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

DISCUSSION PAPER SERIES

IZA DP No. 18062

The Behavioral Signature of GenAI in Scientific Communication

Nikos Askitas

AUGUST 2025

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 18062

The Behavioral Signature of GenAI in Scientific Communication

Nikos Askitas
IZA

AUGUST 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

The Behavioral Signature of GenAI in Scientific Communication*

We examine the uptake and measurable effects of GPT-assisted writing in economics working paper abstracts. Focusing on the IZA discussion paper series, we detect a significant stylistic shift following the public release of ChatGPT-3.5 in March 2023. This shift appears in core textual metrics—including mean word length, type-token ratio, and readability—and reflects growing alignment with machine-generated writing. While the release of ChatGPT constitutes an exogenous technological shock, adoption is endogenous: authors choose whether to incorporate AI assistance. To capture and estimate the magnitude of this behavioral response, we combine stylometric analysis, machine learning classification, and prompt-based similarity testing. Event-study regressions with fixed effects and placebo checks confirm that the observed shift is abrupt, persistent, and not attributable to pre-existing trends. A similarity experiment using OpenAI's API shows that post-ChatGPT abstracts more closely resemble their GPT-optimised counterparts than do pre-ChatGPT texts. A classifier trained on these variants achieves 97% accuracy and increasingly flags post-March 2023 abstracts as GPT-like. Rather than indicating wholesale substitution, our findings suggest selective human–AI augmentation in professional writing. The framework introduced here generalises to other settings where writing plays a central role—including resumes, job descriptions, legal briefs, research proposals, and software documentation.

JEL Classification: C55, C88, O33, C81, L86, J24

Keywords: GPT adoption, academic writing, text analysis, natural language processing (NLP), machine learning, event study, linguistic metrics, AI-assisted writing, diffusion of technology

Corresponding author:
Nikos Askitas
Institute of Labor Economics (IZA)
Schaumburg-Lippe-Strasse 5-9
53113 Bonn
Germany

E-mail: nikos@askitas.com

* Extensive use of the OpenAI API supported aspects of the research as well as the drafting and coding of this paper. Early versions of the code were developed in an exploratory and unstructured manner, then refined post hoc. All content and interpretations are the sole responsibility of the author. I thank participants at working seminars at LISER and the University of Bonn for feedback that improved the manuscript.

1 Introduction

The emergence of large language models (LLMs), particularly ChatGPT, has transformed the landscape of academic and professional writing. Academia—by its culture of experimentation and early adoption—is a natural testbed for studying this transformation. Researchers, especially in data-driven and technology-oriented fields, are both equipped and incentivized to explore new tools that improve clarity, efficiency, or polish. Studying how academic writing evolves in response to these tools offers a window into broader shifts in communication norms and professional expression under AI augmentation.

Recent evidence suggests that human–AI interactions can reshape perception and communication behaviors, even outside conscious awareness (Glickman and Sharot, 2025). Meanwhile, cautionary tails on GPT-derived summaries are emerging Alvarez et al. (2024); Doshi et al. (2024).

While only 5.3% of papers submitted after March 2023 explicitly mention generative AI tools like ChatGPT—based on parsed PDF content[1]—our behavioral estimates suggest much higher adoption rates, ranging from 16% to 19% depending on the method. This discrepancy highlights a fundamental information asymmetry: readers, reviewers, and editors cannot reliably infer AI use from disclosures alone.

Several forces likely contribute to this underreporting. First, incentives to disclose are weak, especially when AI assistance is perceived as minor or stigmatized. Second, normative ambiguity around acceptable use may discourage upfront acknowledgment. Finally, as recent work shows (Glickman and Sharot, 2025), even users themselves may be uncertain or uncomfortable about how to report their reliance on generative tools—reflecting deeper tensions between productivity gains and reputational risk.

Seen in this light, the gap between observed behavior and self-reported use is not a mere measurement artifact, but an early signal of institutional friction: generative AI is reshaping professional writing faster than editorial norms and disclosure standards can adjust.

This paper investigates whether and how GPT-assisted writing has changed the style of research abstracts in economics. We focus on the IZA Discussion Paper series, a long-running and timely corpus of pre-publication research. Our empirical strategy exploits the March 2023 release of ChatGPT-3.5—a widely accessible and improved LLM version—as a technological shock that creates plausibly exogenous variation in AI writing tool availability. Importantly, adoption remains endogenous: authors choose whether to incorporate AI into their writing process. This setup allows us to study behavioral uptake, rather than purely technological capacity.

We adopt a multi-pronged approach to isolate and characterise the resulting stylistic shift. Section 4 introduces core linguistic indicators—mean word length (MWL), type-token ratio (TTR), and several readability metrics—that show a measurable and abrupt break coinciding with ChatGPT's release. Event-study regressions in Section 5, which include rich controls for an exhaustive list of observables in our data (author count, abstract length, JEL codes, keyword density, page count, author gender, and email region), confirm that the shift is persistent and statistically significant. Section 7 establishes temporal specificity: we observe no such shift when running a placebo analysis around a fictitious 2018 cutoff.

To probe deeper, we simulate counterfactuals using the OpenAI API. In Section 6, we prompt GPT-3.5 to optimise each abstract—pre and post—while preserving content and keywords. If post-GPT texts are already shaped by AI, they should resemble these optimisations more closely than pre-GPT texts. Across compression ratio, lexical similarity, and distance metrics, we find strong asymmetry: GPT alters pre-GPT abstracts substantially more than post-GPT ones. In effect, post-GPT abstracts already approximate their AI-enhanced counterparts.

[1]This figure includes both papers about generative AI and those disclosing AI assistance. A targeted search for disclosures of GPT-assisted writing returned no matches.

To estimate the lower bound of real-world adoption, we introduce a placebo-stratified simulation. By artificially applying GPT optimization to subsets of pre-GPT abstracts and comparing effect sizes, we estimate that at least 16% of post-March 2023 abstracts show signs of real GPT use.

In Section 9, we develop an interpretable classifier trained to distinguish original vs. GPT-optimised abstracts. Its high accuracy (97%) on synthetic holdout samples allows us to apply it to real data. The classifier flags a growing share of recent texts as GPT-like, rising from a stable baseline of 13% to over 32% post-release. This pattern is consistent with our regression and simulation-based estimates of selective but nontrivial adoption.

Section 8 adds a panel regression across institutions and time, reinforcing the aggregate and persistent nature of the stylistic break. We conclude in Section 10 with implications for labor substitution, augmentation, and the evolving norms of academic communication.

Our findings suggest that generative AI is already influencing not just the tools of scholarship, but its tone. Rather than signalling wholesale substitution, the evidence points to selective augmentation: researchers appear to adopt AI assistance in a way that subtly but detectably changes how knowledge is presented.

2 Related Literature and Contribution

A rapidly expanding literature investigates how large language models (LLMs), particularly ChatGPT, are reshaping academic and professional writing. In economics, Walther and Dutordoir (2024) analyzes early stylistic shifts in finance publications following the emergence of LLMs. More broadly, Bao et al. (2025) and Lin et al. (2025) track changes in linguistic complexity and convergence toward GPT-like phrasing in academic preprints, suggesting a measurable stylistic influence of generative AI. These studies align with recent findings from psychology and cognitive science indicating that human–AI interaction can reshape communication behaviors, even outside conscious awareness (Glickman and Sharot, 2025).

Outside academia, Humlum and Vestergaard (2024) study ChatGPT adoption across occupational categories in Denmark, finding that use is widespread but moderated by workplace constraints and training barriers. Their work highlights the broader behavioral uptake of generative AI in professional settings, reinforcing the importance of examining LLM influence in high-skilled writing tasks, such as academic research. Other cautionary studies flag concerns about the effects of GPT-generated content on information accuracy, stylistic diversity, and user trust (Alvarez et al., 2024; Doshi et al., 2024).

Several studies document the prevalence of LLM use in scholarly output. Xu (2025) conduct a cross-disciplinary audit of AI acknowledgments in academic publishing, showing that over three-quarters of papers explicitly referencing LLM tools cite ChatGPT. Similarly, Liang et al. (2025) report widespread use of generative models across sectors, underscoring ChatGPT's dominant role. Unlike these self-reported usage studies, our approach infers adoption behaviorally and at scale, based on observable textual features.

From a methodological standpoint, our study complements the growing literature using text-as-data in economics (Gentzkow et al., 2019) and aligns with recent efforts to detect AI-generated content in academic writing (Gao et al., 2022; Korinek, 2023). Our empirical strategy draws on machine learning classification, event-study regressions, and counterfactual generation using the OpenAI API. We apply this toolkit to a curated corpus of economics working paper abstracts, enabling measurement of LLM influence over time.

Studies in the natural and social sciences have highlighted the implications of LLM adoption on authorship and publication practices. Stokel-Walker (2023) and van Dis et al. (2023) express concern about detection and disclosure, while Karpf (2023) discusses the erosion of traditional academic assessments under LLM-enhanced writing. Our findings are related to these debates, offering empirical benchmarks for LLM influence in high-stakes, formal text.

We also contribute to work on readability and linguistic clarity. Using standard metrics such as the Flesch Reading Ease score (Flesch, 1948), SMOG (McLaughlin, 1969), and ARI (Senter and Smith, 1967), we find that readability declines in the post-GPT era. Abstracts become more mechanically structured and less readable according to established indices. This stands in contrast to findings such as those in Imperial and Madabushi (2023), who report that ChatGPT output often scores higher on human-aligned readability metrics. Our results suggest that GPT-assisted writing may sacrifice clarity and fluency in favor of structural regularity and formalism.

Our main contributions are empirical and methodological. We (i) develop a multi-pronged approach to detect LLM use in academic writing; (ii) validate the method through placebo and similarity-based falsification exercises; (iii) estimate lower bounds on adoption; and (iv) show that LLM-driven style shifts correlate with institutional and temporal patterns in a way consistent with real adoption. Our framework offers a generalizable blueprint for tracking LLM influence in other domains such as legal writing, code, or job market materials.

3 Data and Descriptives

We use data from the IZA Discussion Paper Series archive. These metadata are ingested by the RePEc service and contain detailed information on each paper, including title, authorship, abstract, JEL classification, keywords, and document length. We keep data from 2010 onwards as this is when monthly output stabilises.

Metadata records are stored in a simple structured RDF file format. A typical metadata entry is illustrated in Listing 1 below:

```
Template-type: ReDIF-Paper 1.0
Author-Name: Askitas, Nikos
Author-Email: askitas@iza.org
Author-Workplace-Name: IZA
Title: A Hands-on Machine Learning Primer ...
Abstract: This paper addresses the steep learning curve in
Machine Learning faced by non computer scientists, particularly
social scientists, stemming from
...
The objective of this primer is to equip readers with a solid
elementary comprehension of first principles and fire some
trailblazers to the forefront of AI and causal machine learning.
Classification-JEL: C01, C87, C00, C60
Keywords: machine learning  ... universal approximation theorem
Length: 29 pages
Creation-Date: 202405
Number: 17014
File-URL: https://docs.iza.org/dp17014.pdf
File-Format: application/pdf
Handle: RePEc:iza:izadps:dp17014
```

Listing 1: Example RePEc metadata record, In bold the keys used in this paper

We first provide some long-term descriptive trends of our data. In Figure 1 we show the monthly output in time, mostly close to 70 papers per month with the exception of the COVID crisis spike. In Figure 3 we observe a long term convergence to three authors per paper on average while Figure 2 shows a long term trend towards longer papers with a rapid reversion towards the mean starting towards the end of 2023 (which might or might no be GPT related). Figure 4 shows that about 70% of authors are male with a mild long term trend towards more female authors. Figures 5 and 7 illustrate the evolution of average monthly number of JEL codes and keywords, respectively, per paper.

Finally, our key motivating graphs (Figures 8 and 9). Figure 8 shows a significant deviation of mean word length (MWL) and its variance around the ChatGPT-3.5 release date (March

2023). MWL deviates up to three sigmas from its long term stable mean. This is consistent with ChatGPT's preference for longer, more formal word choices (e.g., *utilize* instead of *use*, or *demonstrate* instead of *show*). A different view on the same data is offered by the histograms of Figure 9. We observe that post-GPT the node shifts right, kurtosis becomes flatter and the tails fatter indicating heterogeneous adoption.

Before we proceed investigating more systematically motivated by these plots we take some time to introduce the linguistic metric we will be using in the rest of the paper. Readers with familiarity with these metrics may skip this section.

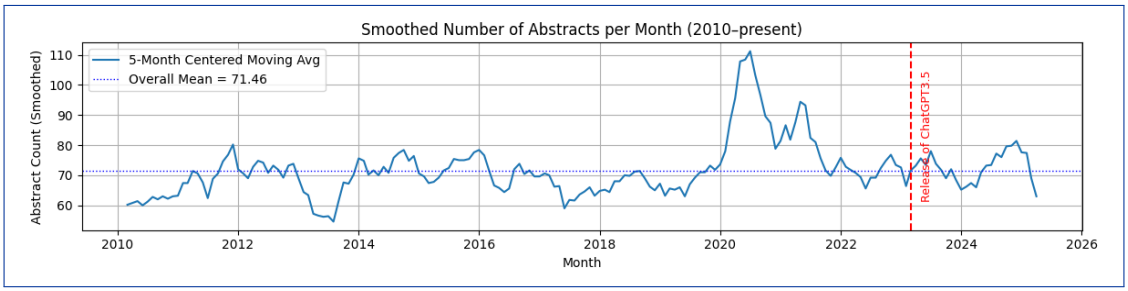

Figure 1: Long-term trends in monthly output.

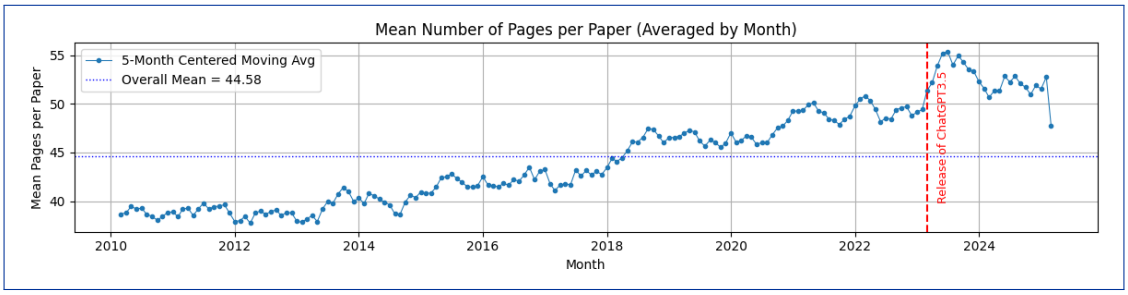

Figure 2: Long term trend towards longer papers.

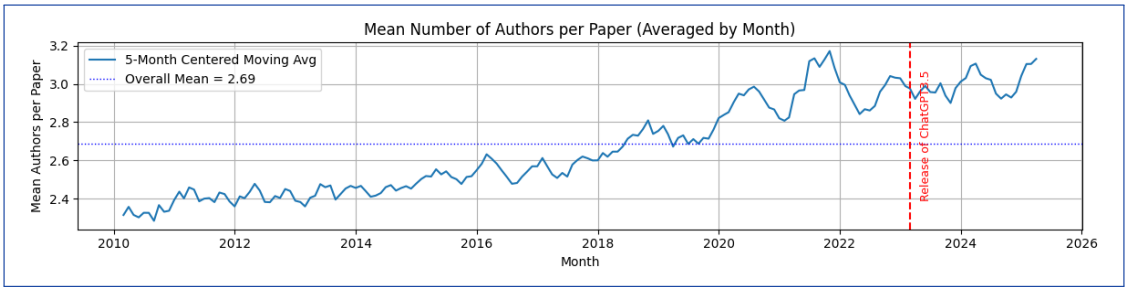

Figure 3: Long term convergence to mostly 3-author papers.

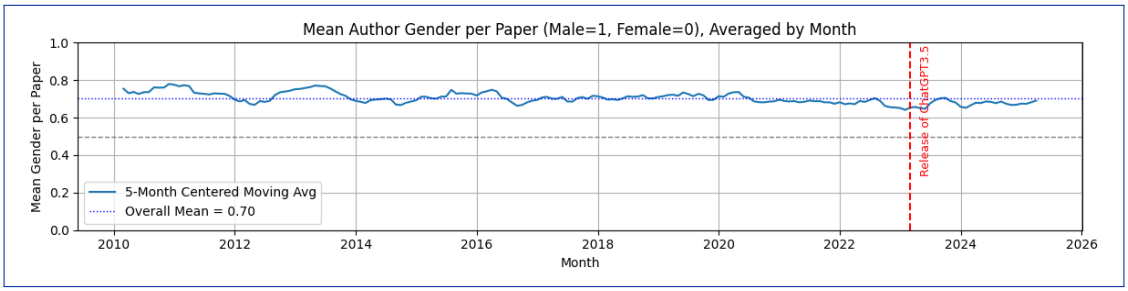

Figure 4: Authorship is male dominated with a mild long term trend towards more female authors

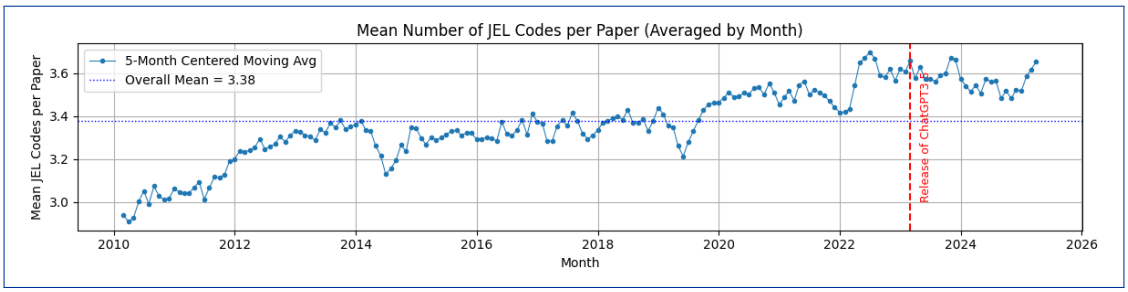

Figure 5: JEL code trends

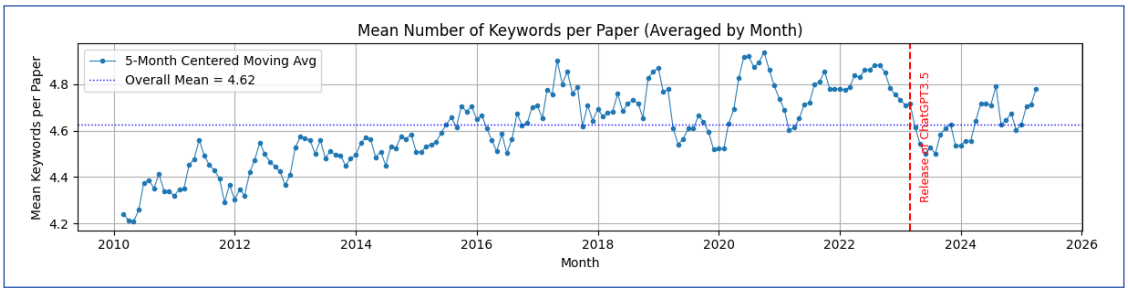

Figure 6: Trends of supplied keywords

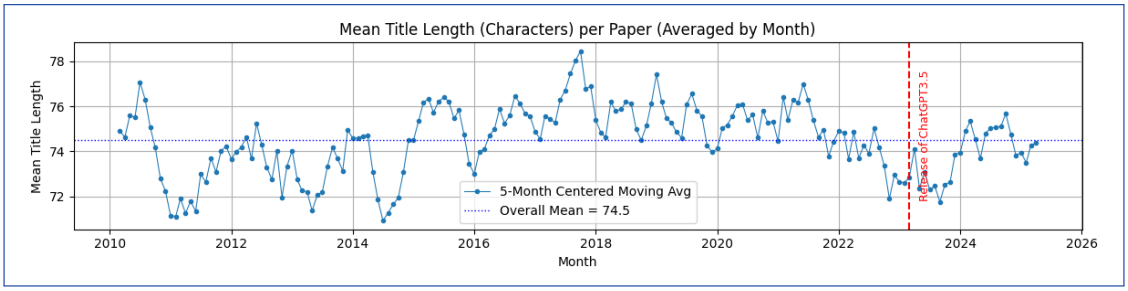

Figure 7: Trends of title lengths

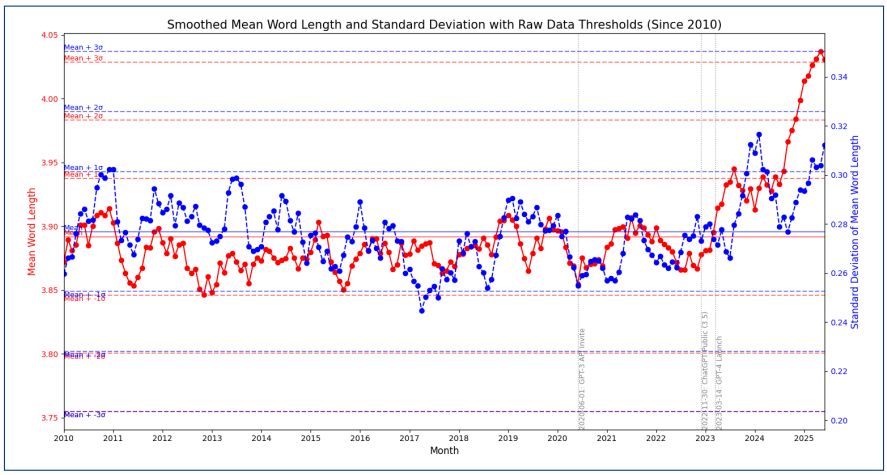

Figure 8: Monthly mean word length (MWL) and variance with ChatGPT-3.5 cutoff indicated.

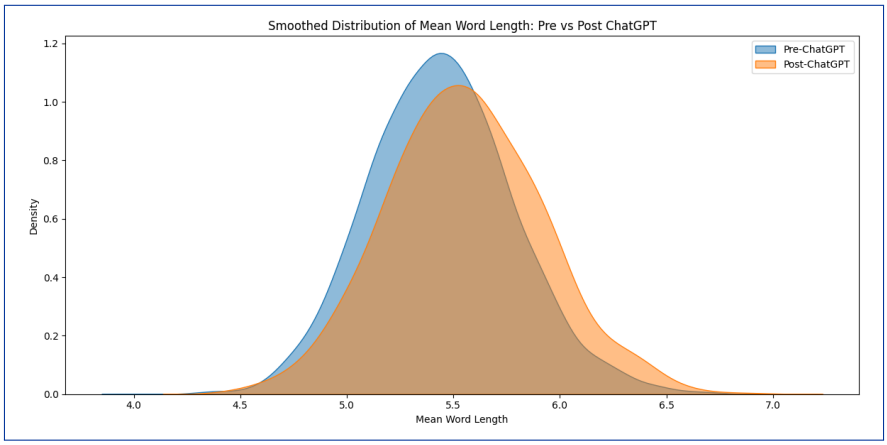

Figure 9: A histogram of mean word length (MWL) for pre- and post- ChatGPT-3.5 abstracts

4 Metrics

We discuss and explain the metrics we will be using in our analysis so that the reader can consult this section as needed.

We use the **Mean Word Length (MWL)**, which is simply the average number of characters per word in an abstract. It captures the tendency toward using shorter versus longer words. A rise in MWL can reflect a preference for more formal or complex word choices (e.g., using *utilize* instead of *use*). The **Type-Token Ratio (TTR)** is defined as the ratio of unique words (types) to the total number of words (tokens) in an abstract:

$$\text{TTR} = \frac{\text{\# unique words}}{\text{\# total words}}.$$

This ratio is always less than or equal to 1, reaching 1 only when every word is used once. A higher TTR reflects greater lexical diversity, while a lower TTR suggests more repetition. We use three reading ease indices. The **Flesch Reading Ease Score** is a classic readability metric computed from sentence length and syllable count:

$$\text{Flesch} = 206.835 - 1.015\left(\frac{\text{words}}{\text{sentences}}\right) - 84.6\left(\frac{\text{syllables}}{\text{words}}\right).$$

Higher values indicate simpler, more accessible texts, while lower values suggest denser or more complex writing.The **SMOG Index**, estimates the years of education required to understand a piece of writing based on the number of polysyllabic words. The **Automated Readability Index (ARI)**, uses character and word counts per sentence to approximate reading difficulty.

Finally we use three text similarity indices. The **Jaccard Similarity** measures the overlap between two sets of words (or other elements):

$$\text{Jaccard}(A, B) = \frac{|A \cap B|}{|A \cup B|}.$$

It ranges from 0 (no shared words) to 1 (identical sets). We use this metric to quantify the similarity between original and GPT-optimised abstracts. The **Compression Ratio** measures how well an abstract can be compressed using algorithms such as gzip:

$$\text{Compression Ratio} = \frac{\text{compressed size}}{\text{original size}}.$$

Highly repetitive texts compress more easily, resulting in a lower ratio, while more complex or diverse texts compress less. The **Levenshtein Distance** is a string-edit metric defined as the minimum number of single-character edits (insertions, deletions, or substitutions) required to transform one text into another. Normalizing by text length gives a relative measure of difference. A smaller distance indicates higher similarity.

Together, these metrics allow us to capture various aspects of abstract style and complexity, ranging from lexical diversity (TTR) and word choice (MWL) to readability (Flesch) and textual similarity (Jaccard, Levenshtein), as well as structural redundancy (compression ratio).

5 Regression Analysis

We estimate the following regression model for each text metric (MWL, TTR, Flesch):

$$Y_i = \alpha + \beta \cdot \text{PostGPT}_i + \gamma \cdot t_i + X_i + \varepsilon_i,$$

where Y_i denotes one of the text metrics—Mean Word Length (MWL), Type-Token Ratio (TTR), or Flesch Reading Ease—for paper i. The variable PostGPT_i is a dummy equal to 1

for papers released after March 2023, while t_i is a linear time trend. The vector X_i collects a set of control variables, including abstract length (log of character count), number of authors, number of keywords, number of JEL codes, number of pages, average author gender, as well as language-group (derived from the Top Level Domain of the author email addresses) and JEL code Letter fixed effects.

We find that β is positive and significant for MWL and TTR, indicating that abstracts became longer and more lexically diverse after the release of ChatGPT. Conversely, β is negative for the Flesch Reading Ease score, pointing to an increase in text complexity (lower readability), consistent with the idea that AI-assisted text generation leads to more sophisticated wording and longer average words.

Table 1: Effect of Post-ChatGPT on Text Metrics

	MWL	TTR	Flesch
PostGPT (β)	0.089***	0.0091***	-2.45***
	(0.011)	(0.002)	(0.35)
Controls	Yes	Yes	Yes
Observations	16,201	16,201	16,201

Notes: Each column reports the estimated coefficient of the PostGPT dummy (β) from separate regressions of the indicated text metric on PostGPT, a linear time trend, and controls X_i (log character length, number of authors, keywords, JEL codes, pages, average gender, language groups, and JEL fixed effects). Robust standard errors in parentheses.
*** $p < 0.01$.

To give the reader an understanding of the magnitude of a coefficient of 0.089 for MWL, note that with a pre-GPT MWL of 5.438, a 200-word abstract would on average be about 18 characters longer post-GPT. This is roughly equivalent to replacing about three shorter words such as "show" with longer alternatives like "demonstrate" which is significant for such a small and highly stylised document type.

6 GPT Optimization Experiment

Our hypothesis is that GPT-assisted abstracts written *after* the release of ChatGPT (post-March 2023) will be **more similar to their original versions** than those written before this cutoff. To test this, we collected 100 pre-GPT and 100 post-GPT abstracts and optimised them using GPT across three prompt styles (conservative, neutral, encouraging), repeating the experiment 10 times (see Table 2).

Since generative AI is a stochastic (essentially auto-regressive) process we expect GPT to alter even its own output so we can only hope to measure differences between post- and pre-GPT abstracts in the similarity of their originals with their treated versions along some similarity metrics. Thus, for MWL and TTR we estimate:

$$\Delta Y_i = \alpha + \beta \cdot \text{PostGPT}_i + X_i + \varepsilon_i,$$

while for Levenshtein, Jaccard, and Compression (which are already binary operators) we use:

$$Y_i = \alpha + \beta \cdot \text{PostGPT}_i + X_i + \varepsilon_i,$$

where the controls X_i are the same as in Section 5.

We perform pooled regression across our results to assess their statistical significance. Table 3 reports the estimated coefficients $\hat{\beta}$ for each metric and prompt style, with standard errors in parentheses.

Prompt Type	Prompt Text
Conservative	"Please review the following research abstract and optimize it for clarity and readability ONLY if changes are clearly necessary. Do not make stylistic edits if the abstract is already effective. Words from the keyword list must not be changed. Here is the abstract:"
Neutral	"Optimize the abstract only if improvements in clarity, structure, or flow are needed. Leave the abstract unchanged if already clear and concise. Do not alter any words that appear in the keyword list. Here is the abstract:"
Encouraging	"Please carefully revise the following abstract for clarity and precision, but ONLY if you believe improvements are warranted. Preserve all words found in the keyword list without modification. Here is the abstract:"

Table 2: Prompts used for GPT optimization of abstracts. All optimizations were performed with `gpt-3.5-turbo` (which is what authors are likely to have used at the time) using a temperature of 0.7, chosen to allow moderate variation in rewriting while maintaining consistency and clarity.

The regression results indicate that GPT-optimised abstracts are more similar to their original versions in the post-GPT period across all of our metrics. The positive coefficients for compression show that the ratio of GPT-optimised to original text size is larger post-GPT, suggesting that fewer structural adjustments are required by GPT, as the original abstracts already resemble GPT's style. Similarly, the negative coefficients for ΔMWL and ΔTTR indicate that differences in mean word length and lexical diversity between GPT rewrites and the originals have diminished, pointing to a closer alignment in word choice and vocabulary richness. The positive Jaccard coefficients confirm that post-GPT abstracts share more unique words with their GPT-optimised counterparts, while the negative Levenshtein coefficients show that fewer character-level edits are needed to go from the original to the GPT-optimised version for post-GPT abstracts, reinforcing the convergence of human writing and GPT-generated phrasing in the post-GPT era.

Metric	Conservative	Neutral	Encouraging	All Prompts
Panel A: Point Estimates				
Compression	0.026***	0.028***	0.023***	0.026***
ΔMWL	-0.111***	-0.111***	-0.110***	-0.111***
ΔTTR	-0.015***	-0.015***	-0.014***	-0.015***
Jaccard	0.018***	0.017***	0.018***	0.018***
Levenshtein	-24.030***	-17.532*	-15.545**	-19.036***
Panel B: Standard Errors				
	(0.006)	(0.009)	(0.006)	(0.005)
	(0.012)	(0.014)	(0.012)	(0.007)
	(0.002)	(0.003)	(0.002)	(0.001)
	(0.003)	(0.005)	(0.003)	(0.002)
	(7.415)	(9.318)	(7.438)	(4.743)

Notes: *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

Table 3: Impact of Post-GPT on GPT Optimization Similarity Metrics

7 Placebo Experiment and Adoption Calibration

To validate the sensitivity of our regression framework, we conducted placebo experiments with a pseudo-cutoff date in March 2018 and GPT-optimised abstracts as artificial treatments restricting to abstracts from 2010 to 2020.

Table 4 shows the results for the placebo experiment for treated and untreated abstracts post the placebo cutoff. For the original abstracts, the *placebo_post* coefficient is essentially zero, indicating no spurious time trend. In contrast, when all post-cutoff abstracts are replaced with their optimised GPT versions, we obtain a large coefficient of approximately 0.513, reflecting the stylistic shift caused by GPT optimization:

$$\text{MWL}_i = \alpha + \beta\, \text{placebo\_post}_i + X_i + \varepsilon_i, \tag{1}$$

where X_i denotes the same controls as in our main specification, but only α and β are reported here.

	Real	Placebo
const	4.9342***	4.9337***
	(0.0796)	(0.0869)
placebo_dummy	0.0001	0.5129***
	(0.0081)	(0.0136)

Table 4: Placebo Regression Results (Cutoff vs. Treated Placebo). All additional controls X_i have been dropped for simplicity; only the constant and placebo dummy are shown.

When we performed a random placebo experiment by randomly replacing abstracts (across all periods) with GPT-optimised versions, the average coefficient remains close to 0.513, confirming that the stylistic shift is unrelated to the time dimension.

We further simulated partial GPT adoption by replacing only a fraction of the post-cutoff abstracts within each month. Averaging over 10 runs for each adoption rate, we find that a 16% placebo adoption produces a coefficient of $\beta \approx 0.0899$, matching our real cutoff estimate.

These findings demonstrate that GPT optimisation introduces a systematic text-style shift consistently detected by our regression framework. The difference in coefficient magnitudes between placebo and real experiments arises because our placebo uses as-is GPT outputs, whereas human authors typically edit and "humanise" ChatGPT content. This adoption estimate provides therefore a conservative lower bound on the adoption rate in reality[2].

8 Panel Regression Analysis

We now aggregate our data on author affiliation level mentioned in the metadata to construct a panel. An abstract might be assigned to more than one institution for multi-author papers. We estimate the impact of the ChatGPT launch (March 2023) on abstract writing by running event-study regressions on two key textual metrics: *type-token ratio* (TTR) and *mean word length* (MWL). Our regressions use institution-level monthly panels and control for various characteristics of the papers. Our model specification is as follows:

$$Y_{i,t} - \bar{Y}_i = \sum_{k \neq -1,0} \beta_k D_k(t) + X_{i,t} + \epsilon_{i,t}, \tag{2}$$

where:

[2] An important caveat is that we are here just optimising abstracts. Real authors might ask ChatGPT to produce an abstract from an Introduction (which ChatGPT itself contributed to) then human-edit and re-optimise in several rounds etc.

- $D_k(t) = 1$ if $t - t_0 = k$, and 0 otherwise.
- i indexes institutions and t indexes months.
- t_0 is March 2023 (the ChatGPT event).
- $D_k(t) = 1$ if $t - t_0 = k$, and 0 otherwise, are event-time dummies for months $k \in [-12, +23]$, excluding $k = -1$ and $k = 0$ as reference periods.
- $X_{i,t}$ denotes control variables: Number of authors, Abstract character length, Number of JEL codes, Number of keywords, Number of pages, Language Family proxied by means of the Top Level Domain of the author email addresses and JEL code letters.

For robustness, we also estimate a specification with a single post-treatment dummy:

$$Y_{i,t} - \bar{Y}_i = \alpha + \delta \, \text{Post}_t + \theta C_t + X_{i,t} + \epsilon_{i,t}, \tag{3}$$

where $\text{Post}_t = 1$ for $t \geq t_0$ and C_t are month fixed effects. Clustering of standard errors is performed at the institution level. Notice that by demeaning the outcome variable we remove time-invariant differences among institutions.

Table 5 reports the estimated coefficients for the event-study dummies and the post-treatment effect δ. The TTR regression exhibits a significant post-event increase of approximately 0.0167 (p-value < 0.01) relative to the pre-treatment baseline. Similarly, the MWL regression finds a post-treatment increase of 0.0495 (p-value < 0.05).

Variable	TTR (demeaned)	MWL (demeaned)
Post (March 2023 onward)	0.0167***	0.0495**
	(0.0044)	(0.0255)
Controls	Yes	Yes
Institution FE	Yes	Yes
Time FE	Yes	Yes
Observations	3,986	3,986
R^2	0.406	0.067

Notes: Standard errors clustered at the institution level.

Table 5: Panel Regression Results (Event Study)

Figures 10 and 11 plot the β_k coefficients from the event-study specification. TTR exhibits a noticeable upward shift post-March 2023, while MWL shows a delayed but statistically significant increase over the following months[3].

[3]In these and later regressions, event-time coefficients exhibit some noisy variation but do not meaningfully affect our overall results. The reader should bear in mind that this is a rapidly evolving field. While we define our main treatment onset as March 2023—the public release of GPT-3.5 via ChatGPT—earlier forms of access existed, including the GPT-3 API invitation program launched in June 2020. Moreover, further significant upgrades occurred after our cutoff, including the release of GPT-4 in March 2023, GPT-4 Turbo in November 2023, and GPT-4o in May 2024. These innovations, and those of competing products, likely triggered new waves of experimentation and adoption across the author population, complicating any attempt to isolate a single treatment point.

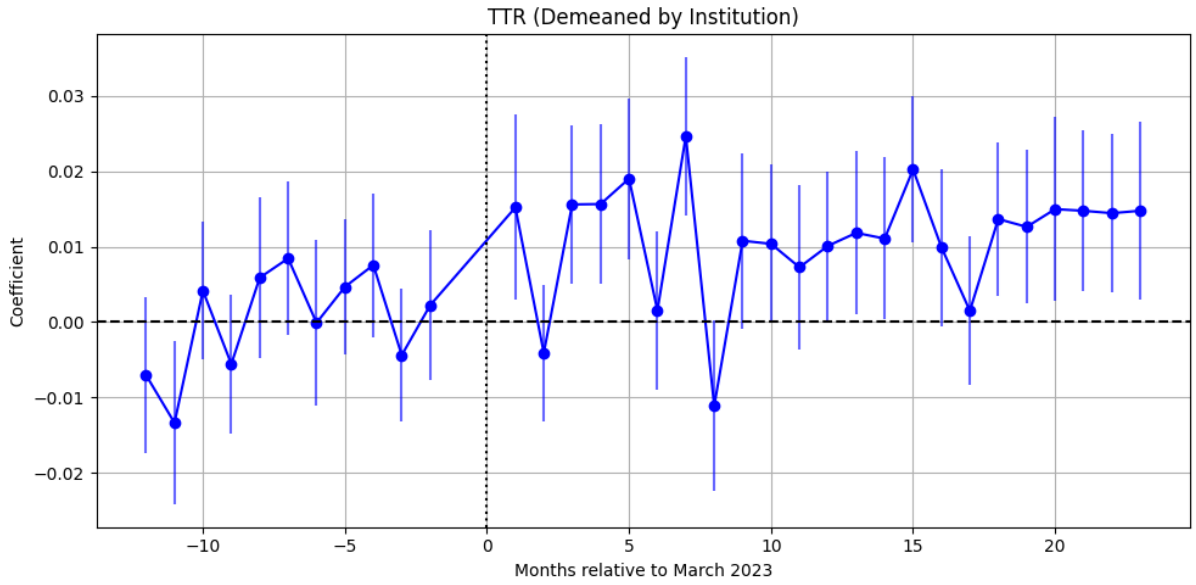

Figure 10: Event study estimates of type-token ration (institution-demeaned) with 95% confidence intervals around March 2023 (month 0). Controls mirror earlier models; institution fixed characteristics like language/TLD group absorbed via demeaning.

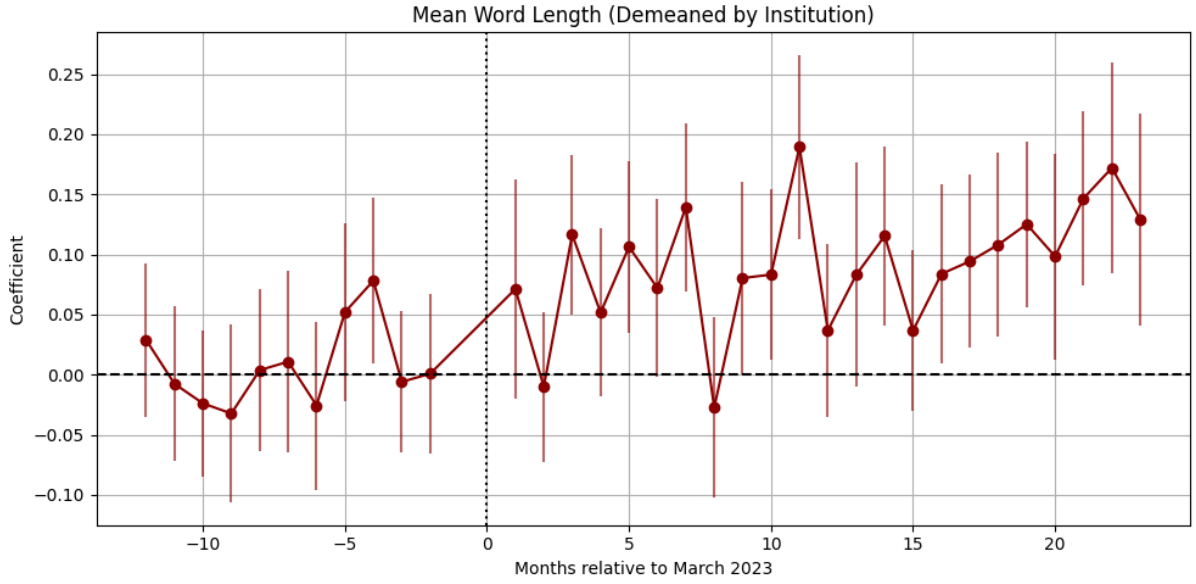

Figure 11: Event study estimates of mean word length (institution-demeaned) with 95% confidence intervals around March 2023 (month 0). Controls mirror earlier models; institution fixed characteristics like language/TLD group absorbed via demeaning.

9 Classifier, Adoption Estimates and Readability

To estimate the adoption of GPT-assisted writing, we trained a classifier on economic paper abstracts from 2010–2020, using human-written originals (label 0) and GPT-optimised versions (label 1) generated in our *placebo* experiment. The classifier was based on TF-IDF features combined with text metrics (e.g., character length, word count, type-token ratio, and Shannon entropy) and achieved an out-of-sample accuracy of 97% (validation set). Analysis of feature importance reveals that the model captures stylistic patterns such as the prevalence of formal connectors (e.g., "furthermore," "findings reveal," "study examines") that are more common in GPT-optimised abstracts. Tables and show top TF-IDF features that predict original or GPT-treated abstracts and the ranking and score of the numeric linguistic variables. Notice how TTR is significant in predicting GPT treatment.

GPT-like Features	**Original-like Features**
study	using
findings	important
notably	use
reveals	large
reveal	evidence
utilizing	effect
examines	paper
impact	used
analysis	people
findings reveal	increases
study examines	labour
significant	estimate
analysis reveals	particular
furthermore	affect
compared	little
additionally	likely
study investigates	results
exhibit	differences
leveraging	main
investigates	finally

Table 6: Top 20 GPT-like vs. Original-like Features Used in Classification

Feature	**Rank (out of 5005)**	**Classifier Weight**
char_length	4656	0.0213
word_count	2994	−0.1349
mean_word_length	2709	−0.1582
ttr	18	4.1906
entropy	1652	−0.2849

Table 7: Custom Feature Importance Rankings from Classifier

Applying the classifier to 2021–2025 abstracts, we compute the monthly average GPT-likeness probability p_t. For the pre-GPT baseline (2010–2020), the classifier assigns an average GPT probability of

$$p_{\text{baseline}} \approx 0.13,$$

reflecting both naturally formal human writing and the false positive rate of the classifier.

Post-March 2023 (following the release of ChatGPT-3.5), we observe a sharp increase in GPT-likeness:

$$\bar{p}_{\text{post}} \approx 0.32,$$

representing an absolute increase of $\Delta p = \bar{p}_{\text{post}} - p_{\text{baseline}} \approx 0.19$ (19 percentage points). A two-sample t-test confirms the significance of this shift ($t = 9.4$, $p < 10^{-10}$). A bootstrap analysis yields a 95% confidence interval for the change of

$$\Delta p \in [0.079,\, 0.120].$$

Assuming the baseline p_{baseline} remains stable, this implies that approximately 19% of recent abstracts (2024–2025) exhibit GPT-like stylistic patterns beyond what is naturally observed in human-written texts. This estimate is close to, but slightly above, our placebo-based lower bound of 16%, indicating robust and growing GPT adoption.

Figure 12 illustrates the monthly GPT-likeness probabilities (5-month moving average), with a vertical line marking the ChatGPT-3.5 release (March 2023).

Our GPT detection model is based on logistic regression, a widely used and interpretable machine learning method that also has strong roots in classical statistics. Logistic regression is a generalised linear model (GLM) that estimates the log-odds of class membership as a linear function of input features. In the context of text classification, the input features include high-dimensional TF-IDF n-grams and low-dimensional style metrics such as type-token ratio and Shannon entropy.

Although logistic regression is often taught as a statistical modelling technique, it is fully compatible with modern machine learning practice: it is trained using supervised learning on labeled data, learns parameters via optimisation (maximum likelihood estimation), and is evaluated on its predictive accuracy on out-of-sample data. As a result, logistic regression remains a strong baseline classifier for natural language processing tasks, offering both predictive power and interpretability.

Using a version of our classifier trained solely on TF-IDF features—excluding linguistic metrics—we estimated the probability that a post-March 2023 abstract resembles GPT-generated text. We then aggregated these probabilities at the institution-month level to construct a measure of institutional GPT-likeness over time. To assess the relationship between GPT-likeness and writing style, we interacted this measure with event-time dummies in a panel event-study regression framework. This can be thought of as a way of reconstructing an unobserved GPT treatment assignment using predicted exposure inferred from text alone.

As dependent variables, we used the institution-demeaned values of three well-established readability metrics: Flesch Reading Ease, SMOG, and the Automated Readability Index (ARI). This demeaning strategy removes time-invariant institutional fixed effects, and we include a within-institution time trend to absorb secular changes in writing style. Unlike our standard regression specifications, we omit additional controls such as abstract length, number of authors, and number of keywords, since these variables are highly correlated with linguistic metrics like mean word length (MWL) and total lexical length (TLL), which directly influence readability. Including them would risk controlling away the stylistic variation we aim to attribute to GPT influence.

Across all three readability specifications, we consistently find that higher GPT-likeness—as predicted by our classifier—is associated with lower readability scores. This suggests that the adoption of GPT-assisted writing may reduce textual clarity and increase complexity, at least as measured by traditional readability indices. Our results are summarised in Figures 13, 14, 15.

[3] It could be argued that for highly formal and compact texts such as economics abstracts, standard readability indices may not fully capture actual ease of understanding. A worsening in scores might just reflect increased textual complexity rather than a straightforward deterioration in readability.

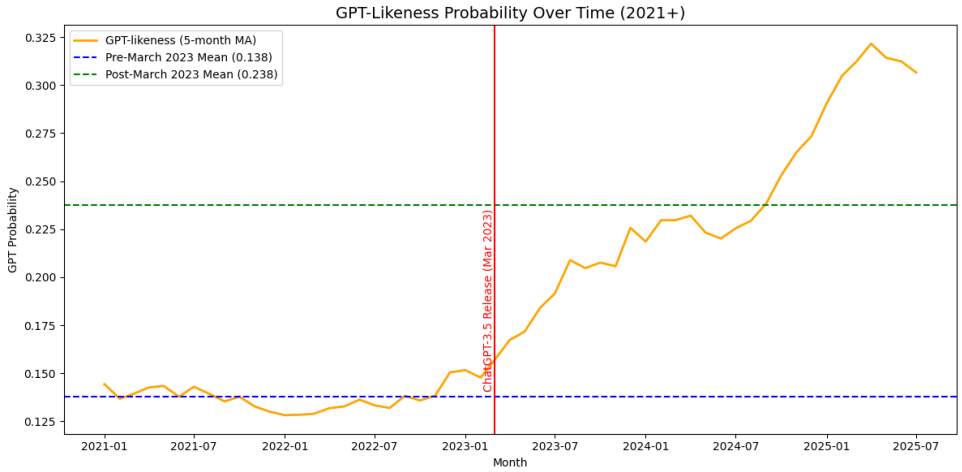

Figure 12: Monthly GPT-likeness probability of economic paper abstracts (2021–2025), estimated using a classifier trained on 2010–2020 human-written and GPT-optimised abstracts. The orange line shows the 5-month moving average. Vertical dashed lines mark key events: the GPT-3 API launch (June 2020), the ChatGPT public release (November 2022), and the ChatGPT-3.5 release (March 2023). A significant post-March 2023 increase in GPT-likeness is observed.

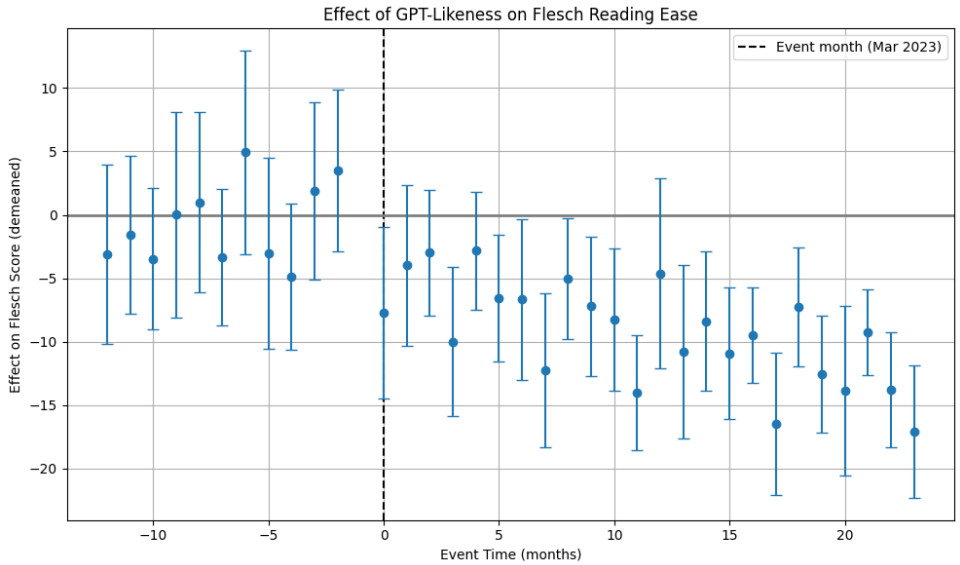

Figure 13: Effect of GPT-Likeness on Flesch Reading Ease (Demeaned). This figure shows estimated coefficients from a panel event-study regression of Flesch Reading Ease on GPT-likeness interacted with event-time dummies, using institution-demeaned Flesch scores as the dependent variable. The vertical dashed line indicates the release of ChatGPT-3.5 (March 2023). Each point represents the estimated effect of GPT-likeness at a given event time, with 95% confidence intervals clustered at the institution level. Negative coefficients after the event suggest a decline in readability among GPT-like writing.

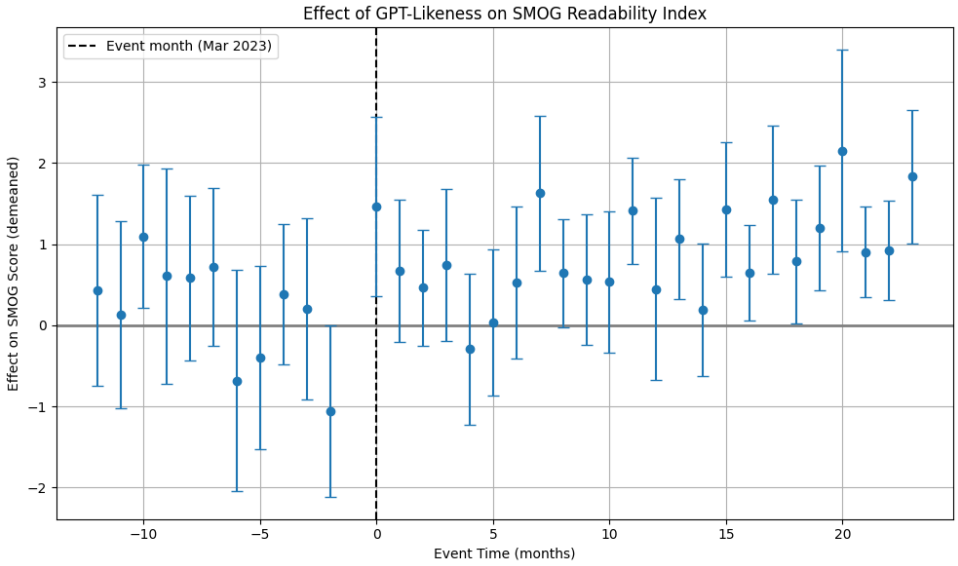

Figure 14: Effect of GPT-Likeness on SMOG readability index (Demeaned). This figure shows estimated coefficients from a panel event-study regression of SMOG readability index on GPT-likeness interacted with event-time dummies, using institution-demeaned SMOG readability indices as the dependent variable. The vertical dashed line indicates the release of ChatGPT-3.5 (March 2023). Each point represents the estimated effect of GPT-likeness at a given event time, with 95% confidence intervals clustered at the institution level. Positive coefficients after the event suggest a decline in readability among GPT-like writing.

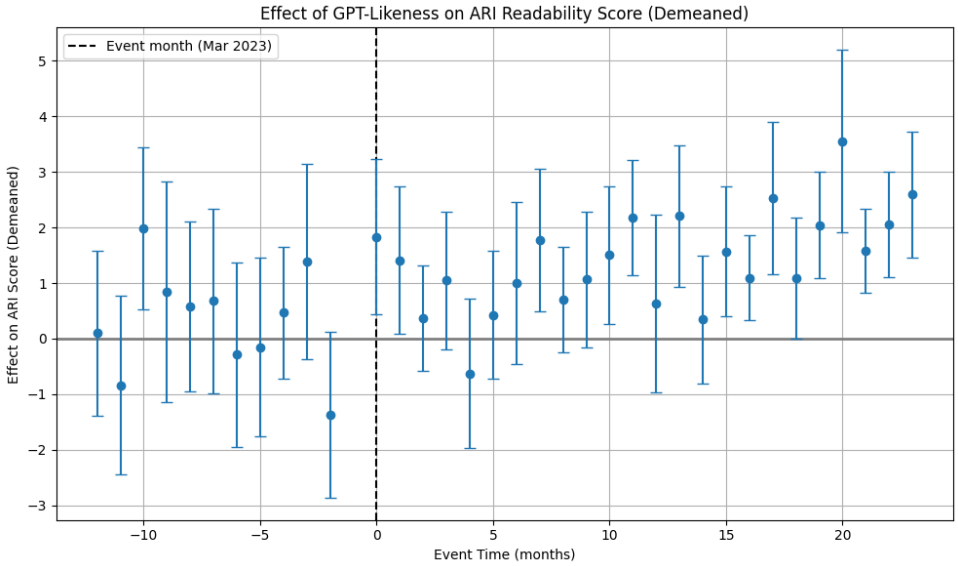

Figure 15: Effect of GPT-Likeness on ARI readability index (Demeaned). This figure shows estimated coefficients from a panel event-study regression of ARI readability index on GPT-likeness interacted with event-time dummies, using institution-demeaned ARI readability indices as the dependent variable. The vertical dashed line indicates the release of ChatGPT-3.5 (March 2023). Each point represents the estimated effect of GPT-likeness at a given event time, with 95% confidence intervals clustered at the institution level. Positive coefficients after the event suggest a decline in readability among GPT-like writing.

10 Conclusion

Large language models such as ChatGPT have made advanced language generation tools broadly accessible. This paper provides empirical evidence that these tools are already reshaping academic writing. Focusing on economics working paper abstracts from the IZA Discussion Paper series, we detect a significant and abrupt shift in stylistic features—such as word length, lexical diversity, and readability—coinciding with the public release of ChatGPT-3.5 in March 2023.

By combining a suite of complementary methods—including event-study regressions with rich controls, placebo tests, similarity metrics using the OpenAI API, and interpretable machine learning—we isolate a persistent change in writing style that is consistent with the uptake of generative AI tools. The triangulation of findings across these approaches reinforces the interpretation that this shift reflects behavioral adoption rather than background trends or confounders.

While the release of ChatGPT constitutes an exogenous shock, actual usage remains endogenous: researchers choose whether and how to integrate AI into their writing workflow. Our findings suggest that this integration is selective rather than wholesale—pointing to a pattern of human-AI augmentation rather than substitution. Abstracts are not being written entirely by machines, but they are increasingly shaped by them.

The framework developed here generalizes readily to other professional writing settings. With appropriate adaptations, it can be applied to resumes, cover letters, job ads, research proposals, code documentation, and beyond—any domain where writing carries meaning and consequence. Future work could expand the scope to full papers, explore topic-specific heterogeneity, or extend the analysis across the wider RePEc corpus to uncover institutional and disciplinary variation.

In capturing this early phase of AI adoption in academic writing, we provide a template for tracing how generative tools alter the mechanics of communication—and, by extension, the way expertise and knowledge are presented and perceived.

References

Alvarez, R., Holcombe, A. O., and Bhatia, S. (2024). Opportunities and risks of generative AI for scientific communication. *Nature Human Behaviour*, 8:675–679.

Bao, T., Zhao, Y., Mao, J., and Zhang, C. (2025). Examining linguistic shifts in academic writing before and after the launch of ChatGPT: A study on preprint papers.

Doshi, S., Bhatia, S., and Ungar, L. H. (2024). Generative AI improves output quality but reduces stylistic diversity. *Science Advances*, 10(18):eadn5290.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233.

Gao, C., Howard, F. M., Markov, N., Dyer, E. C., Smith, A., and Watson, D. S. (2022). Comparing scientific abstracts generated by chatgpt to original abstracts using blinded human reviewers. *bioRxiv*. Preprint.

Gentzkow, M., Kelly, B., and Taddy, M. (2019). Text as data. *Journal of Economic Literature*, 57(3):535–574.

Glickman, M. and Sharot, T. (2025). AI feedback loops alter perception and emotional response. *Nature Human Behaviour*, 9(1):12–19.

Humlum, A. and Vestergaard, C. (2024). The adoption of ChatGPT. *SSRN*. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4807516.

Imperial, J. M. and Madabushi, H. T. (2023). Flesch or fumble? evaluating readability standard alignment of instruction-tuned language models.

Karpf, D. (2023). Why GPT is different: Ai writing and the collapse of the college essay. *Communication and the Public*, 8(1):4–9.

Korinek, A. (2023). Language models and cognitive automation for economic research. *Brookings Papers on Economic Activity*. Forthcoming.

Liang, W., Zhang, Y., Codreanu, M., Wang, J., Cao, H., and Zou, J. (2025). The widespread adoption of large language model-assisted writing across society.

Lin, D., Zhao, N., Tian, D., and Li, J. (2025). ChatGPT as linguistic equalizer? quantifying LLM-Driven lexical shifts in academic writing.

McLaughlin, H. G. (1969). SMOG Grading – a new readability formula. *Journal of Reading*, 12(8):639–646.

Senter, R. J. and Smith, E. A. (1967). Automated Readability Index. (AMRL-TR-66-220). Technical Report.

Stokel-Walker, C. (2023). AI bot ChatGPT writes smart essays — should professors worry? *Nature*, 613:620–621.

van Dis, E., Bollen, J., Zuidema, W., van Rooij, R., and Bockting, C. (2023). ChatGPT: Five priorities for research. *Nature*, 614(7947):224–226.

Walther, T. and Dutordoir, M. (2024). Certainly! generative AI and its impact on academic writing (in finance). *SSRN*. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5317993.

Xu, Z. (2025). Patterns and purposes: A cross-journal analysis of AI tool usage in academic writing.