automated Readability Index by Authorship Group for the Last Five Weeks

	Native		**Mixed**		**Nonnative**	
Date	Impact	Std. Impact	Impact	Std. Impact	Impact	Std. Impact
2024-10-06	0.7106	0.2054	0.6062	0.1752	0.9992	0.2889
2024-10-13	0.6215	0.1797	0.5250	0.1518	0.8143	0.2354
2024-10-20	0.6658	0.1925	0.5691	0.1645	0.7894	0.2282
2024-10-27	0.3707	0.1072	0.7183	0.2077	0.9515	0.2751
2024-11-03	0.4183	0.1209	0.7089	0.2049	0.8416	0.2433
Average	0.5574	0.1611	0.6255	0.1808	0.8792	0.2542

Notes: This table summarizes the estimated pointwise impacts on the Automated Readability Index (ARI) over the last five weeks in our sample. The table categorizes the results by three authorship groups: Native, Mixed, and Nonnative. The "Impact" columns show the raw estimated effects of the intervention on ARI for each group, while "Std. Impact" columns report the standardized impacts, calculated by dividing the raw estimates by the standard deviation of ARI in the sample. The estimates correspond to the last five weeks of pointwise impacts displayed in Figures 7, 8, and 9, shown in the middle panels of these figures.

4.3 Additional Analyses and Tests

In this subsection, we assess the robustness of our findings. First, we explore alternative writing metrics to evaluate lexical diversity and complexity. Second, we perform a placebo test to verify the stability of our dynamic forecasts. Third, we explore changes in the composition of articles.

4.3.1 Alternative Writing Metrics

Herdan's C: We evaluated the lexical diversity of articles using Herdan's C (Herdan, 1964) (log-TTR) as an alternative to TTR. Herdan's C is less sensitive to text length than TTR, making it a more robust measure of lexical diversity for varying abstract sizes.

Figure B.1 illustrates the quarterly averages of Herdan's C for native, mixed, and nonnative authored articles. Herdan's C increased most dramatically for the nonnative authored articles. Again, we observe that lexical diversity increases across all language groups. Figures B.2, B.3, and B.4 illustrate the BSTS impact estimates. The findings

align with the TTR analysis, supporting our conclusion that lexical diversity increased the most in nonnative-authored articles.

Table B.1 presents the pointwise impacts for the last five weeks. The estimated pointwise impact on Herdan's C is approximately 0.002 for native-authored articles, 0.004 for mixed-authored articles, and 0.006 for nonnative-authored articles. These impacts correspond to increases of approximately 0.10, 0.19, and 0.28 standard deviations, respectively. These standardized impacts are greater than those found for TTR.

Gunning Fog Index: Next, we use the Gunning Fog Index as an alternative measure of text complexity to the ARI (Gunning, 1952). While both indices assess readability, they differ in their emphasis: the ARI focuses more on sentence structure and word length, whereas the Gunning Fog Index places greater weight on the use of complex words.

The results presented in Figure B.5 illustrate the quarterly averages of the Gunning Fog Index for native, mixed, and nonnative authored articles. The readability index changed most dramatically for the nonnative authored articles. Moreover, we observe that all language groups converge to a similar level of readability by November 2024. Figures B.6, B.7, and B.8 illustrate the BSTS impact estimates. These findings closely align with those in Section 4.2. As with the ARI, the impact is most pronounced for nonnative-authored articles compared to the other groups.

Table B.2 presents the pointwise impacts for the last five weeks. The average estimated pointwise impact for native-authored articles is approximately 0.25, while it is 0.40 for mixed-authored articles and 0.59 for nonnative-authored articles. These impacts correspond to increases of approximately 0.09, 0.14, and 0.21 standard deviations, respectively.

4.3.2 Placebo Test

To test the stability of our BSTS model, we conducted a placebo test. The main objective of this test was to verify whether the observed effect in the post-ChatGPT period could be genuinely attributed to the release of ChatGPT, rather than arising from instability in long-term dynamic forecasting.

In the placebo analysis, we used a pre-ChatGPT period where no known intervention took place, followed by a fictitious post-ChatGPT period. Specifically, we defined a placebo post-ChatGPT period from 30 November 2021 (one year before the release of ChatGPT), to 30 November 2022 (the release date of ChatGPT). The placebo test follows the same methodological approach as the main analysis. We estimate the impacts on TTR and ARI until the release of ChatGPT. The method predicts what would have happened during the placebo post-intervention period based on the model, which we compare to

the actual data during that period to assess the predictive accuracy of our approach.

The results from our placebo test are reported in Appendix C. We find no significant pointwise impacts for any group across all metrics analyzed. While cumulative impacts appear to be credibly non-zero for native and nonnative-authored articles in some cases, these effects are in the opposite direction to our main findings for TTR and are negligible in magnitude for both TTR and ARI. This suggests that any bias introduced by the BSTS model is small and does not undermine the validity of our results.

4.3.3 Changes in Composition of Articles

It is possible that LLMs changed the composition of articles, for example by benefiting certain disciplines more than others. If certain disciplines have lower or higher TTR and ARI scores, and/or native and nonnative representation, a change in composition could potentially explain our findings.

To investigate whether composition did change, we applied the BSTS model using the weekly average proportion of arXiv.org articles by subject as the outcome variable. The model generated a counterfactual scenario to estimate what subject proportions would have been if ChatGPT had not been released. Figure 10 shows that, following ChatGPT's release, the proportion of mathematics and physics articles decreased, while computer science articles increased.

Given the increased proportion of computer science publications, and the possibility that computer science abstracts may exhibit richer lexical diversity and greater lexical complexity, shifts in article composition could potentially explain differences between native and nonnative authored articles. To address this concern, we conducted an additional analysis using a partially linear regression model to estimate smoothed trends for the selected outcome variables whilst controlling for subject field. The results, detailed and demonstrated in Appendix D, are consistent with our main results and suggest the estimated effects are robust to controlling for subject field.

The results are reported in Appendix E. Consistent with earlier studies, we find that writing metrics for computer scientists change more dramatically across all three authorship groups. Specifically, the impact on lexical diversity (TTR) is approximately twice as large for computer science articles compared to other fields. The effect on lexical complexity (ARI) is even greater for computer scientists than other researchers. The disparity between native (and mixed) and nonnative authored articles remains persistent in both computer science and non-computer science articles.

These findings highlight a striking disparity in the adoption and impact of AI tools

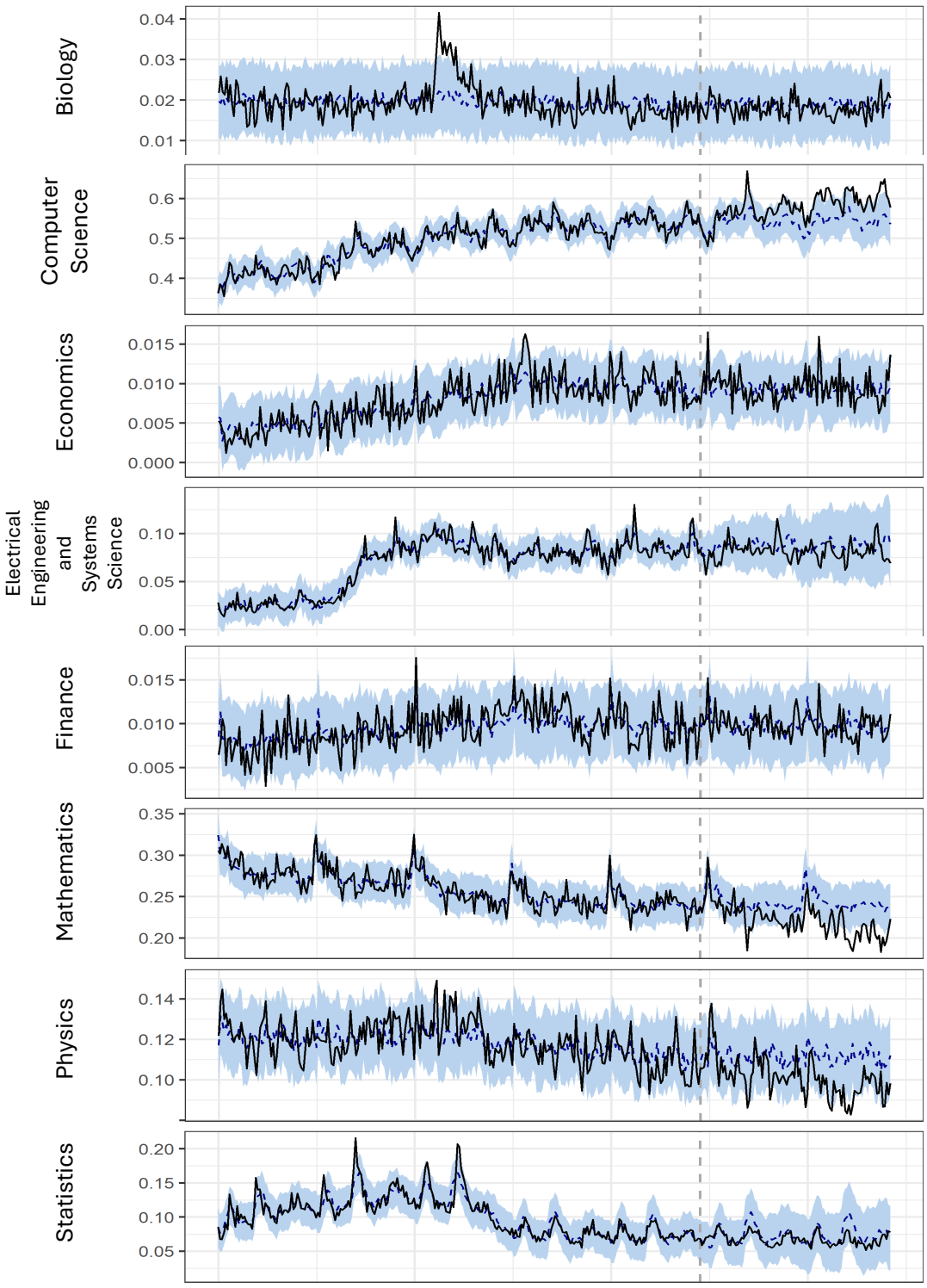

Figure 10: Changes in the Proportion of Subjects Published on arXiv.org Following the Release of ChatGPT

Notes: The figure, generated using the `CausalImpact` package (Brodersen et al., 2017), demonstrates the impact of ChatGPT (and other LLMs) on the proportion of articles published in various subjects on arxiv.org from January 1, 2018, to November 14, 2024. Each panel presents the observed data (solid black line) and counterfactual predictions (blue line), with the shaded regions representing 95% credible intervals. The dashed vertical line indicates the release date of ChatGPT.

across disciplines. Our results are consistent with previous studies that document slower uptake of AI-assisted writing tools among non-computer scientists. This difference may be attributed to varying levels of enthusiasm, hesitation, or technical proficiency across disciplines. Researchers in other fields might exhibit greater skepticism toward these tools or face challenges in integrating them effectively into their workflows. Alternatively, the disparity could reflect differences in the perceived utility or relevance of AI tools for their specific writing needs.

5 Discussion

Our results demonstrate that, since the launch of ChatGPT, there has been a marked increase in lexical diversity and complexity in academic abstracts, and both effects have been strongest for articles with nonnative English-speaking authors. For lexical diversity, this effect has helped to close the gap between native and nonnative authors. For complexity, there has been an increased divergence, with nonnative authored articles being more complex than native authored articles.

While we cannot attribute these recent trends to LLMs with certainty, there is much to support this interpretation. One challenge to our interpretation is that in 2022 the world was still recovering from the Covid-19 pandemic, and this likely influenced authorship dynamics. However, in our descriptive figures for writing metrics we see a sharp trend break that coincides strongly with the ChatGPT launch date, which mirrors the patterns observed in our data and elsewhere (e.g., Kobak et al., 2024) on AI detection. The existence of this structural break is also supported by formal statistical methods. If this effect is due to other forces, the timing would be highly coincidental.

We motivated our analysis by pointing to the disadvantages faced by nonnative English-speaking researchers. For example, Amano et al. (2023) finds that nonnative English-speaking researchers take 51% more time to write articles and are asked to revise their manuscripts 12.5 times more frequently than native speakers. Ramírez-Castañeda (2020) notes that writing a research article in English requires an average of 96.86 extra labor hours compared to writing in Spanish. Our results suggest that LLMs have potentially helped to alleviate these and other challenges faced by nonnative English-speaking researchers. On some measures of writing style (i.e., TTR and Gunning Fog Index) native and nonnative authors have largely converged, which could imply a more even treatment from readers (although there has been increased divergence for ARI).

An important limitation to our analysis is that it does not speak directly to the question of *quality* of writing. The impact of LLMs on academic writing is accompanied by important growing concerns (Schlagwein & Willcocks, 2023). The fact that abstracts have become

more diverse and complex since the launch of ChatGPT implies that they require a higher level of language proficiency to comprehend than in the past. This may make research even less accessible to non-specialist readers. It is also possible that abstracts written with the aid of LLMs contain errors and biases if the technology is used indiscriminately.

As the quality of the AI-generated text continues to improve, the integration of AI in academic writing appears to be evolving, with generated text becoming increasingly sophisticated and harder to differentiate from human writing (Liang, Izzo, et al., 2024; Geng et al., 2024). Liang et al. (2023) found that AI-writing detectors frequently misclassify nonnative English writing as AI generated. Furthermore, Novy-Marx & Velikov (2024) demonstrated the potential of LLMs to automate the production of hundreds of academic finance articles on stock return predictability using simple prompts. These developments suggest that while AI holds transformative potential for academia, its broader impacts on research and publication practices remain uncertain.

We see our results as an important first step towards understanding the impact of LLMs on academic writing, with future research to consider the broader implications for research quality and careers. Our results imply that future research should continue to focus on the different experiences of native and nonnative English-speakers.

Our study took place two years after ChatGPT's release, which is a relatively short timeframe to fully observe the impact of LLMs. Indeed, our estimates show the effect of LLMs on writing to be continuously increasing over the period we studied, implying potentially larger effects to come. The integration of LLMs into teaching and language learning practices could further help alleviate language barrier challenges. While there are some challenges in integrating AI into learning frameworks (Ma et al., 2024; Rahimi & Sevilla-Pavón, 2024), the overall opinion remains positive regarding its prospects for teaching and language learning. Li et al. (2024) found that ChatGPT enhances self-directed learning and teacher workflows, while Xu et al. (2024) demonstrated its ability to improve foreign language self-efficacy and enjoyment among students. Additionally, Jeon et al. (2023) highlighted the innovative potential of AI-driven chatbots, suggesting their utility in multimodal and goal-oriented educational contexts.

Beyond implications for language skills and learning, AI could have profound effects on labor markets. Eloundou et al. (2024) identify LLMs as general-purpose technologies with broad economic, social, and policy implications, particularly in higher-income jobs. At a more granular level, Noy & Zhang (2023) and Brynjolfsson et al. (2023) highlight productivity boosts from generative AI, narrowing skill gaps for less experienced workers in tasks like professional writing and customer support. Acemoglu (2024) cautions that while AI can improve productivity in certain low-skill tasks, it may exacerbate inequality

if the creation of high-value tasks lags. He finds that AI's impact is more evenly distributed across demographic groups compared to previous automation technologies and concludes that AI advances are less likely to exacerbate inequality to the same extent. We find notable disparities in the adoption of AI across disciplines, with computer scientists emerging as the leaders in leveraging AI within their research outputs (see Section 4.3.3). This implies that the impact of AI may not be evenly distributed.

6 Conclusion

This study provided empirical evidence that LLMs are influencing academic writing, particularly for articles written by nonnative English-speakers, potentially helping to bridge gaps between natives and nonnatives. Our analysis is also enabled by use of LLMs to classify authors as native and nonnative English speakers. Although this approach is not without limitations and errors, our methodology demonstrated LLMs' other benefits in improving research prospects.

Across various analyses, we consistently find more pronounced improvements in lexical diversity and lexical complexity of articles authored by nonnative English speakers. These findings highlight the transformative potential of LLMs in addressing long-standing inequities in academic careers and publishing. However, more research is needed on the influence of LLMs on the quality and impact of academic writing to evaluate its benefits (and potential costs). One thing is clear – LLMs are being widely used, especially by nonnative English-speaking authors. This highlights the importance of ensuring researchers are educated on how to utilize this technology productively and responsibly.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Acknowledgements

The authors acknowledge the use of OpenAI's `gpt-4o` to enhance writing quality, automate the creation of LaTeX tables, and perform text classification in this research. The last application is central to the methodology, as detailed in this paper. All AI-generated suggestions and code were critically reviewed, edited, and validated by the authors.

Funding

This research was supported by Royal Holloway, University of London, which provided £401.11 to cover the classification costs associated with this project.

References

Acemoglu, D. (2024, May). *The simple macroeconomics of ai* (Working Paper No. 32487). National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w32487 doi: 10.3386/w32487

Al-Zaabi, A., Al-Amri, A., Albalushi, H., Aljabri, R., & Aal Abdulsalam, A. (2023). Chatgpt applications in academic research: A review of benefits, concerns, and recommendations. *Biorxiv*, 2023–08.

Amano, T., González-Varo, J. P., & Sutherland, W. J. (2016). Languages are still a major barrier to global science. *PLoS biology*, *14*(12), e2000933.

Amano, T., Ramírez-Castañeda, V., Berdejo-Espinola, V., Borokini, I., Chowdhury, S., Golivets, M., ... others (2023). The manifold costs of being a non-native english speaker in science. *PLoS Biology*, *21*(7), e3002184.

arXiv.org Submitters. (2024). *arxiv dataset.* Kaggle. Retrieved from https://www.kaggle.com/dsv/7548853 (Accessed:) doi: 10.34740/KAGGLE/DSV/7548853

Berdejo-Espinola, V., & Amano, T. (2023). Ai tools can improve equity in science. *Science*, *379*(6636), 991–991.

Bisi, T., Risser, A., Clavert, P., Migaud, H., & Dartus, J. (2023). What is the rate of text generated by artificial intelligence over a year of publication in orthopedics & traumatology: Surgery & research? analysis of 425 articles before versus after the launch of chatgpt in november 2022. *Orthopaedics & Traumatology: Surgery & Research*, *109*(8), 103694.

Brodersen, K. H., Gallusser, F., Koehler, J., Remy, N., & Scott, S. L. (2015). Inferring causal impact using Bayesian structural time-series models. *The Annals of Applied Statistics*, *9*(1), 247 – 274. Retrieved from https://doi.org/10.1214/14-AOAS788 doi: 10.1214/14-AOAS788

Brodersen, K. H., Hauser, A., & Hauser, M. A. (2017). Package causalimpact. *Google LLC: Mountain View, CA, USA*.

Brynjolfsson, E., Li, D. R., & Lindsey. (2023). Generative ai at work. *arXiv preprint arXiv:2304.11771*.

Chotlos, J. W. (1944). Iv. a statistical and comparative analysis of individual written language samples. *Psychological Monographs*, *56*(2), 75.

Duarte, F. (2024). *Number of chatgpt users (nov 2024).* Retrieved from https://explodingtopics.com/blog/chatgpt-users (Accessed: 2024-11-28)

Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2024). Gpts are gpts: Labor market impact potential of llms. *Science*, *384*(6702), 1306–1308.

Feld, J., Lines, C., & Ross, L. (2024). Writing matters. *Journal of Economic Behavior & Organization*, *217*, 378–397.

Geng, M., Chen, C., Wu, Y., Chen, D., Wan, Y., & Zhou, P. (2024). The impact of large language models in academia: from writing to speaking. *arXiv preprint arXiv:2409.13686*.

Gunning, R. (1952). *The technique of clear writing.* New York: McGraw-Hill.

Hanauer, D. I., Sheridan, C. L., & Englander, K. (2019). Linguistic injustice in the writing of research articles in english as a second language: Data from taiwanese and mexican researchers. *Written Communication*, *36*(1), 136–154.

Herdan, G. (1964). Quantitative linguistics or generative grammar? *Linguistics*, *2*(4), 56–65. doi: 10.1515/ling.1964.2.4.56

Huang, J., Yang, D. M., Rong, R., Nezafati, K., Treager, C., Chi, Z., ... others (2024). A critical assessment of using chatgpt for extracting structured data from clinical notes. *npj Digital Medicine*, *7*(1), 106.

Hwang, S. I., Lim, J. S., Lee, R. W., Matsui, Y., Iguchi, T., Hiraki, T., & Ahn, H. (2023). Is chatgpt a "fire of prometheus" for non-native english-speaking researchers in academic writing? *Korean Journal of Radiology*, *24*(10), 952.

Jeon, J., Lee, S., & Choe, H. (2023). Beyond chatgpt: A conceptual framework and systematic review of speech-recognition chatbots for language learning. *Computers & Education*, 104898.

Kobak, D., Márquez, R. G., Horvát, E.-Á., & Lause, J. (2024). Delving into chatgpt usage in academic writing through excess vocabulary. *arXiv preprint arXiv:2406.07016*.

Kreitmeir, D., & Raschky, P. A. (2024). The heterogeneous productivity effects of generative ai. *arXiv preprint arXiv:2403.01964*.

Lamichhane, B. (2023). Evaluation of chatgpt for nlp-based mental health applications. *arXiv preprint arXiv:2303.15727*.

Li, B., Lowell, V. L., Wang, C., & Li, X. (2024). A systematic review of the first year of publications on chatgpt and language education: Examining research on chatgpt's use in language learning and teaching. *Computers and Education: Artificial Intelligence*, 100266.

Liang, W., Izzo, Z., Zhang, Y., Lepp, H., Cao, H., Zhao, X., ... others (2024). Monitoring ai-modified content at scale: A case study on the impact of chatgpt on ai conference peer reviews. *arXiv preprint arXiv:2403.07183*.

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). Gpt detectors are biased against non-native english writers. *Patterns*, *4*(7).

Liang, W., Zhang, Y., Wu, Z., Lepp, H., Ji, W., Zhao, X., ... others (2024). Mapping the increasing use of llms in scientific papers. *arXiv preprint arXiv:2404.01268*.

Lingard, L., Chandritilake, M., de Heer, M., Klasen, J., Maulina, F., Olmos-Vega, F., & St-Onge, C. (2023). Will chatgpt's free language editing service level the playing field in science communication?: Insights from a collaborative project with non-native english scholars. *Perspectives on Medical Education*, *12*(1), 565.

Liu, W. (2017). The changing role of non-english papers in scholarly communication: Evidence from web of science's three journal citation indexes. *Learned Publishing*, *30*(2), 115–123.

Ma, Q., Crosthwaite, P., Sun, D., & Zou, D. (2024). Exploring chatgpt literacy in language education: A global perspective and comprehensive approach. *Computers and education: Artificial intelligence*, *7*, 100278.

Nishant, R., Schneckenberg, D., & Ravishankar, M. (2024). The formal rationality of artificial intelligence-based algorithms and the problem of bias. *Journal of Information Technology*, *39*(1), 19–40. doi: 10.1177/02683962231176842

Novy-Marx, R., & Velikov, M. (2024). Ai-powered (finance) scholarship. *Available at SSRN*.

Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science (New York, N.Y.)*, *381*(6654), 187-192.

Rahimi, A. R., & Sevilla-Pavón, A. (2024). The role of chatgpt readiness in shaping language teachers' language teaching innovation and meeting accountability: A bisymmetric approach. *Computers and Education: Artificial Intelligence*, *7*, 100258.

Ramírez-Castañeda, V. (2020). Disadvantages in preparing and publishing scientific papers caused by the dominance of the english language in science: The case of colombian researchers in biological sciences. *PloS one*, *15*(9), e0238372.

Reuters. (2023). *Chatgpt sets record for fastest growing user base - analyst note.* Retrieved from https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/ (Accessed: 2024-11-28)

Schlagwein, D., & Willcocks, L. (2023). 'chatgpt et al.': The ethics of using (generative) artificial intelligence in research and science. *Journal of Information Technology*, *38*(3), 232-238. doi: 10.1177/02683962231200411

Smith, E. A., & Senter, R. (1967). *Automated readability index* (Vol. 66) (No. 220). Aerospace Medical Research Laboratories, Aerospace Medical Division, Air ....

Stokel-Walker, C. (2023). Chatgpt listed as author on research papers: many scientists disapprove. *Nature*, *613*(7945), 620–621.

Templin, M. C. (1957). *Certain language skills in children; their development and interrelationships.* University of Minnesota Press.

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. *Nature medicine*, *29*(8), 1930–1940.

Tweedie, F. J., & Baayen, R. H. (1998). How variable may a constant be? measures of lexical richness in perspective. *Computers and the Humanities*, *32*, 323–352.

Uribe, S. E., & Maldupa, I. (2024). Estimating the use of chatgpt in dental research publications. *Journal of Dentistry*, *149*, 105275.

Van Noorden, R., & Perkel, J. M. (2023). Ai and science: what 1,600 researchers think. *Nature*, *621*(7980), 672–675.

Xu, S., Ye, X., Zhang, M., & Wang, P. (2024). Impact of chatgpt on the writing style of condensed matter physicists. *arXiv preprint arXiv:2408.17325*.

A Author Classification with LLM

In this study, we used the model `gpt-4o-2024-08-06` to categorize author names based on gender and language background, leveraging name features for prediction. The model was instructed to classify author names into one of two categories: native English speaker, nonnative English speaker. The categorization task involved dividing the names into batches (with size set to 10) to minimize our cost and processing them sequentially. For each batch, the model was prompted with:

```
You are an expert in onomastics, the study of names and
   their origins. I will provide you with a list of 10 names
   of individuals who have published articles on arXiv.org.
   For each name, predict whether the individual is likely a
   native English speaker or a nonnative English speaker
   based on the name. You might know the person because
   there is a chance that your model is trained with the
   articles published on arXiv; if not, try to use
   etymological information in the names and surnames. Your
   outputs must strictly be one of the following two
   options: native or nonnative. List each prediction on a
   new line, in the same order as the names provided,
   without quotation marks or additional words. Do not
   provide any explanations or comments. Under no
   circumstances should you skip any name or provide fewer
   than 10 predictions.

Depending on the 10, example output format will be like this:
native
nonnative
native
nonnative
...
native
```

To optimize the model's performance, the"temperature" parameter was set to 0 to ensure the model's predictions were highly deterministic.

This prompt ensured that the model adhered to the classification structure and output format required for our analysis. Responses that do not adhere to this output structure are fixed in a subsequent prompt. These errors occur because some articles include

organization names, and the LLM does not always respond as instructed, even though it was explicitly told to follow the output format. In some cases, the output is returned in a stylized structure (for example, with bullet points). Although these cases are rare, they cause parsing errors, which is why we reclassify all authors in those batches. In total, 231 batches (2,310 authors) are reclassified in a subsequent process.

B Alternative Writing Metrics

See Section 4.3.1 for discussions of alternative writing metrics.

- **Herdan's C** (Herdan, 1964).
  - **Figure B.1** shows the quarterly averages of Herdan's C for native, mixed, and nonnative-authored articles.
  - **Figure B.2** shows the impact estimations for Herdan's C for native-authored articles.
  - **Figure B.3** shows the impact estimations for Herdan's C for mixed-authored articles.
  - **Figure B.4** shows the impact estimations for Herdan's C for nonnative-authored articles.
  - **Table B.1** summarizes the estimated pointwise impacts for the last five weeks in our sample.
- **Gunning Fog Index** (Gunning, 1952).
  - **Figure B.5** shows the quarterly averages of the Gunning Fog Index for native, mixed, and nonnative-authored articles.
  - **Figure B.6** shows the impact estimations for the Gunning Fog Index for native-authored articles.
  - **Figure B.7** shows the impact estimations for the Gunning Fog Index for mixed-authored articles.
  - **Figure B.8** shows the impact estimations for the Gunning Fog Index for nonnative-authored articles.
  - **Table B.2** summarizes the estimated pointwise impacts for the last five weeks in our sample.

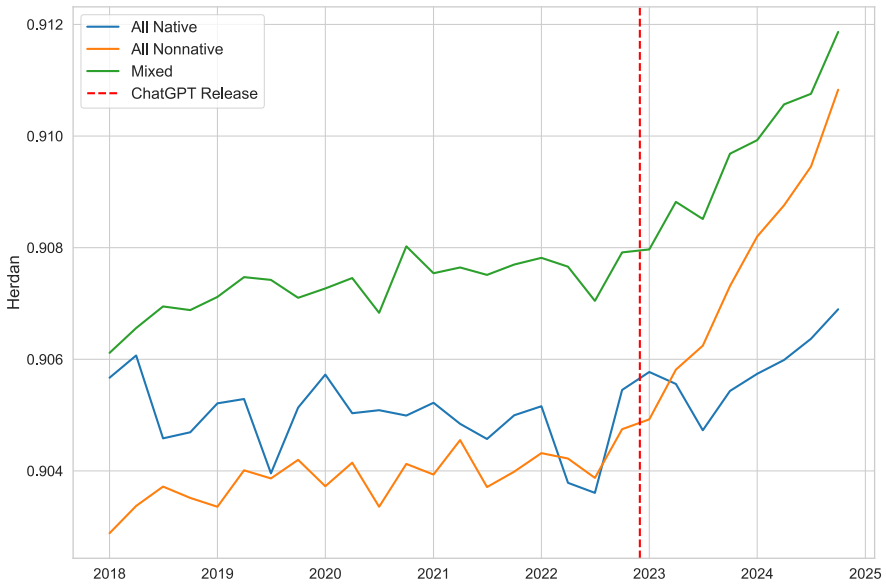

Figure B.1: Quarterly Average Herdan's C by Language Categorization

Notes: This figure displays the quarterly average Herdan's C calculated from abstracts of articles published on arxiv.org from January 1, 2018, to November 14, 2024. Herdan's C (Herdan, 1964) (log-TTR) serves as an alternative to TTR and is less sensitive to text length, making it a more robust measure of lexical diversity across varying abstract sizes. Higher values of Herdan's C indicate a broader range of unique vocabulary relative to the text length, providing a clearer understanding of lexical diversity. The articles are categorized into three authorship groups: "All Native" refers to articles written entirely by native English-speaking authors; "All Nonnative" includes articles authored solely by non-native English speakers; "Mixed" represents articles written by a combination of native and non-native authors. The red dashed vertical line marks the release date of ChatGPT on 30 November 2022.

Table B.1: Pointwise (Standardized) Estimated Impacts on Herdan's C by Authorship Group for the Last Five Weeks

	Native		**Mixed**		**Nonnative**	
	Impact	Std. Impact	Impact	Std. Impact	Impact	Std. Impact
2024-10-06	0.0044	0.1924	0.0053	0.2307	0.0072	0.3114
2024-10-13	0.0011	0.0469	0.0040	0.1737	0.0063	0.2743
2024-10-20	0.0023	0.1016	0.0039	0.1701	0.0066	0.2891
2024-10-27	0.0008	0.0356	0.0040	0.1739	0.0056	0.2456
2024-11-03	0.0030	0.1323	0.0042	0.1830	0.0060	0.2616
Average	0.0023	0.1018	0.0043	0.1863	0.0064	0.2764

Notes: This table summarizes the estimated pointwise impacts on Herdan's C over the last five weeks in our sample. The table categorizes the results by three authorship groups: Native, Mixed, and Nonnative. The "Impact" columns show the raw estimated effects of the intervention on Herdan's C for each group, while "Std. Impact" columns report the standardized impacts, calculated by dividing the raw estimates by the standard deviation of Herdan's C in the sample. The estimates correspond to the last five weeks of pointwise impacts displayed in Figures B.2, B.3, and B.4, shown in the middle panels of these figures.

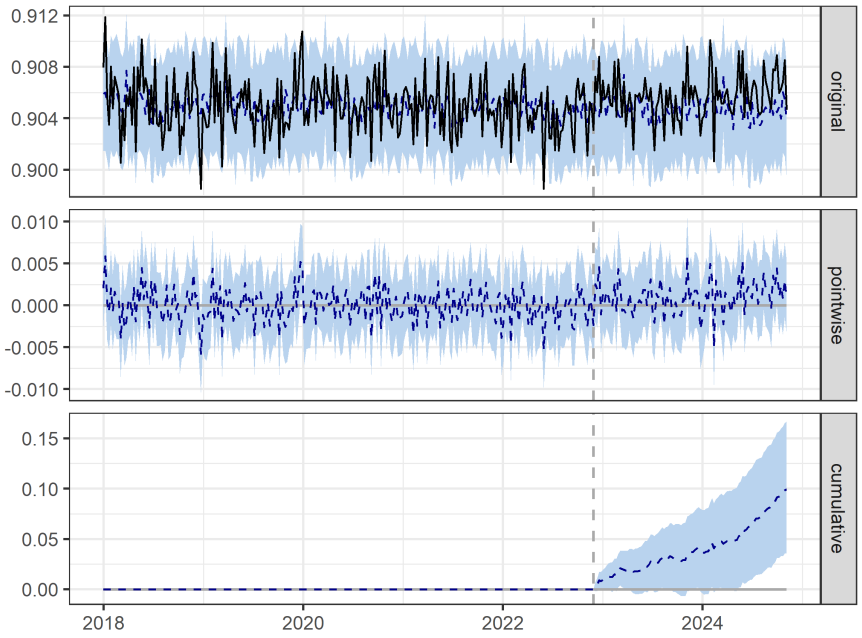

Figure B.2: The Impact of ChatGPT on the Average Herdan's C for Articles Authored by Native English Speakers

Notes: This figure, generated using the CausalImpact package (Brodersen et al., 2017), illustrates the effect of ChatGPT's release on Herdan's C (log-TTR) for articles authored exclusively by native English speakers, calculated from abstracts of articles published on arxiv.org from January 1, 2018, to November 14, 2024. Herdan's C (Herdan, 1964) serves as a measure of lexical diversity, with higher values indicating a broader range of unique vocabulary relative to the text length. The top panel shows the observed data (solid black line) and the counterfactual predictions (blue line), with the shaded region representing the 95% credible intervals. The dashed vertical line marks the release date of ChatGPT, 30 November 2022. The middle panel displays the pointwise impact, calculated as the difference between observed and predicted values. The bottom panel visualizes the cumulative impact over time.

Table B.2: Pointwise (Standardized) Estimated Impacts on the Gunning Fog Index by Authorship Group for the Last Five Weeks

	Native		**Mixed**		**Nonnative**	
	Impact	Std. Impact	Impact	Std. Impact	Impact	Std. Impact
2024-10-06	0.4375	0.1531	0.4168	0.1459	0.6629	0.2320
2024-10-13	0.0753	0.0264	0.3212	0.1124	0.5542	0.1940
2024-10-20	0.4899	0.1715	0.3364	0.1177	0.5338	0.1868
2024-10-27	0.1324	0.0463	0.4488	0.1571	0.6447	0.2257
2024-11-03	0.1226	0.0429	0.4594	0.1608	0.5407	0.1893
Average	0.2515	0.0880	0.3965	0.1388	0.5872	0.2055

Notes: This table summarizes the estimated pointwise impacts on the Gunning Fog Index over the last five weeks in our sample. The table categorizes the results by three authorship groups: Native, Mixed, and Nonnative. The "Impact" columns show the raw estimated effects of the intervention on the Gunning Fog Index for each group, while "Std. Impact" columns report the standardized impacts, calculated by dividing the raw estimates by the standard deviation of the Gunning Fog Index in the sample. The estimates correspond to the last five weeks of pointwise impacts displayed in Figures B.6, B.7, and B.8, shown in the middle panels of these figures.

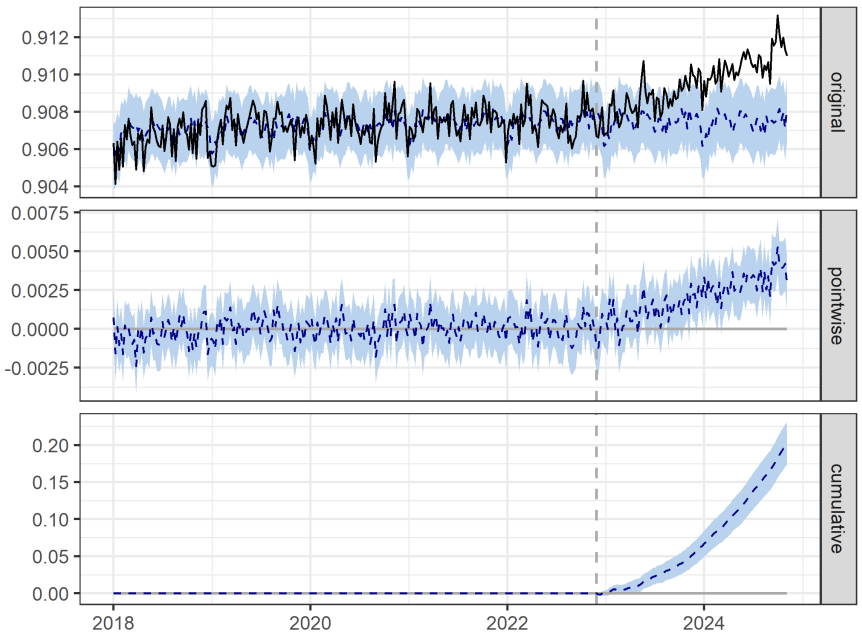

Figure B.3: The Impact of ChatGPT on the Average Herdan's C for Articles Authored by Native and Nonnative English Speakers

Notes: This figure illustrates the effect of ChatGPT's release on the Herdan's C Index for articles authored by native and nonnative English-speaking researchers. For interpretation, refer to the figure note provided with Figure B.2.

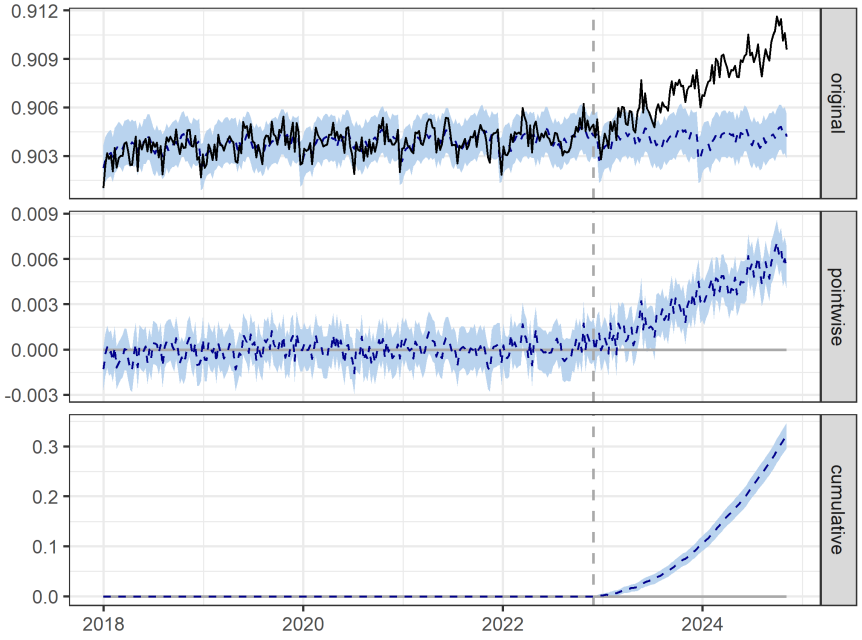

Figure B.4: The Impact of ChatGPT on the Average Herdan's C for Articles Authored by Nonnative English Speakers

Notes: This figure illustrates the effect of ChatGPT's release on the Herdan's C Index for articles authored exclusively by nonnative English-speaking researchers. For interpretation, refer to the figure note provided with Figure B.2.

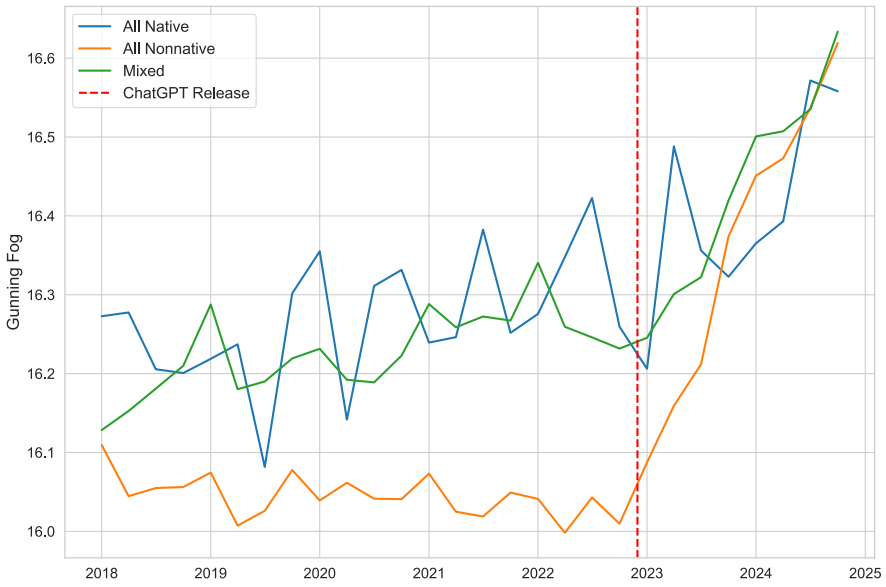

Figure B.5: Quarterly Average Gunning Fog Index by Language Categorization

Notes: This figure displays the quarterly average Gunning Fog Index calculated from abstracts of articles published on arxiv.org from January 1, 2018, to November 14, 2024. The Gunning Fog Index (Gunning, 1952) measures the readability of text, with higher values indicating greater complexity and difficulty. The articles are categorized into three authorship groups: "All Native" refers to articles written entirely by native English-speaking authors; "All Nonnative" includes articles authored solely by non-native English speakers; "Mixed" represents articles written by a combination of native and non-native authors. The red dashed vertical line marks the release date of ChatGPT on 30 November 2022.

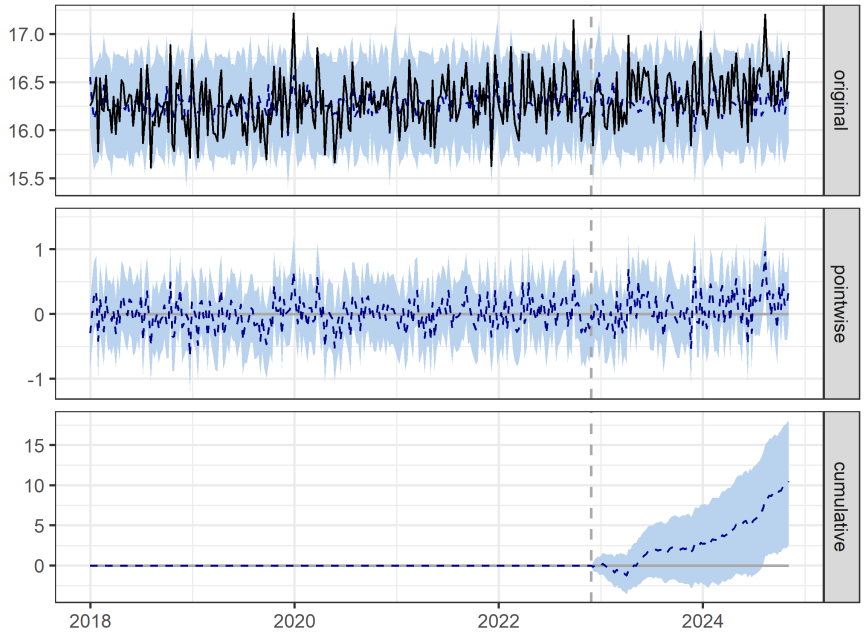

Figure B.6: The Impact of ChatGPT on the Gunning Fog Index for Articles Authored by Native English Speakers

Notes: This figure, generated using the CausalImpact package (Brodersen et al., 2017), illustrates the effect of ChatGPT's release on the Gunning Fog Index for articles authored exclusively by native English speakers, calculated from abstracts of articles published on arxiv.org from January 1, 2018, to November 14, 2024. The Gunning Fog Index (Gunning, 1952) measures the readability of text, with higher values indicating greater complexity and difficulty. The top panel shows the observed data (solid black line) and the counterfactual predictions (blue line), with the shaded region representing the 95% credible intervals. The dashed vertical line marks the release date of ChatGPT, 30 November 2022. The middle panel displays the pointwise impact, calculated as the difference between observed and predicted values. The bottom panel visualizes the cumulative impact over time.

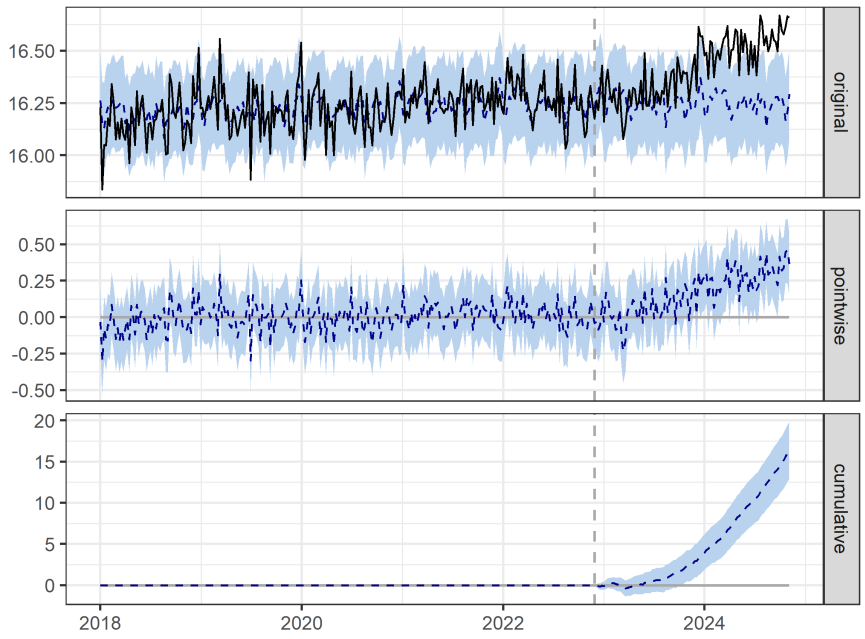

Figure B.7: The Impact of ChatGPT on the Gunning Fog Index for Articles Authored by Native and Nonnative English Speakers

Notes: This figure illustrates the effect of ChatGPT's release on the Gunning Fog Index for articles authored by native and nonnative English-speaking researchers. For interpretation, refer to the figure note provided with Figure B.6.

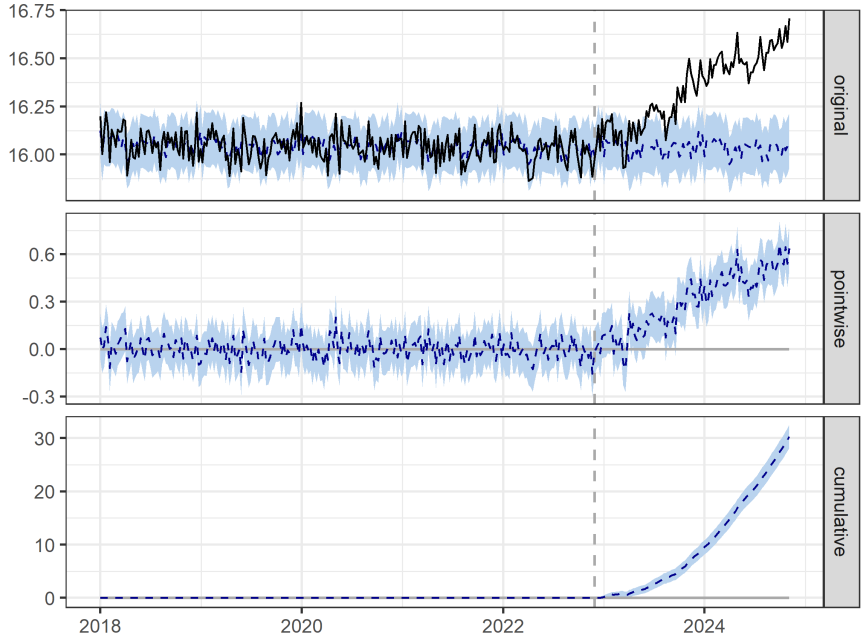

Figure B.8: The Impact of ChatGPT on the Gunning Fog Index for Articles Authored by Nonnative English Speakers

Notes: This figure illustrates the effect of ChatGPT's release on the Gunning Fog Index for articles authored exclusively by nonnative English-speaking researchers. For interpretation, refer to the figure note provided with Figure B.6.

C Placebo Test Results

- **Type-Token Ratio (TTR)**
  - **Figure C.1** shows the placebo test results for the impact of ChatGPT on TTR for articles authored by native English speakers.
  - **Figure C.2** shows the placebo test results for the impact of ChatGPT on TTR for articles authored by mixed native and nonnative English speakers.
  - **Figure C.3** shows the placebo test results for the impact of ChatGPT on TTR for articles authored by nonnative English speakers.
- **Automated Readability Index (ARI)**
  - **Figure C.4** shows the placebo test results for the impact of ChatGPT on the ARI for articles authored by native English speakers.
  - **Figure C.5** shows the placebo test results for the impact of ChatGPT on the ARI for articles authored by mixed native and nonnative English speakers.
  - **Figure C.6** shows the placebo test results for the impact of ChatGPT on the ARI for articles authored by nonnative English speakers.

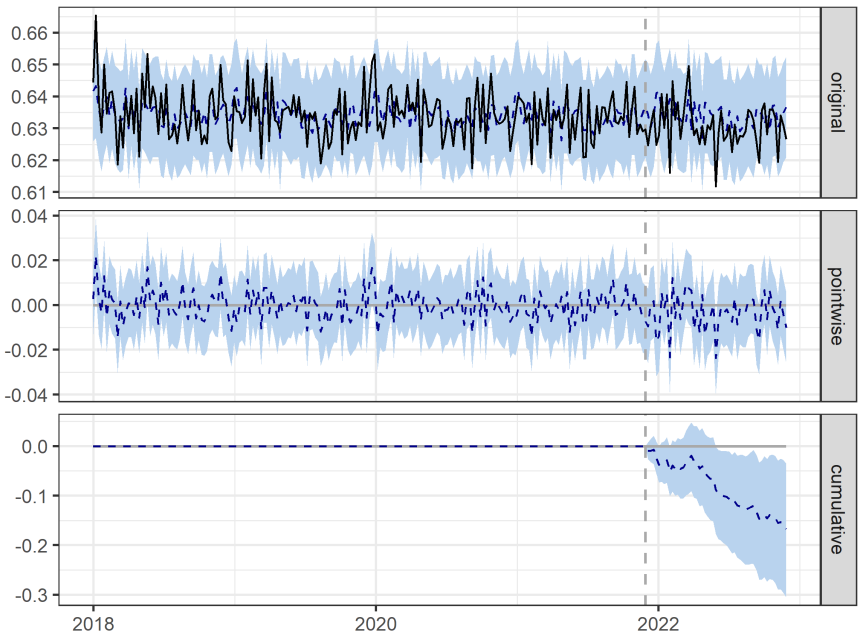

Figure C.1: Placebo Test: The Impact of ChatGPT on the TTR for Articles Authored by Native English Speakers

Notes: This figure, generated using the `CausalImpact` package (Brodersen et al., 2017), illustrates the placebo effect of ChatGPT's release on the Type-Token Ratio (TTR) for articles authored exclusively by native English speakers, calculated from abstracts of articles published on arxiv.org from January 1, 2018, to November 14, 2024. TTR, which measures lexical diversity, is defined as the number of unique words (types) divided by the total number of words (tokens). The top panel shows the observed data (solid black line) and the counterfactual predictions (blue line), with the shaded region indicating the 95% credible intervals. The dashed vertical line marks 30 November 2021, exactly one year before ChatGPT's release. The middle panel depicts the pointwise impact (the difference between observed and predicted values). The bottom panel visualizes the cumulative impact over time.

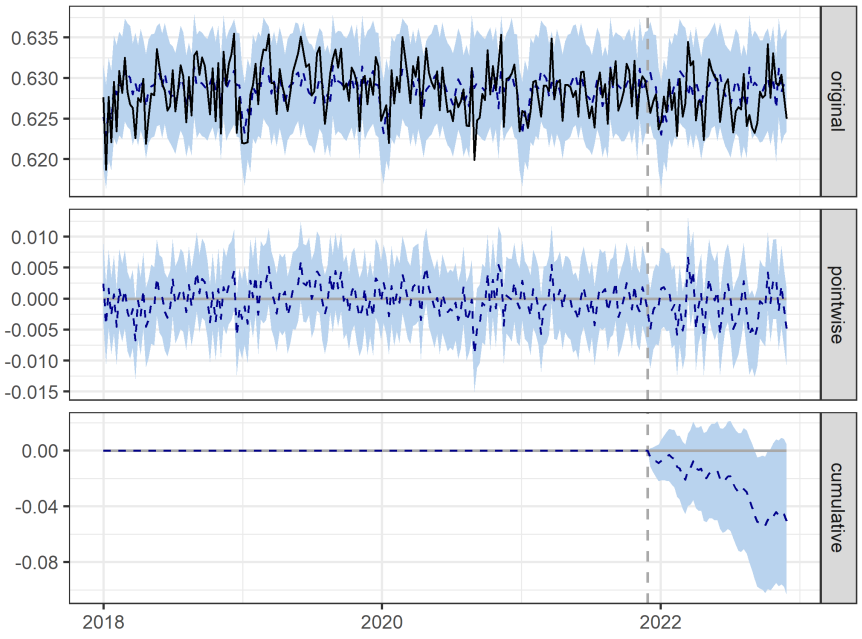

Figure C.2: Placebo Test: The Impact of ChatGPT on the TTR for Articles Authored by Native and Nonnative English Speakers

Notes: This figure illustrates the placebo effect of ChatGPT's release on the Type-Token Ratio (TTR) for articles authored by native and nonnative English speakers. or interpretation, refer to the figure note provided with Figure C.1.

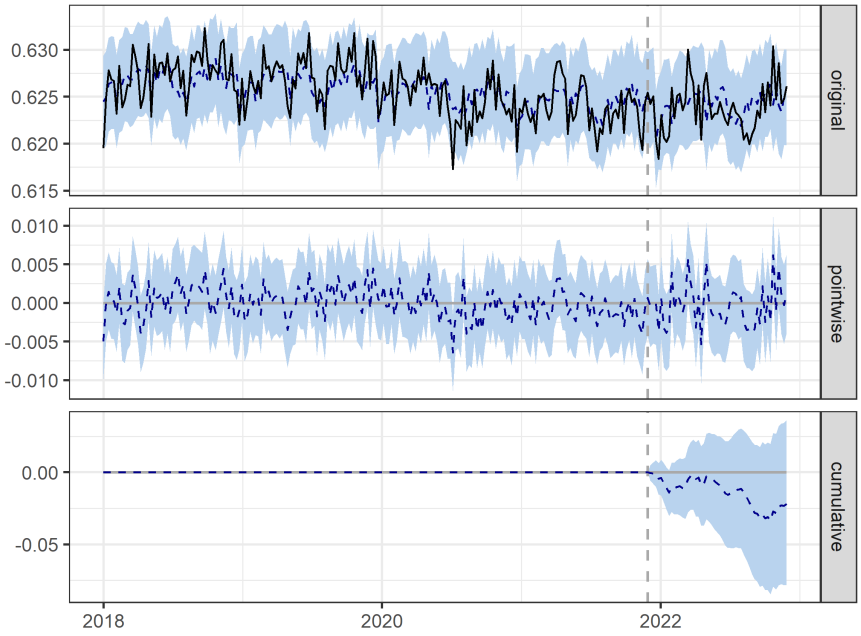

Figure C.3: Placebo Test: The Impact of ChatGPT on the TTR for Articles Authored by Nonnative English Speakers

Notes: This figure illustrates the placebo effect of ChatGPT's release on the Type-Token Ratio (TTR) for articles authored exclusively by nonnative English speakers. or interpretation, refer to the figure note provided with Figure C.1.

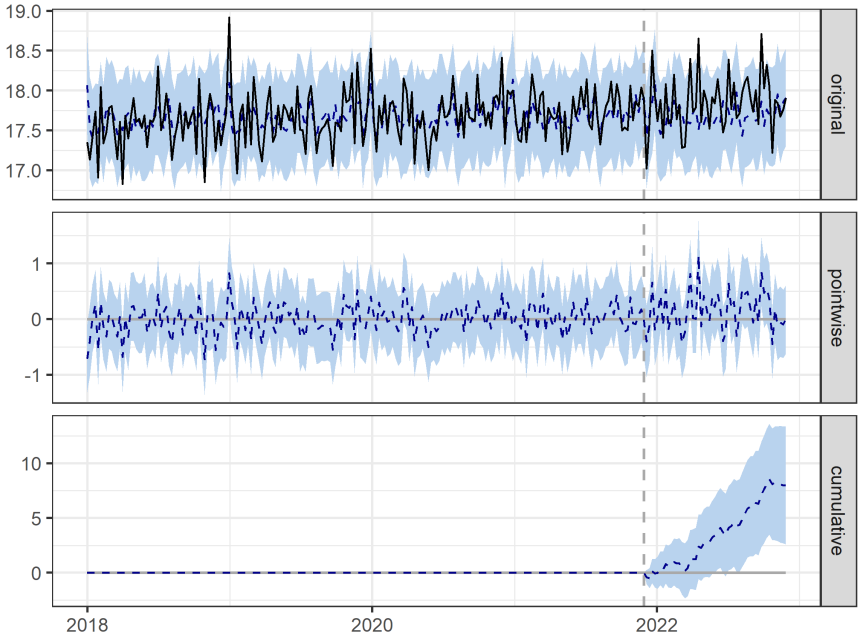

Figure C.4: Placebo Test: The Impact of ChatGPT on the Automated Readability Index for Articles Authored by Native English Speakers

Notes: This figure, generated using the CausalImpact package (Brodersen et al., 2017), illustrates the placebo effect of ChatGPT's release on the Automated Readability Index (ARI) for articles authored exclusively by native English speakers, calculated from abstracts of articles published on arxiv.org from January 1, 2018, to November 14, 2024. ARI, which measures text readability and complexity (Smith & Senter, 1967). The top panel shows the observed data (solid black line) and the counterfactual predictions (blue line), with the shaded region representing the 95% credible intervals. The dashed vertical line marks 30 November 2021, exactly one year before ChatGPT's release. The middle panel displays the pointwise impact, calculated as the difference between observed and predicted values. The bottom panel visualizes the cumulative impact over time.

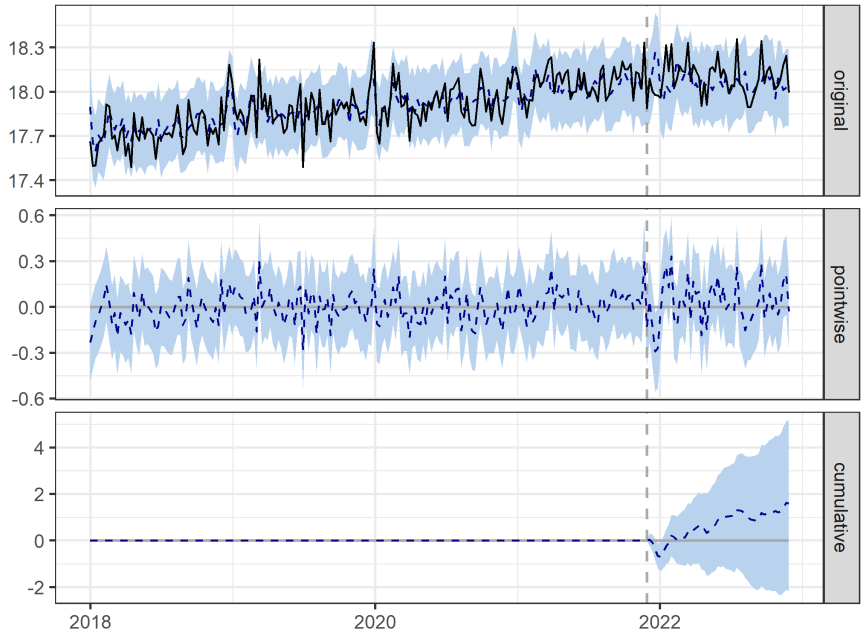

Figure C.5: Placebo Test: The Impact of ChatGPT on the the Automated Readability Index for Articles Authored by Native and Nonnative English Speakers

Notes: This figure illustrates the placebo effect of ChatGPT's release on the Automated Readability Index (ARI) for articles authored by native and nonnative English speakers. or interpretation, refer to the figure note provided with Figure C.4.

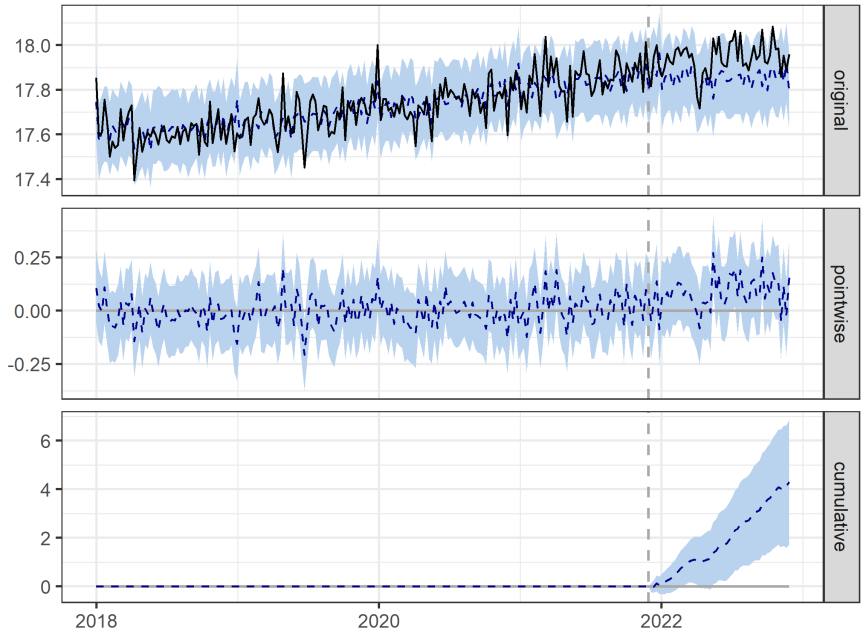

Figure C.6: Placebo Test: The Impact of ChatGPT on the the Automated Readability Index for Articles Authored by Nonnative English Speakers

Notes: This figure illustrates the placebo effect of ChatGPT's release on the Automated Readability Index (ARI) for articles authored exclusively by nonnative English speakers. or interpretation, refer to the figure note provided with Figure C.4.

D Smoothed Trends with Partially Linear Regression

In this analysis, instead of aggregating the data on a weekly basis, we use a predictive model that includes a flexible trend component for each article's score while controlling for subject fields. The partially linear regression model is specified as follows:

$$y_{it} = f(t) + \beta X_{it} + \epsilon_{it}, \tag{4}$$

where y_{it} represents the outcome variable for article i published on date t, X consists of dummy variables indicating the subject field of the article, and $f(t)$ is a smooth function of time (days) to capture underlying trends. The nonparametric component, $f(t)$, allows for flexible modeling of the time trend without imposing a strict parametric form, while the subject field indicators enter the model linearly.

The nonparametric component $f(t)$ is modeled as a penalized regression spline, which provides a flexible, data-driven way of estimating the effect of time on the outcome. We use a thin plate regression spline, as it adapts to the underlying data without requiring manual selection of knots. The smoothness of $f(t)$ is controlled by a penalty that prevents overfitting, automatically selected by the model to balance the trade-off between fit and smoothness. This allows $f(t)$ to capture non-linear trends over time while avoiding complexity.

By estimating this model, we allow the trend component to vary smoothly over time while controlling for field-specific effects through the parametric component, enabling us to visually inspect whether there are noticeable shifts in the time trends in the post-ChatGPT period. This approach avoids the need for aggregating the data and provides a clearer view of any potential changes in the composition of articles over time. Doing a non-aggregated estimation is more efficient than our baseline BSTS estimation, but this semi-parametric approach does not allow us to produce counterfactuals and changes in trends are only visually observable. Therefore, we use these results for the purpose of validating our overall conclusions.

In Figure D.1 there is a discernible change in the trend trajectory across all three groups, consistent with results in Section 4.1. The effect is most pronounced for nonnative-authored articles and smallest for native-authored articles. While this model does not generate counterfactuals, assuming a constant pre-ChatGPT trend, the pointwise estimated impacts remain qualitatively consistent with BSTS estimates. Therefore, we conclude that the observed impact on TTR is robust to controlling for subject field.

Figure D.2, which reports estimates for ARI, conveys similar conclusions to those in Section 4.2, with the most significant impact observed for nonnative-authored articles compared to the other groups. Assuming a constant pre-ChatGPT trend, the pointwise estimated impacts qualitatively align with the findings from the BSTS estimates.

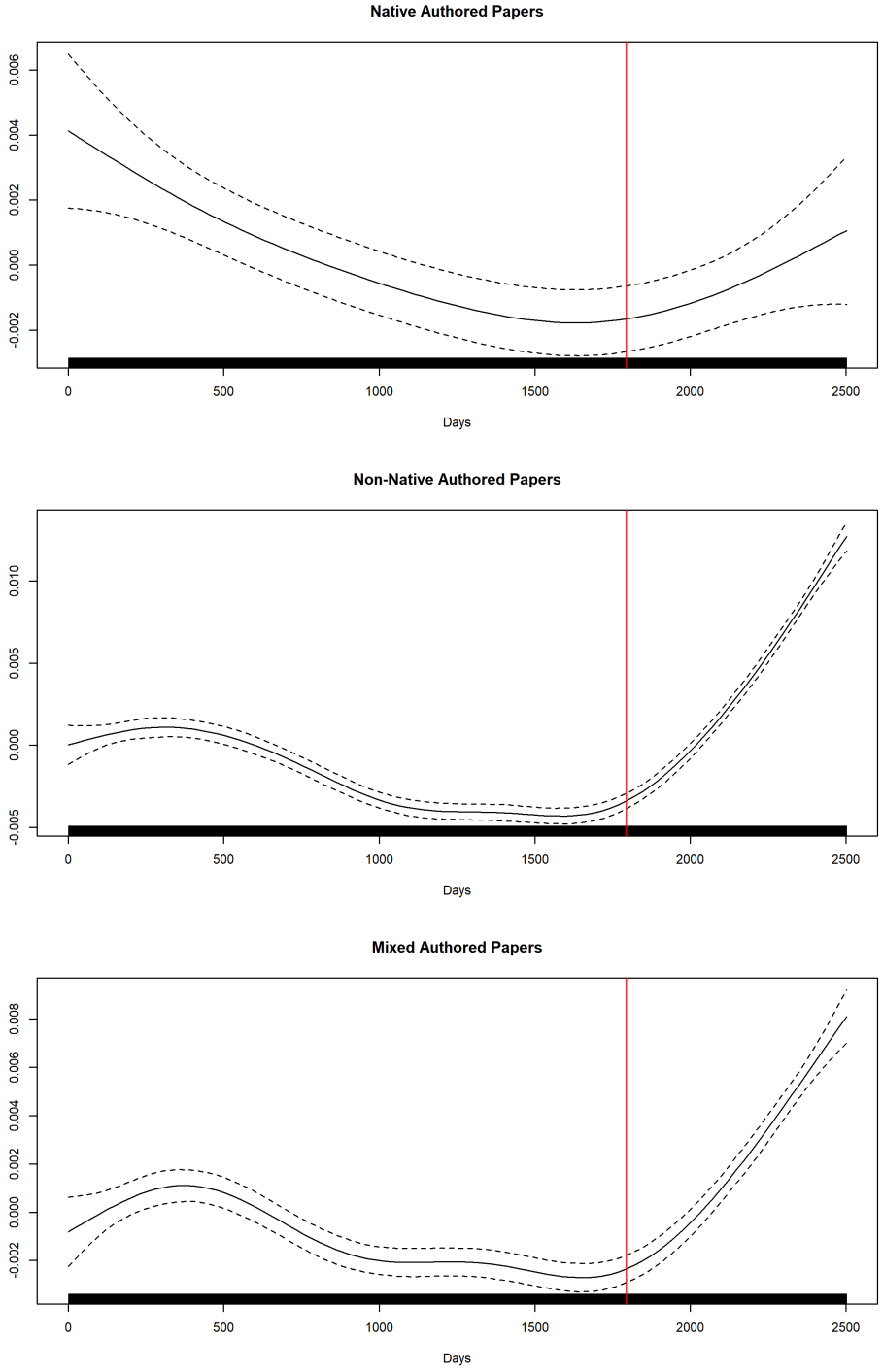

Figure D.1: TTR: Estimated Trend Terms for Articles Authored by Native, Mixed, and Nonnative English Speakers

Notes: The graphs illustrate the smoothed trends in the Type-Token Ratio (TTR), calculated from abstracts of articles published on arxiv.org from January 1, 2018, to November 14, 2024. TTR, which measures lexical diversity, is defined as the number of unique words (types) divided by the total number of words (tokens). The trends were estimated using partially linear regression to show the variations in lexical diversity over time. The top panel represents articles authored by native English speakers, the middle panel shows those authored by nonnative English speakers, and the bottom panel highlights articles with mixed authorship. Red vertical lines mark the release date of ChatGPT, our reference point for potential shifts in trends. The solid lines represent the smoothed trend, and the dashed lines indicate the 95% confidence intervals.

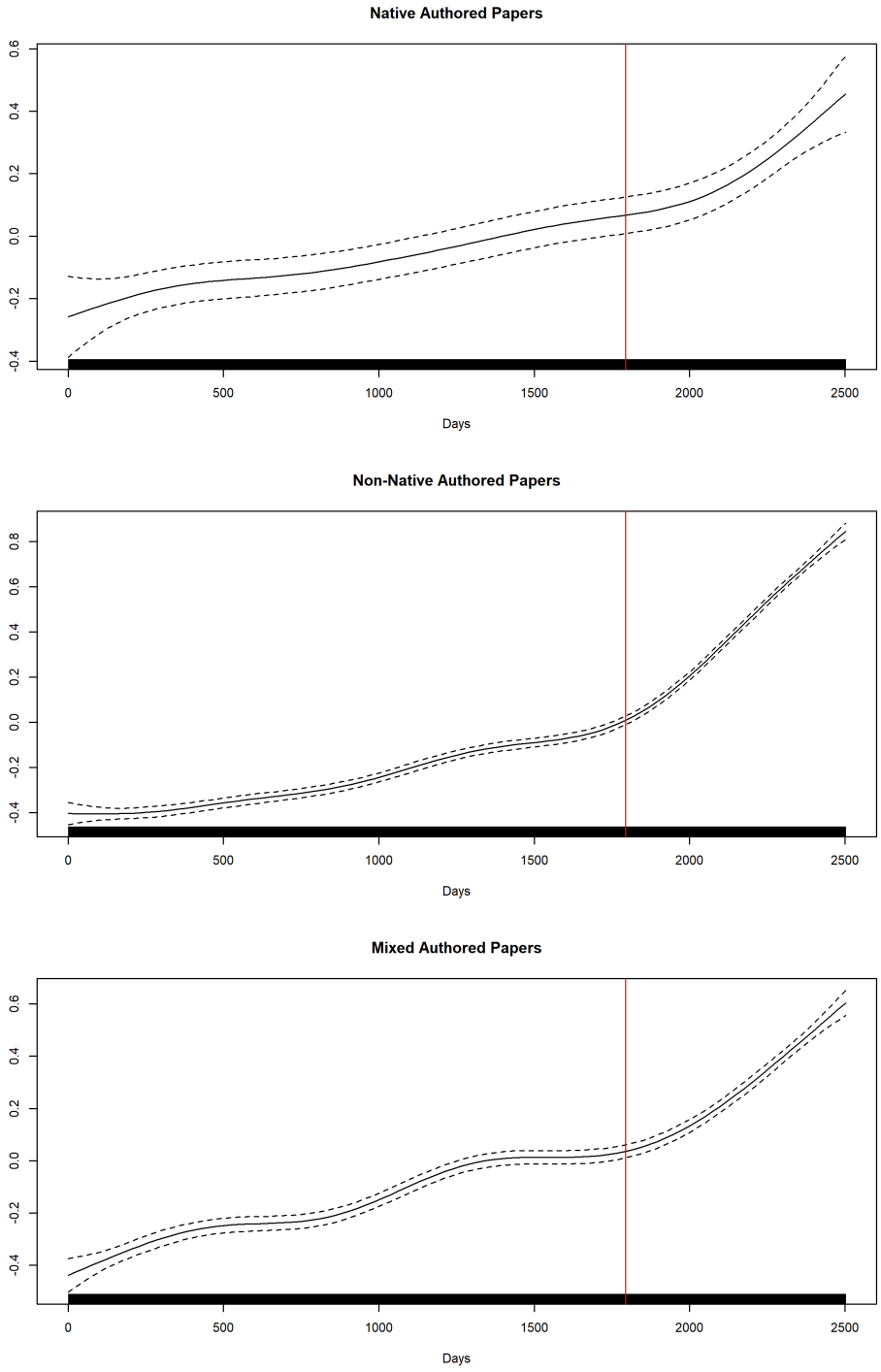

Figure D.2: Automated Readability Index: Estimated Trend Terms for Articles Authored by Native, Mixed, and Nonnative English Speakers

Notes: The graphs illustrate the smoothed trends in the Automated Readability Index (ARI), calculated from abstracts of articles published on arxiv.org from January 1, 2018, to November 14, 2024. ARI, which measures text readability and complexity (Smith & Senter, 1967). The top panel represents articles authored by native English speakers, the middle panel shows those authored by nonnative English speakers, and the bottom panel highlights articles with mixed authorship. Red vertical lines mark the release date of ChatGPT, our reference point for potential shifts in trends. The solid lines represent the smoothed trend, and the dashed lines indicate the 95% confidence intervals.

E Computer Science and Other Subjects

This section presents the BSTS estimation results for abstracts in our sample, categorized by subject area. The sample is divided into two subsamples: one consisting of computer science articles and the other comprising articles from all other subjects (biology, economics, electrical engineering and systems science, finance, mathematics, physics, and statistics).

Results are provided for both the Type-Token Ratio (TTR) and the Automated Readability Index (ARI). For each authorship group, we compare the findings for computer science articles with those for non-computer science articles.

- **Type-Token Ratio (TTR)**
  - **Figure E.1** shows the impact of ChatGPT on TTR for Computer Science articles authored by native speakers.
  - **Figure E.2** shows the impact of ChatGPT on TTR for Non-Computer Science articles authored by native speakers.
  - **Figure E.3** shows the impact of ChatGPT on TTR for Computer Science articles authored by both native and nonnative speakers.
  - **Figure E.4** shows the impact of ChatGPT on TTR for Non-Computer Science articles authored by both native and nonnative speakers.
  - **Figure E.5** shows the impact of ChatGPT on TTR for Computer Science articles authored by nonnative speakers.
  - **Figure E.6** shows the impact of ChatGPT on TTR for Non-Computer Science articles authored by nonnative speakers.
- **Automated Readability Index (ARI)**
  - **Figure E.7** shows the impact of ChatGPT on the ARI for Computer Science articles authored by native speakers.
  - **Figure E.8** shows the impact of ChatGPT on the ARI for Non-Computer Science articles authored by native speakers.
  - **Figure E.9** shows the impact of ChatGPT on the ARI for Computer Science articles authored by both native and nonnative speakers.
  - **Figure E.10** shows the impact of ChatGPT on the ARI for Non-Computer

Science articles authored by both native and nonnative speakers.

– **Figure E.11** shows the impact of ChatGPT on the ARI for Computer Science articles authored by nonnative speakers.

– **Figure E.12** shows the impact of ChatGPT on the ARI for Non-Computer Science articles authored by nonnative speakers.

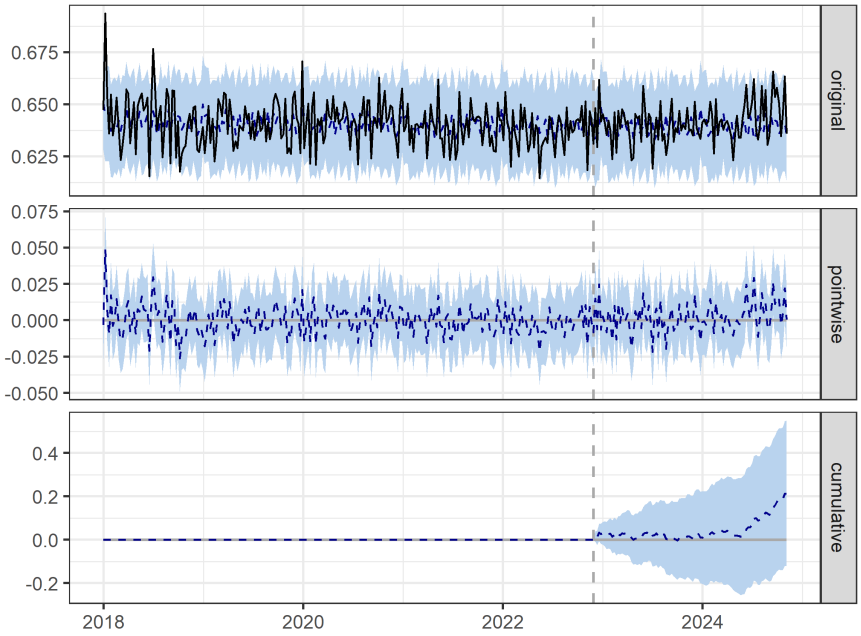

Figure E.1: The Impact of ChatGPT on TTR for Computer Science Articles Authored by Native Speakers

Notes: This figure, generated using the `CausalImpact` package (Brodersen et al., 2017), illustrates the effect of ChatGPT's release on the Type-Token Ratio (TTR) for articles authored exclusively by native English speakers in the field of computer science, calculated from abstracts of articles published on arxiv.org from January 1, 2018, to November 14, 2024. TTR, which measures lexical diversity, is defined as the number of unique words (types) divided by the total number of words (tokens). The top panel shows the observed data (solid black line) and the counterfactual predictions (blue line), with the shaded region representing the 95% credible intervals. The dashed vertical line marks the release date of ChatGPT, 30 November 2022. The middle panel displays the pointwise impact, calculated as the difference between observed and predicted values. The bottom panel visualizes the cumulative impact over time.

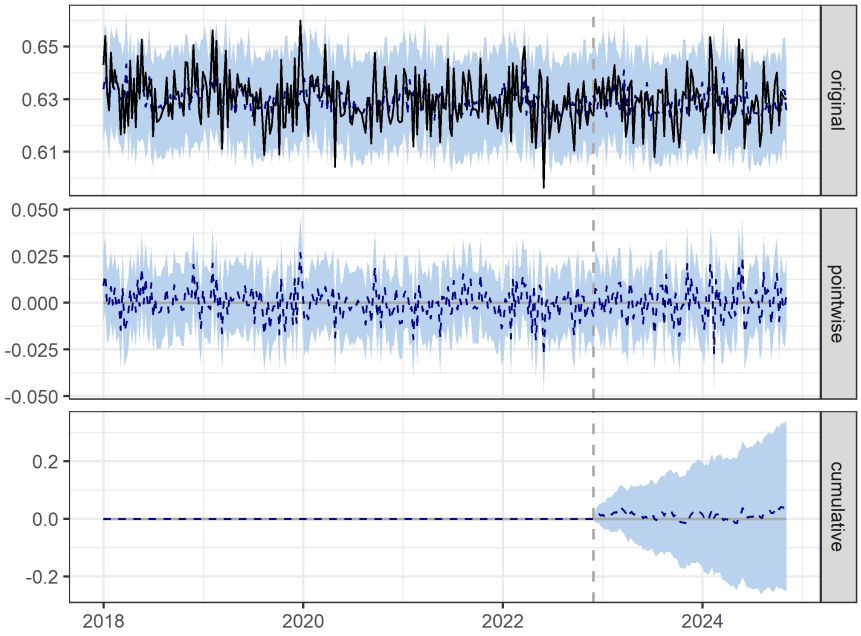

Figure E.2: The Impact of ChatGPT on TTR for Non-Computer Science Articles Authored by Native Speakers

Notes: This figure illustrates the effect of ChatGPT's release on the Type-Token Ratio (TTR) for articles authored exclusively by native English speakers in subjects other than computer science. The figure follows the same structure and interpretation as described in the figure note provided with Figure E.1.

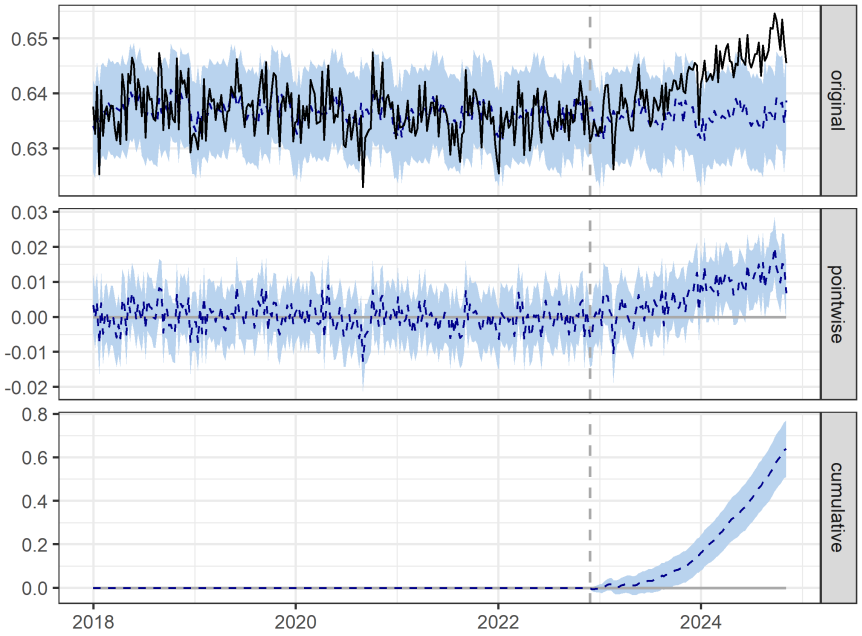

Figure E.3: The Impact of ChatGPT on TTR for Computer Science Articles Authored by Native and Nonnative Speakers

Notes: This figure illustrates the effect of ChatGPT's release on the Type-Token Ratio (TTR) for articles authored by native and nonnative English speakers in the field of computer science. The figure follows the same structure and interpretation as described in the figure note provided with Figure E.1.

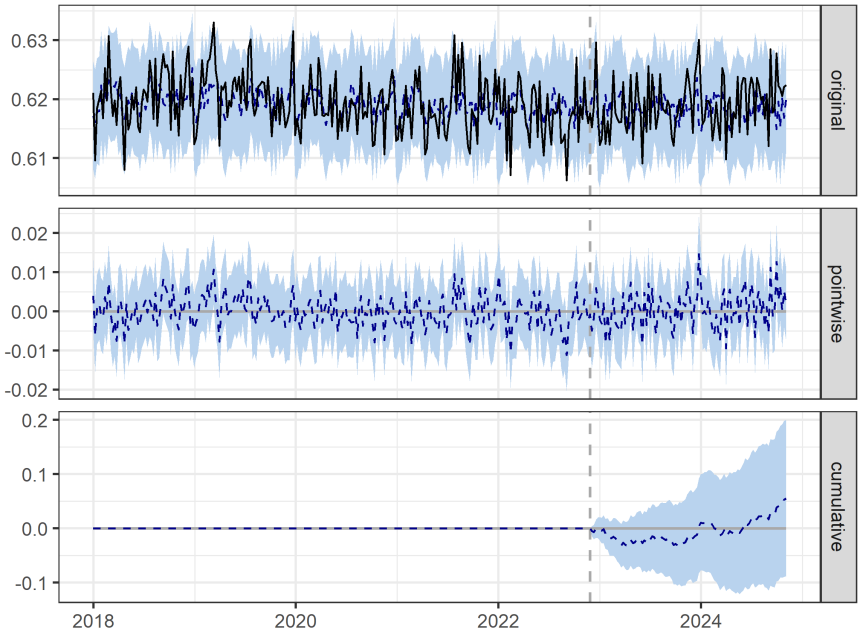

Figure E.4: The Impact of ChatGPT on TTR for Non-Computer Science Articles Authored by Native and Nonnative Speakers

Notes: This figure illustrates the effect of ChatGPT's release on the Type-Token Ratio (TTR) for articles authored by native and nonnative English speakers in subjects other than computer science. The figure follows the same structure and interpretation as described in the figure note provided with Figure E.1.

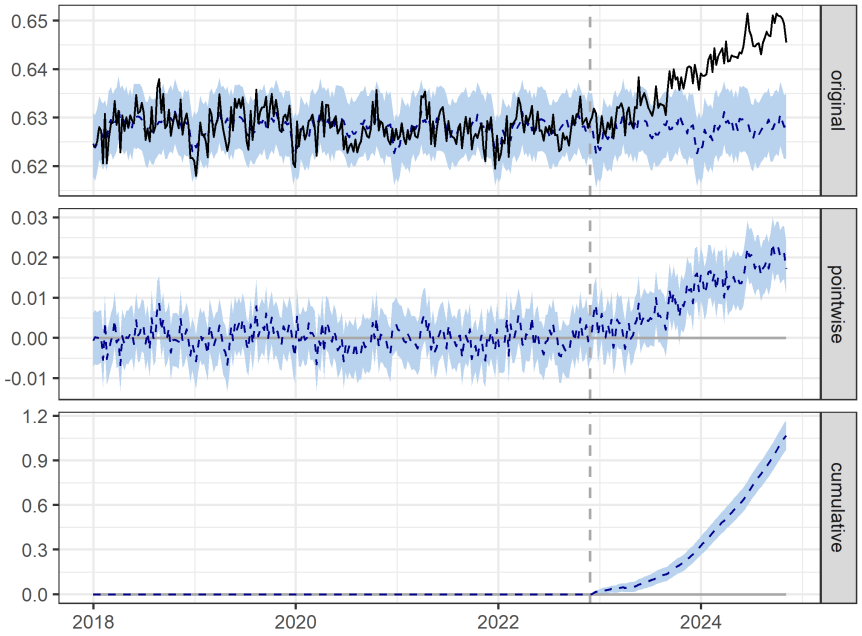

Figure E.5: The Impact of ChatGPT on TTR for Computer Science Articles Authored by Nonnative Speakers

Notes: This figure illustrates the effect of ChatGPT's release on the Type-Token Ratio (TTR) for articles authored exclusively by nonnative English speakers in the field of computer science. The figure follows the same structure and interpretation as described in the figure note provided with Figure E.1.

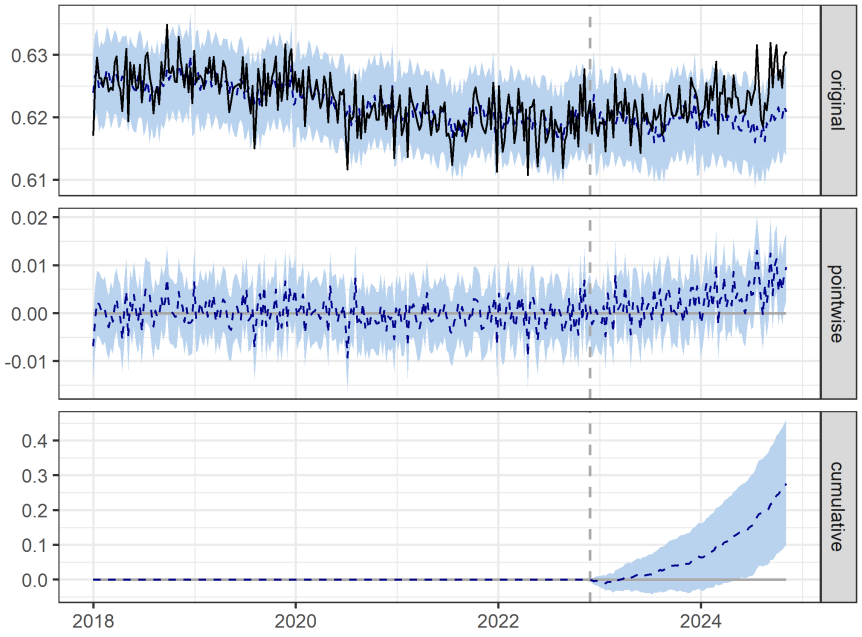

Figure E.6: The Impact of ChatGPT on TTR for Non-Computer Science Articles Authored by Nonnative Speakers

Notes: This figure illustrates the effect of ChatGPT's release on the Type-Token Ratio (TTR) for articles authored exclusively nonnative English speakers in subjects other than computer science. The figure follows the same structure and interpretation as described in the figure note provided with Figure E.1.

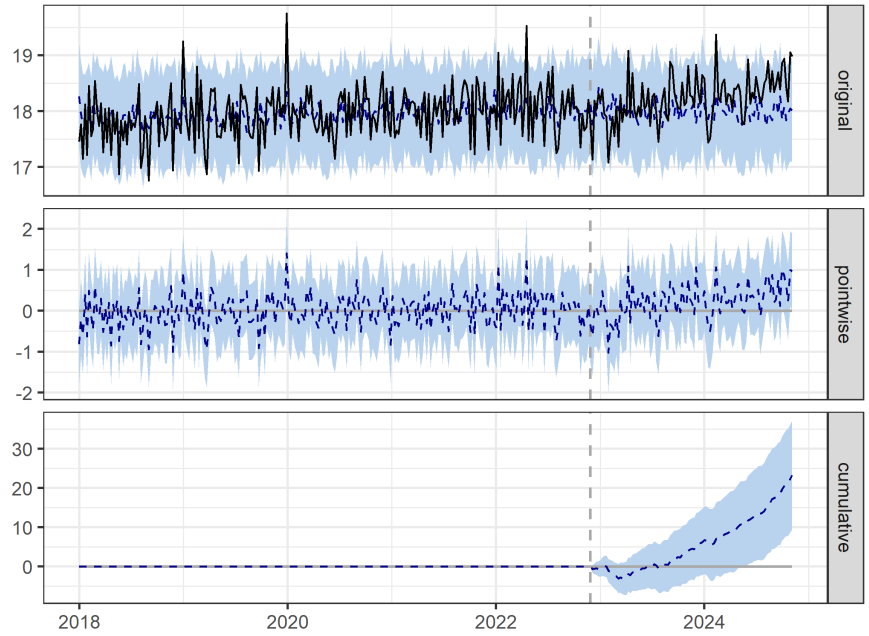

Figure E.7: The Impact of ChatGPT on the Automated Readability Index for Computer Science Articles Authored by Native Speakers

Notes: This figure, generated using the CausalImpact package (Brodersen et al., 2017), illustrates the effect of ChatGPT's release on the Automated Readability Index (ARI) for articles authored exclusively by native English speakers in the field of computer science, calculated from abstracts of articles published on arxiv.org from January 1, 2018, to November 14, 2024. ARI, which measures text readability and complexity (Smith & Senter, 1967). The top panel shows the observed data (solid black line) and the counterfactual predictions (blue line), with the shaded region representing the 95% credible intervals. The dashed vertical line marks the release date of ChatGPT, 30 November 2022. The middle panel displays the pointwise impact, calculated as the difference between observed and predicted values. The bottom panel visualizes the cumulative impact over time.

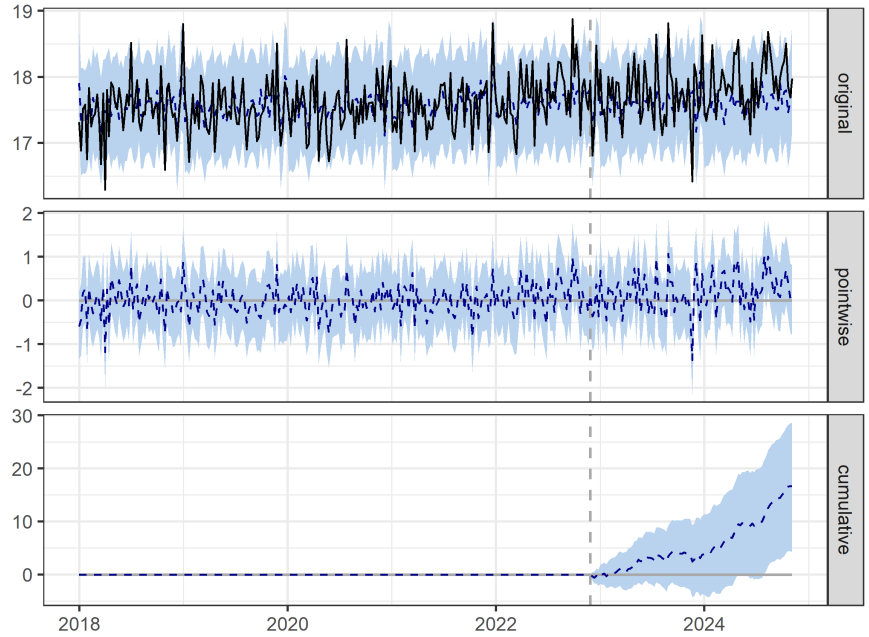

Figure E.8: The Impact of ChatGPT on the Automated Readability Index for Non-Computer Science Articles Authored by Native Speakers

Notes: This figure illustrates the effect of ChatGPT's release on the Automated Readability Index (ARI) for articles authored exclusively by native English speakers in subjects other than computer science. The figure follows the same structure and interpretation as described in the figure note provided with Figure E.7.

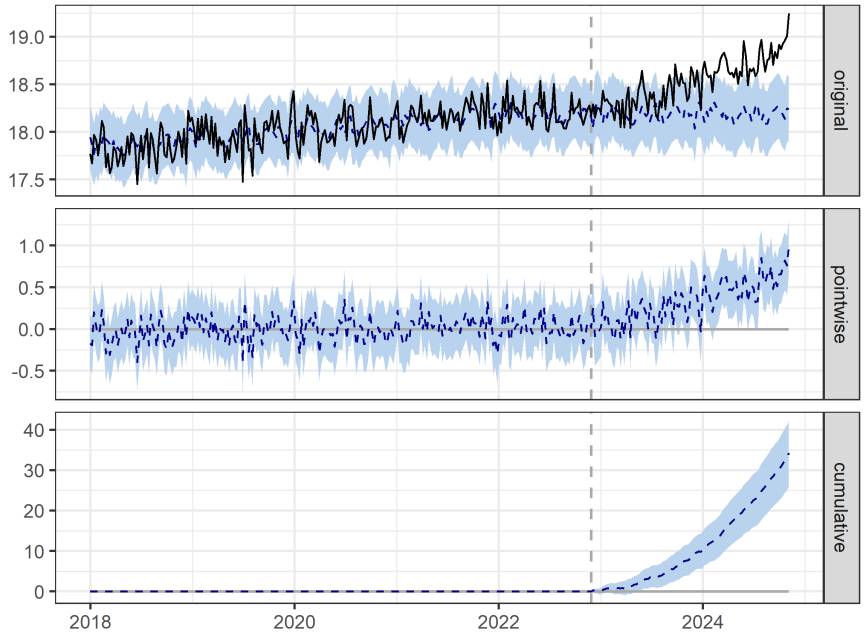

Figure E.9: The Impact of ChatGPT on the Automated Readability Index for Computer Science Articles Authored by Native and Nonnative Speakers

Notes: This figure illustrates the effect of ChatGPT's release on the Automated Readability Index (ARI) for articles authored by native and nonnative English speakers in the field of computer science. The figure follows the same structure and interpretation as described in the figure note provided with Figure E.7.

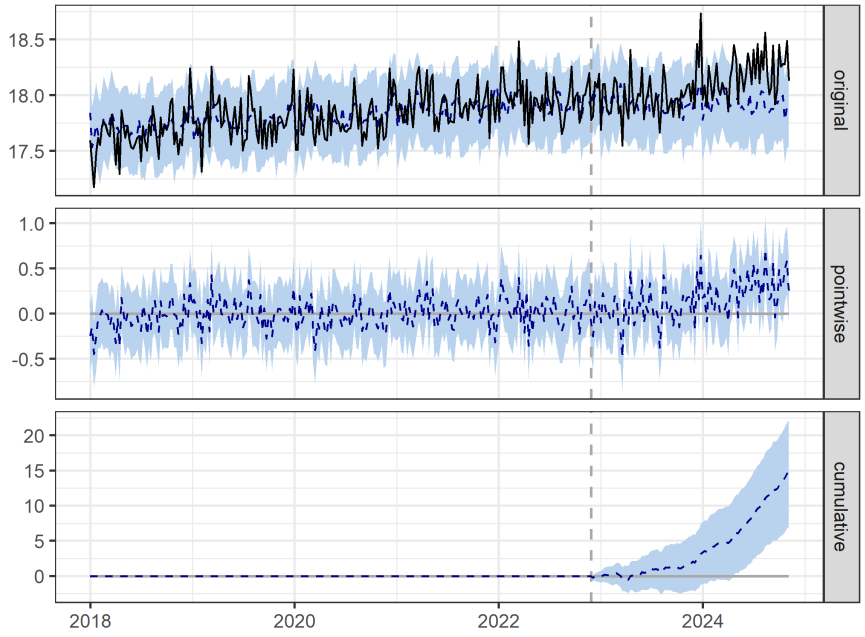

Figure E.10: The Impact of ChatGPT on the Automated Readability Index for Non-Computer Science Articles Authored by Native and Nonnative Speakers

Notes: This figure illustrates the effect of ChatGPT's release on the Automated Readability Index (ARI) for articles authored by native and nonnative English speakers in subjects other than computer science. The figure follows the same structure and interpretation as described in the figure note provided with Figure E.7.

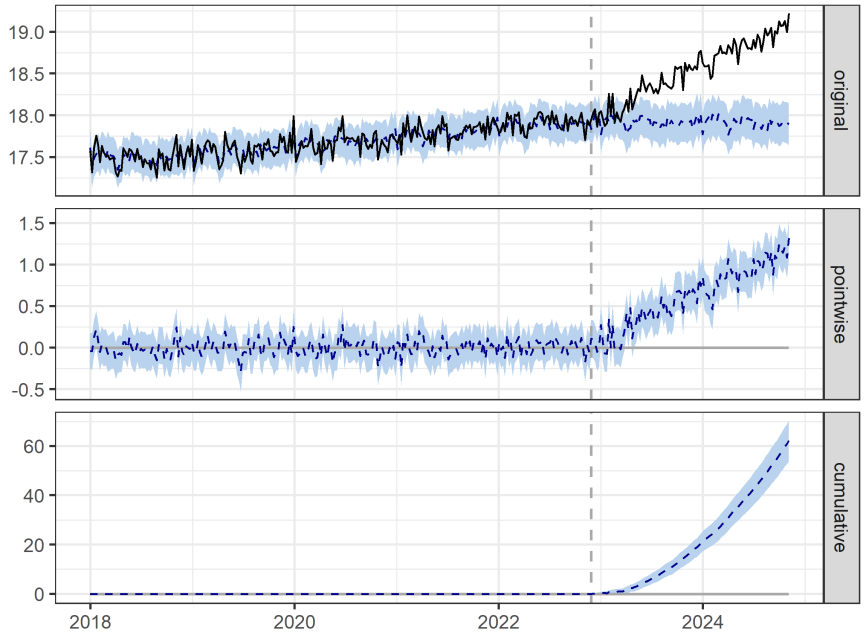

Figure E.11: The Impact of ChatGPT on the Automated Readability Index for Computer Science Articles Authored by Nonnative Speakers

Notes: This figure illustrates the effect of ChatGPT's release on the Automated Readability Index (ARI) for articles authored exclusively by nonnative English speakers in the field of computer science. The figure follows the same structure and interpretation as described in the figure note provided with Figure E.7.

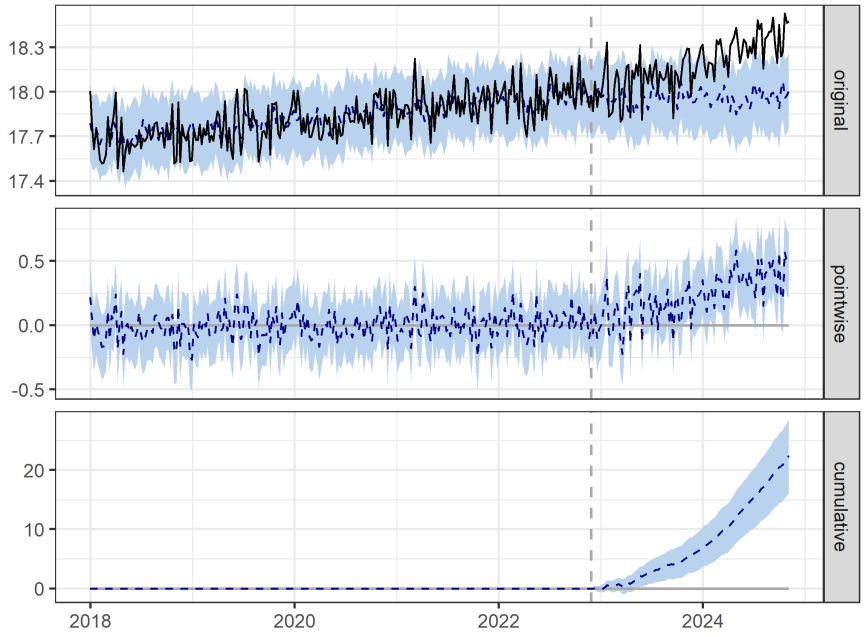

Figure E.12: The Impact of ChatGPT on the Automated Readability Index for Non-Computer Science Articles Authored by Nonnative Speakers

Notes: This figure illustrates the effect of ChatGPT's release on the Automated Readability Index (ARI) for articles authored exclusively by nonnative English speakers in subjects other than computer science. The figure follows the same structure and interpretation as described in the figure note provided with Figure E.7.