PUEY UNGPHAKORN INSTITUTE
FOR ECONOMIC RESEARCH

Experimenting with Generative AI: Does ChatGPT Really Increase Everyone's Productivity?

by

Voraprapa Nakavachara, Tanapong Potipiti, Thanee Chaiwat

January 2025

Discussion Paper

No. 229

Experimenting with Generative AI: Does ChatGPT Really Increase Everyone's Productivity?

Voraprapa Nakavachara, Tanapong Potipiti[1], Thanee Chaiwat
Faculty of Economics, Chulalongkorn University

6 January 2025

Generative AI technologies such as ChatGPT, Gemini, and MidJourney have made remarkable progress in recent years. Recent literature has documented ChatGPT's positive impact on productivity in areas where it has strong expertise—attributable to extensive training datasets—such as the English language and Python/SQL programming. However, the literature is still limited regarding ChatGPT's performance in areas where its capabilities could still be further enhanced. In this paper, we asked participants to perform writing analysis tasks in a non-English language (specifically, Thai) and math & data analysis tasks using a less frequently used programming package (specifically, Stata). The findings suggest that, on average, participants performed better using ChatGPT in terms of scores and time taken to complete the tasks. However, a detailed examination reveals that 34% of participants saw no improvement in writing analysis tasks, and 42% did not improve in math & data analysis tasks when employing ChatGPT. Further investigation indicated that higher-ability participants, as proxied by their econometrics grades, were the ones who performed worse in writing analysis tasks when using ChatGPT. We also found evidence that participants with better digital skills performed better with ChatGPT. This research provides insights on the impact of generative AI. Thus, relevant parties can make informed decisions regarding appropriate strategies, policies, and educational systems. It also highlights the critical role of human skills in addressing and complementing the limitations of AI.

JEL Codes: A20, D24, J24, O33
Keywords: ChatGPT, Generative AI, Large Language Models, Labor Productivity

[1] Corresponding Author: tanapong.p@chula.ac.th

We are grateful for the funding provided by the Puey Ungphakorn Institute for Economic Research at the Bank of Thailand. We also extend our sincere appreciation for the research assistance provided by Chanalak Chaisrilak and Nathatai Kullapa-anunchan. The research conducted in this paper was approved by Chulalongkorn University's Research Ethics Review Committee for Research Involving Human Subjects.

1. Introduction

Generative AI technologies such as ChatGPT, Gemini, and MidJourney are experiencing rapid advancements and are anticipated to improve significantly in the coming years. Currently, they have the capability to execute a diverse array of tasks, including but not limited to composing text, generating ideas, writing codes, and creating artwork. Major industry players, such as Microsoft/OpenAI Google and Meta, are channeling substantial investments into these technologies, ensuring their evolution.

Recent literature has begun to explore how these Generative AI technologies can impact labor productivity and the economy. Brynjolfsson et al. (2023) asked customer support agents to use AI-enabled tools and found a 14 percent improvement in their productivity. Noy and Zhang (2023) conducted an experiment with mid-level professional writing tasks and observed that participants using ChatGPT completed the tasks faster and with higher quality. Korinek (2023) noted that ChatGPT can be useful for a variety of tasks, including ideation and feedback, writing, background research, coding, data analysis, and mathematics. Cheng et al. (2023) explored whether GPT-4 could serve as an effective data analyst. The authors found that GPT-4 can outperform an entry-level data analyst in terms of efficiency and cost, while also delivering results more quickly.

Most of the literature has documented ChatGPT's positive impact on productivity, particularly in tasks that fall within its areas of strong expertise—attributable to extensive training datasets—such as the English language and Python/SQL programming. However, there are still a limited number of papers addressing ChatGPT's performance in areas where its capabilities could be further enhanced.

In this study, we recruited undergraduate economics students who had completed basic courses in statistics and econometrics. To focus on a less commonly used programming package, specifically Stata, we required that the participants have prior experience using Stata. This selection criterion limited the number of participants available for the study. Consequently, we were unable to randomly assign participants to different groups. Therefore, we employed regression analysis as our primary methodological approach.

The participants were asked to perform writing analysis tasks in Thai and math and data analysis tasks using Stata. This study contributes to the emerging research regarding ChatGPT by exploring its usage in non-English and non-Python/SQL environments, areas where ChatGPT still has significant room for improvement.

The results indicate that, on average, participants demonstrated improved performance in both quality and time taken to complete the tasks when utilizing ChatGPT. Yet, a closer analysis unveils that 34% of participants saw no improvement in writing analysis tasks, and 42% did not improve in math & data analysis tasks when employing ChatGPT. Further scrutiny revealed that higher-ability participants, as proxied by their econometrics grades, were the ones who performed worse in writing analysis when using ChatGPT. Additionally, our findings suggest that participants possessing advanced digital skills experienced improved performance when using ChatGPT.

This research offers valuable insights for educators, policymakers, businesses, and workers, enabling them to comprehend and predict the ways in which generative AI technologies may enhance or diminish the performance of individuals. Consequently, they can make informed decisions to adapt strategies or implement appropriate policies. Moreover, the study contributes to the education sector by offering insights on how students perform in the advent of generative AI. A critical takeaway is the enduring importance of human skills in recognizing the limitations of

these technologies and compensating for them, highlighting the synergy between human expertise and artificial intelligence.

2. Literature Review

With the rapid advancement of generative AI, scholars from various disciplines have explored its diverse impacts. Economists have examined how generative AI affects worker productivity, focusing on its potential to automate tasks and augment human capabilities. Educators have investigated its effects on teaching methodologies, learning outcomes, and student performance. Additionally, researchers in other fields have assessed broader risks associated with generative AI, including risks, biases, and societal implications.

Brynjolfsson et al. (2023) conducted an experiment with 5,179 customer support agents, half of whom were allowed to use tools linked to ChatGPT. The study found that the group using ChatGPT experienced a 14 percent improvement in productivity, as measured by the number of cases solved per hour. Similarly, Noy and Zhang (2023) conducted an experiment involving 444 workers engaged in mid-level professional writing tasks. They discovered that the group using ChatGPT experienced a decrease in the time taken by 0.8 standard deviations, while the output quality increased by 0.4 standard deviations. Both studies highlighted that ChatGPT helped improve the productivity of inexperienced individuals more than that of experienced ones.

Cheng et al. (2023) investigated the potential of GPT-4 as a competent data analyst, finding that GPT-4 surpasses the performance of an entry-level data analyst in efficiency and cost-effectiveness, providing faster outcomes. Dell'Acqua et al. (2023) carried out an experiment involving consultants from Boston Consulting Group, categorizing them into three groups: those not permitted to use AI, those using GPT-4, and those using GPT-4 with an overview. For tasks

within AI's strengths, such as generating and developing new product ideas, consultants using AI demonstrated notable increases in productivity. Conversely, for tasks outside AI's capabilities, such as solving business problems with quantitative data and conducting customer and company interviews, consultants using AI performed slightly worse than their counterparts without AI.

In terms of education, Fauzi et al. (2023) found that ChatGPT significantly enhances student productivity by offering useful information, improving language skills, facilitating collaboration, increasing efficiency, and providing support and motivation. Choi and Schwarcz (2024) studied the use of ChatGPT by legal students during exams. They observed that GPT-4 significantly improved performance on straightforward multiple-choice questions but had little effect on complex essay questions. The impact of GPT-4 varied greatly based on students' initial skill levels; those with lower starting points experienced substantial improvements with AI assistance, whereas top-performing students encountered declines in performance. Kim and Moon (2024) found that ChatGPT worsened the math performance of individuals, highlighting a paradox where the tool intended to assist users instead became a stumbling block in this context. Raman et al. (2024) examined the adoption and societal implications of ChatGPT, revealing that Gen Z students perceive ChatGPT as innovative, compatible, and user-friendly, making it a valuable tool for pursuing their educational goals. The students expressed satisfaction with using ChatGPT to support their learning.

ChatGPT has also been explored in other different research contexts. In risk analysis, Kim et al. (2023) found that ChatGPT's evaluation of firm risk from quarterly earnings call transcripts positively correlates with stock price volatility. This demonstrates its potential utility in financial risk assessment. Korinek (2023) emphasized ChatGPT's ability in various research tasks such as ideation, feedback, writing, background research, coding, data analysis, and mathematics, though

its effectiveness can vary based on the task. Meanwhile, Rutinowski et al. (2024) explored ChatGPT's self-perception and political biases, uncovering a tendency towards progressive and libertarian views and revealing that the model perceives itself as highly open and agreeable.

3. Methodological Design

In this paper, we explore how ChatGPT affects participants' performance in writing analysis tasks in Thai and math & data analysis tasks in Stata. As mentioned earlier, we required that our participants (undergraduate economics students) have taken basic statistics and econometrics classes. In addition, we required that they have previously used Stata. Consequently, we only have a limited pool of eligible partcipants and cannot randomly assigned them into groups. Therefore, we rely on regression analysis for our main methodology.

The main research focus is to compare the performance of undergraduate economics students in various skills, including econometric abilities. The pool of available subjects is limited to undergraduate economics students who have taken 3rd or 4th year economics courses and are familiar with Stata. Stata was selected as the econometrics software for the experiment because it is the most popular Econometrics software used at Chulalongkorn and Thammasat universities. Note that from our survey about 60% of economic students at Chula and Thammasat use Stata. Based on our estimates, the total subject pool at Chulalongkorn University and Thammasat University's Rangsit campus is approximately 400 students. Given this relatively small population size, we cannot be highly selective in our sampling method. Therefore, we employed convenience sampling for this study.

This study involved 121 undergraduate economics students from Chulalongkorn University and Thammasat University in Thailand.[2] The writing analysis tasks (in Thai language) encompassed brainstorming, reading and providing feedback on texts, composing tweets, and summarizing documents. The math & data analysis tasks included Stata coding for data visualization, variable generation, conducting regression analyses, hypothesis testing, and equation derivation.

At the beginning of the study, participants were informed about the session structure. They were asked to log in to their ChatGPT accounts, and time was allocated for account creation if needed. Participants began with two sessions of writing analysis tasks, comparing performance with and without ChatGPT. They then proceeded to two sessions of math & data analysis tasks, again with and without ChatGPT. The ChatGPT used in this study is version 3.5 (during January/February 2024).

Throughout all sessions, participants had access to a browser and the internet but were prohibited from using other Large Language Model (LLM) platforms. Administrative staff monitored their screens to ensure compliance with these restrictions and to verify that ChatGPT was only used during the permitted sessions.

To accurately assess the impact of ChatGPT, we prepared two sets of tasks (Set A and Set B) of equal difficulty for both writing analysis and math & data analysis. This design allowed for a direct comparison of individual performance with and without ChatGPT.To eliminate any bias from the difference in problem sets or the order of using ChatGPT, participants were allocated into four groups as detailed in Table 1. For instance, Group 1 started with Set A of writing analysis

[2] The sessions were conducted during 22-25 January 2024 at Chulalongkorn University and on 7 February 2024 at Thammasat University (Rangsit Campus). Participants were asked not to reveal and/or discuss the tasks until after all sessions were completed.

without ChatGPT, followed by Set B with ChatGPT, then proceeded to Set A of math & data analysis without ChatGPT, and finished with Set B with ChatGPT. The sequences for the other groups are outlined in Table 1.

Table 1: Participant Groups

	Writing Analysis Tasks		Math & Data Analysis Tasks	
	Session I	Session II	Session I	Session II
Group 1 (21 Participants)	Set A No ChatGPT	Set B With ChatGPT	Set A No ChatGPT	Set B With ChatGPT
Group 2 (34 Participants)	Set A With ChatGPT	Set B No ChatGPT	Set A With ChatGPT	Set B No ChatGPT
Group 3 (44 Participants)	Set B No ChatGPT	Set A With ChatGPT	Set B No ChatGPT	Set A With ChatGPT
Group 4 (22 Participants)	Set B With ChatGPT	Set A No ChatGPT	Set B With ChatGPT	Set A No ChatGPT

We assessed productivity in two dimensions – quality and time. We utilized two metrics: a quality score out of 10 (averaged from the assessments of three experts) and the time required to complete tasks. The grading criteria for the writing analysis tasks included accuracy, completeness, and the absence of redundancy. For the math and data analysis tasks, the grading criteria were more straightforward, as there was only one correct answer for each task. However, partial credit was awarded if participants demonstrated progress toward the correct solution by reaching certain milestones.

Participants were awarded monetary rewards based on their quality score (with higher scores preferred) and completion time (with quicker times preferred). Each participant received a base show-up fee plus a performance fee determined by both the time spent and the quality of their work. Each task session had a 20-minute limit. Participants could submit their work before this time, but were required to submit whatever they had at the 20-minute mark. Following all sessions, participants completed a questionnaire detailing their academic background (including GPAX and

econometrics grade) and self-assessed their reading, writing, math, and digital skills on a scale from 1 to 5.

4. Data Overview and Preliminary Findings

Table 2 provides a summary statistics for the 121 participants involved in this study. Participants are undergraduate economics students in their 3rd and 4th years. Their average cumulative grade point average (GPAX) is 3.32, with an average econometrics grade of 3.00. About 36% of them are male. They rated their abilities in reading, writing, math, and digital skills on a scale from 1 to 5 (where 5 is the highest), with average scores of 3.46, 3.07, 3.26, and 3.20, respectively. In preparation for our main regression analysis, detailed in Section 5, we defined a 'ChatGPT proficiency' variable based on 'ChatGPT usage per week,' with thresholds at 30 minutes, 1 hour, and 2 hours to categorize three levels of expertise with ChatGPT.

Table 2: Summary Statistics

Panel A: Overall Statistics

Variable	Obs	Mean	Std. Dev.	Min	Max
Male	121	0.36	0.48	0.00	1.00
CollegeYear	121	3.28	0.49	2.00	5.00
GPAX	121	3.32	0.33	2.50	3.99
Econometrics Grade	121	3.00	0.78	1.00	4.00
Self-Evaluated Reading Skills	121	3.46	0.82	1.00	5.00
Self-Evaluated Writing Skills	121	3.07	0.85	1.00	5.00
Self-Evaluated Math Skills	121	3.26	1.01	1.00	5.00
Self-Evaluated Digital Skills	121	3.20	0.97	1.00	5.00

Panel B: Breakdown of ChatGPT Usage per Week

Variable	Obs	Percent
ChatGPT Usage per Week	121	100.00
Never Used ChatGPT	8	6.61
Upto <30 Minutes	55	45.45
30 to <60 Minutes	25	20.66
1 to <2 Hours	15	12.40
2 Hours or More	18	14.87

Table 3 revealed our initial findings. In Panel A, it can be observed that, on average, participants achieved higher scores when permitted to use ChatGPT for both writing analysis and math & data analysis tasks. Panel B illustrated that, on average, participants completed tasks, both writing analysis and math & data analysis, more quickly when using ChatGPT.

Table 3: Preliminary Results

Panel A: Score (Maximum is 10)

Variable	Obs	Mean	Std. Dev.	Min	Max
Writing Analysis - No ChatGPT	121	5.33	1.47	1.42	8.92
Writing Analysis - With ChatGPT	121	5.75	1.47	1.42	8.75
Math & Data Analysis - No ChatGPT	121	3.09	2.91	0.00	10.00
Math & Data Analysis - With ChatGPT	121	4.74	2.97	0.00	10.00

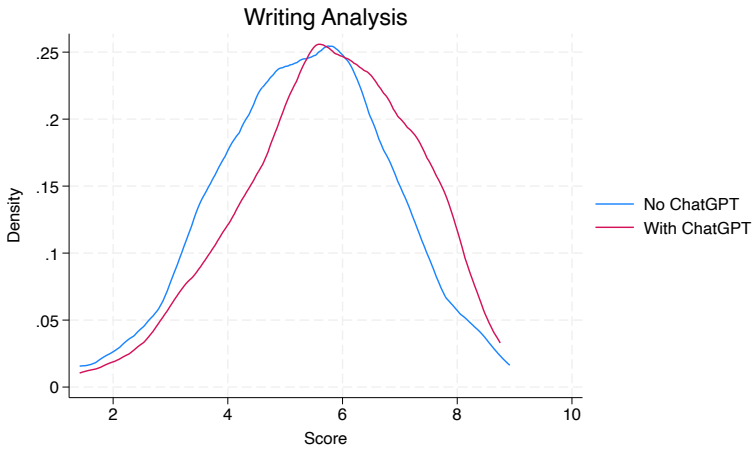
Panel B: Time Spent (Maximum is 20 Minutes)

Variable	Obs	Mean	Std. Dev.	Min	Max
Writing Analysis - No ChatGPT	121	19.19	1.62	9.00	20.00
Writing Analysis - With ChatGPT	121	17.46	3.09	7.00	20.00
Math & Data Analysis - No ChatGPT	121	18.26	2.57	5.00	20.00
Math & Data Analysis - With ChatGPT	121	17.27	3.04	6.00	20.00

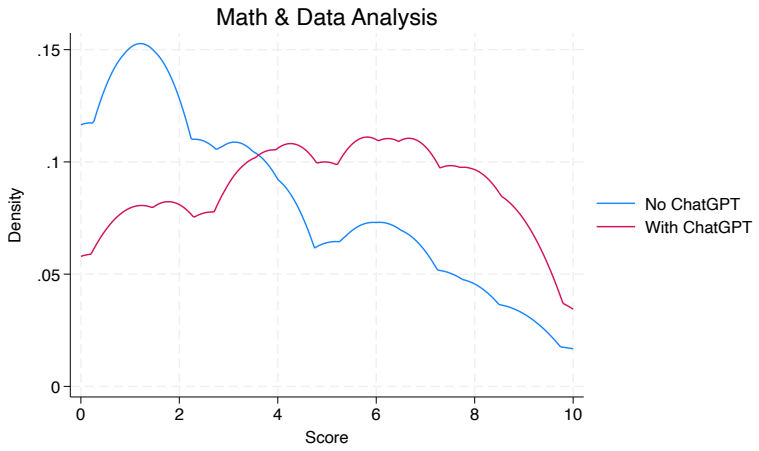
Figure 1 displayed the score and time distribution for both sets of tasks. For scores, the distribution for participants using ChatGPT lies on the right, indicating higher scores compared to those not using ChatGPT. Conversely, for time, the distribution for participants using ChatGPT lies towards the left, suggesting they completed tasks more swiftly than those not using ChatGPT.

Figure 1: Preliminary Results

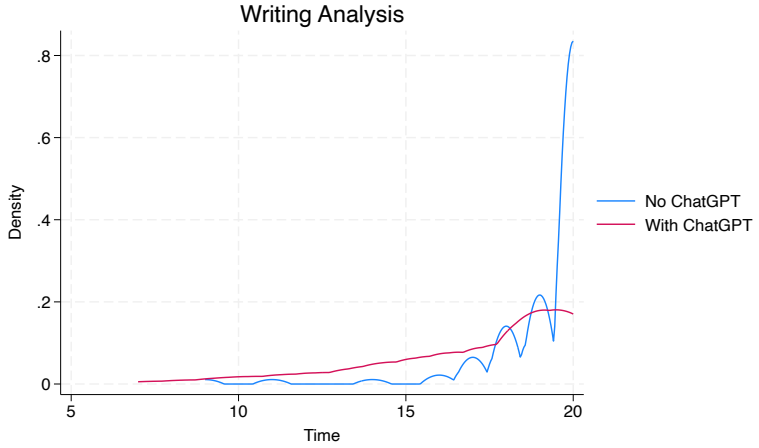
Panel A: Writing Analysis Score
No ChatGPT vs. With ChatGPT



Panel B: Math & Data Analysis Score
No ChatGPT vs. With ChatGPT



Panel C: Writing Analysis Time Spent
No ChatGPT vs. With ChatGPT



Panel D: Math & Data Analysis Time Spent
No ChatGPT vs. With ChatGPT

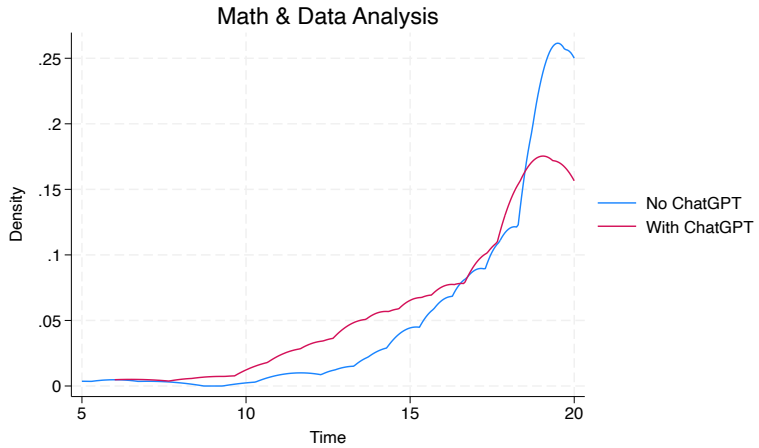


Table 4 delves deeper into the analysis of scores for writing tasks. 'Score Diff' is defined as the score a participant achieved when using ChatGPT minus the score achieved when not using ChatGPT. On the other hand, 'Time Diff' is defined as the time taken when not using ChatGPT minus the time taken when using ChatGPT. On average, the score with ChatGPT is higher than the score without ChatGPT by 0.43 (out of a total possible score of 10), while the time to complete tasks with ChatGPT is shorter than without by 1.73 minutes. Despite the overall improvement in scores with ChatGPT, 41 out of 121 participants (approximately 34%) did not see improved outcomes. This discrepancy underscores the importance of investigating the characteristics that differentiate participants who benefited from ChatGPT from those who did not. Interestingly, participants who did not benefit from ChatGPT appeared to have higher ability as proxied by higher GPAX and econometrics grades. They also self-report higher reading and writing skills. In contrast, participants who benefited from ChatGPT exhibit stronger math and digital skills. In testing for differences in means, only the Econometrics grade showed statistical significance at the 10% level.

Table 4: Detailed Results -- Writing Analysis

Panel A: Overall Difference

Variable	Obs	Mean	Std. Dev.	Min	Max
Score Diff = With ChatGPT - No ChatGPT	121	0.43	1.98	-5.00	5.17
Time Diff = No ChatGPT - With ChatGPT	121	1.73	3.43	-9.00	13.00

Panel B: Students with Score Diff > 0

Variable	Obs	Mean	Std. Dev.	Min	Max
Male	80	0.40	0.49	0.00	1.00
CollegeYear	80	3.34	0.53	2.00	5.00
GPAX	80	3.31	0.34	2.53	3.99
Econometrics Grade	80	2.91	0.78	1.00	4.00
Self-Evaluated Reading Skills	80	3.41	0.82	1.00	5.00
Self-Evaluated Writing Skills	80	3.05	0.91	1.00	5.00
Self-Evaluated Math Skills	80	3.29	1.03	1.00	5.00
Self-Evaluated Digital Skills	80	3.23	0.97	1.00	5.00

Panel C: Students with Score Diff <= 0

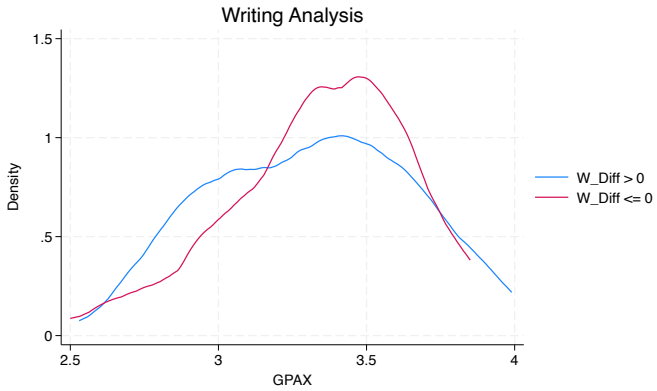
Variable	Obs	Mean	Std. Dev.	Min	Max
Male	41	0.29	0.46	0.00	1.00
CollegeYear	41	3.17	0.38	3.00	4.00
GPAX	41	3.34	0.30	2.50	3.85
Econometrics Grade	41	3.17	0.76	1.50	4.00
Self-Evaluated Reading Skills	41	3.56	0.81	2.00	5.00
Self-Evaluated Writing Skills	41	3.10	0.74	2.00	5.00
Self-Evaluated Math Skills	41	3.20	0.98	1.00	5.00
Self-Evaluated Digital Skills	41	3.15	0.99	1.00	5.00

Figure 2 explores this issue from a distributional perspective, contrasting participants who improved with ChatGPT against those who did not. Panels A through F display GPAX, econometrics grades, reading, writing, math, and digital skills. Panel A shows that the GPAX distribution for participants who did not benefit from ChatGPT shifts slightly right compared to their counterparts. Similary Panel B indicates a rightward shift in economics grade for the same group. The patterns for other skills are inconclusive.

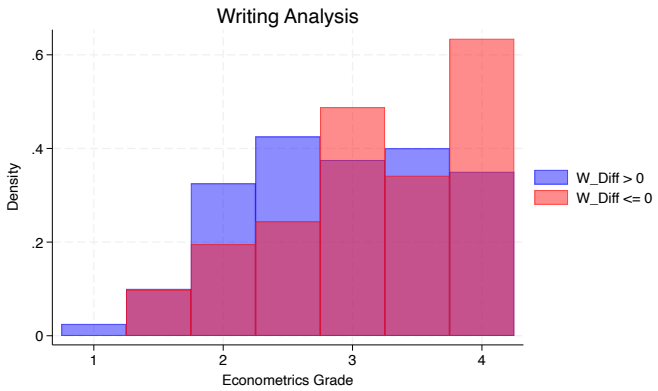
Table 5 looks into the details of math & data analysis scores. On average, scores with ChatGPT are higher than those without ChatGPT by 1.65 (out of a total score of 10), while the time taken with ChatGPT is shorter than without by 0.99 minutes. Despite the general improvement in scores with ChatGPT, 51 out of 121 participants (approximately 42%) did not perform better. Further observation reveals that, on average, participants who did not improve with ChatGPT tend to have higher GPAX, econometrics grades, and self-assessed reading and writing skills. Conversely, participants who improved with ChatGPT demonstrate better math and digital skills. However, none of the t-tests reached statistical significance. Figure 3 explores this issue through the distribution of scores, but the patterns remain inconclusive.

Figure 2: Detailed Results -- Writing Analysis

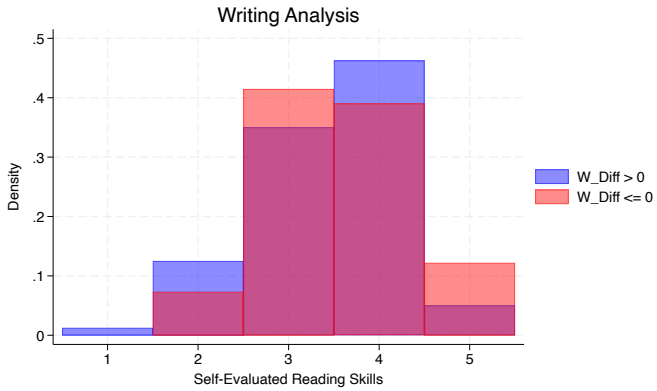
Panel A: Distribution of GPAX
Score Diff > 0 vs. Score Diff <= 0



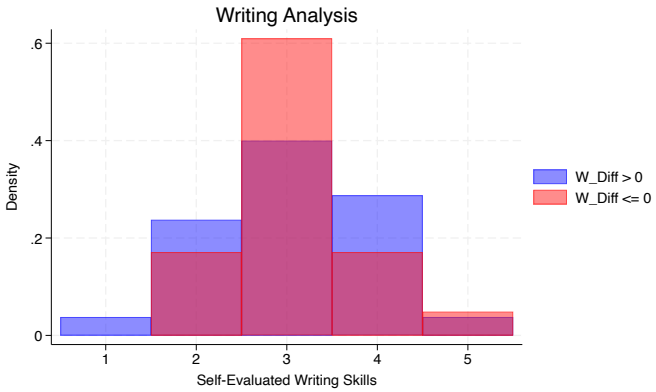
Panel B: Distribution of Econometrics Grade
Score Diff > 0 vs. Score Diff <= 0



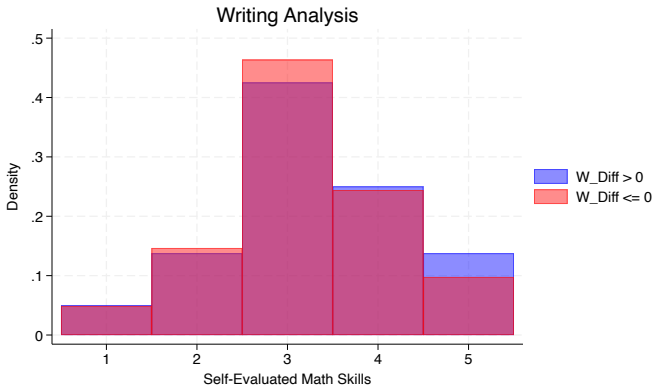
Panel C: Distribution of Self-Evaluated Reading Skills
Score Diff > 0 vs. Score Diff <= 0



Panel D: Distribution of Self-Evaluated Writing Skills
Score Diff > 0 vs. Score Diff <= 0



Panel E: Distribution of Self-Evaluated Math Skills
Score Diff > 0 vs. Score Diff <= 0



Panel F: Distribution of Self-Evaluated Digital Skills
Score Diff > 0 vs. Score Diff <= 0

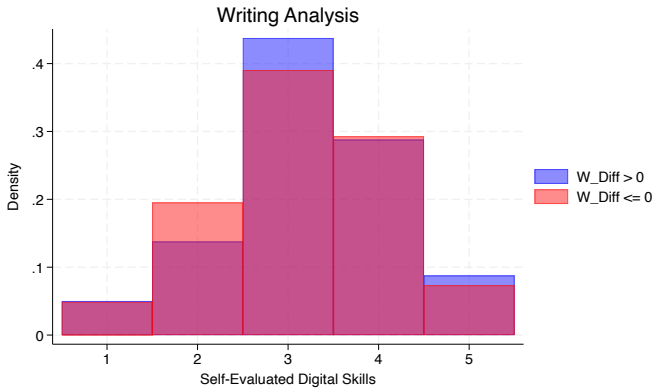


Table 5: Detailed Results -- Math & Data Analysis

Panel A: Overall Difference

Variable	Obs	Mean	Std. Dev.	Min	Max
Score Diff = With ChatGPT - No ChatGPT	121	1.65	3.67	-7.50	8.75
Time Diff = No ChatGPT - With ChatGPT	121	0.99	2.90	-7.00	10.00

Panel B: Students with Score Diff > 0

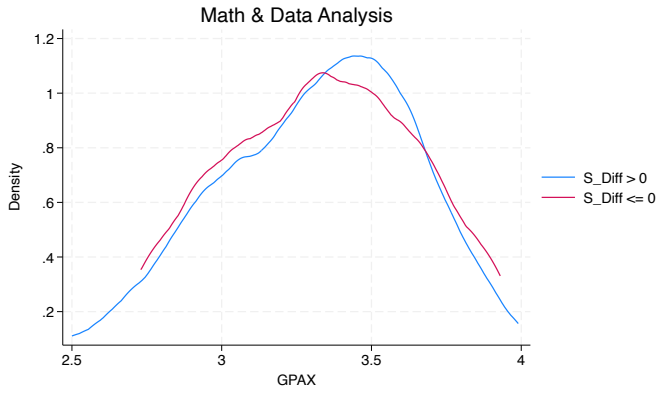
Variable	Obs	Mean	Std. Dev.	Min	Max
Male	70	0.36	0.48	0.00	1.00
CollegeYear	70	3.29	0.51	2.00	5.00
GPAX	70	3.31	0.33	2.50	3.99
Econometrics Grade	70	2.96	0.80	1.00	4.00
Self-Evaluated Reading Skills	70	3.43	0.84	1.00	5.00
Self-Evaluated Writing Skills	70	3.03	0.85	1.00	5.00
Self-Evaluated Math Skills	70	3.26	0.93	1.00	5.00
Self-Evaluated Digital Skills	70	3.30	0.86	1.00	5.00

Panel C: Students with Score Diff <= 0

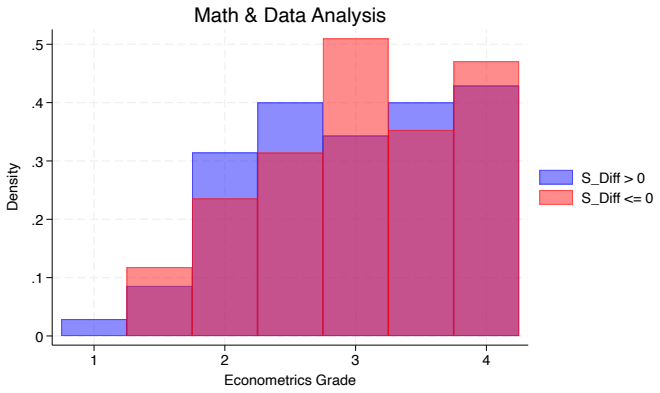
Variable	Obs	Mean	Std. Dev.	Min	Max
Male	51	0.37	0.49	0.00	1.00
CollegeYear	51	3.27	0.45	3.00	4.00
GPAX	51	3.33	0.32	2.73	3.93
Econometrics Grade	51	3.04	0.76	1.50	4.00
Self-Evaluated Reading Skills	51	3.51	0.78	2.00	5.00
Self-Evaluated Writing Skills	51	3.12	0.86	2.00	5.00
Self-Evaluated Math Skills	51	3.25	1.13	1.00	5.00
Self-Evaluated Digital Skills	51	3.06	1.10	1.00	5.00

Figure 3: Detailed Results – Math & Data Analysis

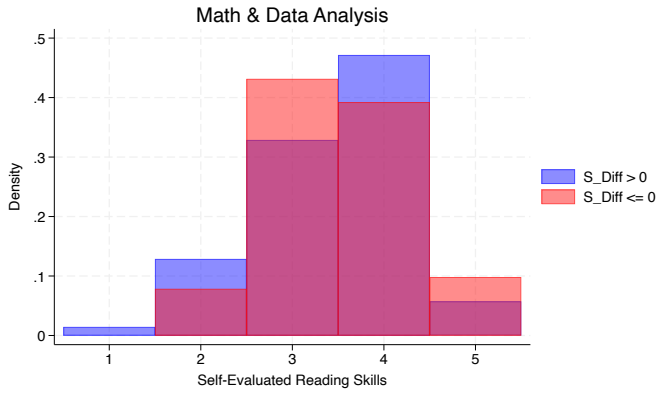
Panel A: Distribution of GPAX
Score Diff > 0 vs. Score Diff <= 0



Panel B: Distribution of Econometrics Grade
Score Diff > 0 vs. Score Diff <= 0

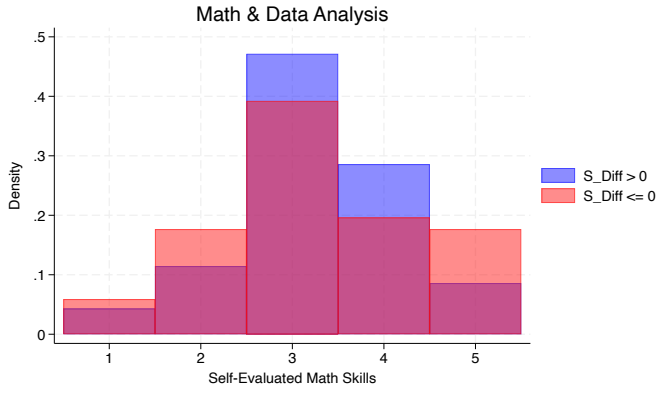


Panel C: Distribution of Self-Evaluated Reading Skills
Score Diff > 0 vs. Score Diff <= 0

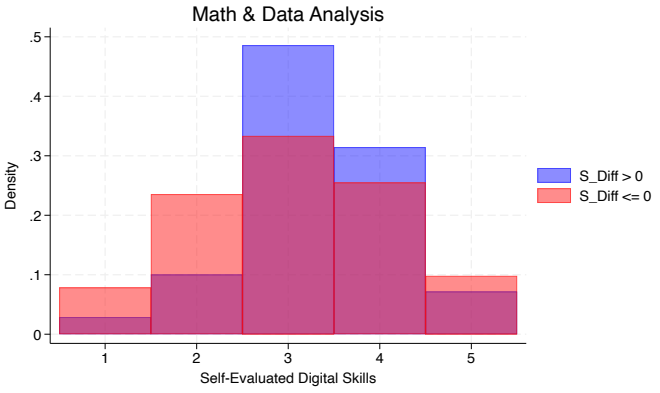


Panel D: Distribution of Self-Evaluated Writing Skills
Score Diff > 0 vs. Score Diff <= 0

Panel E: Distribution of Self-Evaluated Math Skills
Score Diff > 0 vs. Score Diff <= 0



Panel F: Distribution of Self-Evaluated Digital Skills
Score Diff > 0 vs. Score Diff <= 0



5. Empirical Models and Results

For our main analyses, we seek to investigate the factors contributing to individuals' performance improvement with ChatGPT, we further employ Linear Probability Model (LPM) regression analysis.

$$Improved\_Performance_i = \alpha + \beta_1 ChatExp_i + \beta_2 Econo_i + \beta_3 GPAX_i + \gamma \cdot \boldsymbol{Skills}_i + \delta \cdot \boldsymbol{x}_i + \varepsilon_i \quad (1)$$

Improved_Performance represents a dummy variable that equals to one if participants performed better when using ChatGPT and zero otherwise. *ChatExp* is a dummy variable indicating 'ChatGPT proficiency,' determined by whether participants possess expertise in using ChatGPT. This expertise is assessed based on 'ChatGPT usage per week,' with thresholds at 30 minutes, 1 hour, and 2 hours to categorize three levels of proficiency with ChatGPT. It is widely recognized within the LLM user community that prompting skills and expertise significantly affect LLM performance. Recent studies, such as the work by Chen, Banghao, et al. (2023) have demonstrated that a user's experience with LLMs and their prompt engineering skills are crucial factors in determining LLM performance on complex tasks.

Econo represents the student's grade in econometrics. *GPAX* denotes the cumulative grade point average. *Skills* is a vector comprising self-evaluated skills in four domains: reading, writing, math, and digital. *x* represents a vector of other control variables, including dummy variables for male[3], participant group (whether they belong to group 1 or 2 or 3 or 4 per Table 1 – this identifies the sequence the tasks they are given for their sessions), study program, university, and year. This

[3] Previous studies, such as the work by Goswami and Dutta (2015), have indicated that men tend to demonstrate higher levels of technological proficiency and are more likely to adopt new technologies compared to women.

analysis is conducted separately for writing analysis tasks and math & data analysis tasks. We employ robust standard errors in all our regressions to ensure the reliability of our findings.

In the supplemental analyses, we conducted the logistic regression model for both sets of tasks to compare outcomes. Robust standard errors were utilized in all regressions to ensure the accuracy and reliability of the results.

Table 6 presents our main regression findings. Columns 1-4 display results for the writing analysis tasks. Column 1 excludes the *ChatExp* variable, while Columns 2-4 incorporate *ChatExp* with thresholds of 30 minutes or more, 1 hour or more, and 2 hours or more per week, respectively. The *Econo* variable is consistently negative and significant across all specifications at the 5% level, suggesting that participants with higher econometrics grades are likely to perform worse when using ChatGPT, with the probability of 13.6% to 15.2%. Similarly, Columns 5-8 detail results for the math & data analysis tasks. Column 5 omits the *ChatExp* variable, whereas Columns 6-8 include *ChatExp* at thresholds of 30 minutes or more, 1 hour or more, and 2 hours or more per week, respectively. The *Digital skills* variable is positive and significant at the 5% to 10% level in all models, implying that participants with higher digital skills are likely to perform better when using ChatGPT, with the probability of 12.1% to 13.5%.

Table 7 showcases the supplemental regression outcomes employing the logistic model. In the analysis of writing tasks, presented in Columns 1-4, the *Econo* variable consistently appears negative and significant at the 5% level across all specifications. This pattern suggests that participants with higher econometrics grades tend to perform worse when utilizing ChatGPT. For the math & data analysis tasks, detailed in Columns 5-8, the *Digital Skills* variable again shows a positive and significant relationship, at levels ranging from 5% to 10%, across all specifications. This reaffirms that participants with enhanced digital skills tend to excel when using ChatGPT.

Table 6:Main Regression Results -- Linear Probability Model

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Writing Analysis				Math & Data Analysis			
VARIABLES	Improved Performance	Improved Performance	Improved Performance	Improved Performance	Improved Performance	Improved Performance	Improved Performance	Improved Performance
ChatExp_30m		-0.0846				0.0760		
		(0.0786)				(0.0995)		
ChatExp_1hr			-0.0710				-0.00953	
			(0.0820)				(0.105)	
ChatExp_2hr				-0.128				0.00540
				(0.0815)				(0.139)
Econo	-0.142**	-0.136**	-0.143**	-0.152**	-0.0361	-0.0408	-0.0363	-0.0357
	(0.0640)	(0.0629)	(0.0650)	(0.0660)	(0.0786)	(0.0799)	(0.0788)	(0.0819)
GPAX	0.147	0.125	0.153	0.156	-0.0265	-0.00665	-0.0257	-0.0269
	(0.162)	(0.162)	(0.164)	(0.163)	(0.154)	(0.159)	(0.155)	(0.156)
Skills_Reading	-0.0540	-0.0608	-0.0592	-0.0579	0.0180	0.0241	0.0173	0.0181
	(0.0565)	(0.0569)	(0.0564)	(0.0565)	(0.0597)	(0.0600)	(0.0610)	(0.0598)
Skills_Writing	0.0165	0.0166	0.0193	0.0147	-0.0424	-0.0425	-0.0420	-0.0423
	(0.0549)	(0.0548)	(0.0552)	(0.0560)	(0.0614)	(0.0615)	(0.0619)	(0.0621)
Skills_Math	-0.0102	-0.0121	-0.0106	-0.00241	-0.0366	-0.0349	-0.0366	-0.0369
	(0.0517)	(0.0518)	(0.0518)	(0.0528)	(0.0548)	(0.0557)	(0.0551)	(0.0552)
Skills_Digital	0.0539	0.0691	0.0579	0.0569	0.135**	0.121*	0.135**	0.135**
	(0.0462)	(0.0468)	(0.0468)	(0.0464)	(0.0626)	(0.0676)	(0.0629)	(0.0637)
X_Male	0.199*	0.192*	0.194*	0.201*	0.0438	0.0496	0.0431	0.0437
	(0.101)	(0.103)	(0.103)	(0.102)	(0.101)	(0.0987)	(0.0994)	(0.101)
Constant	0.643*	0.659*	0.628*	0.609*	0.611	0.597	0.609	0.612
	(0.353)	(0.358)	(0.356)	(0.357)	(0.396)	(0.403)	(0.399)	(0.397)
Model	LPM	LPM	LPM	LPM	LPM	LPM	LPM	LPM
Observations	121	121	121	121	121	121	121	121
R-squared	0.322	0.328	0.325	0.330	0.267	0.271	0.267	0.267

Other fixed-effect variables included in the regressions but not shown in the table: Participant group, Study Program, University, Year

Robust standard errors in parentheses

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

Table 7: Supplemental Regression Results -- Logistic Model

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Writing Analysis				Math & Data Analysis			
VARIABLES	Improved Performance	Improved Performance	Improved Performance	Improved Performance	Improved Performance	Improved Performance	Improved Performance	Improved Performance
ChatExp_30m		-0.393				0.425		
		(0.548)				(0.513)		
ChatExp_1hr			-0.195				0.00737	
			(0.600)				(0.570)	
ChatExp_2hr				-0.726				0.0265
				(0.623)				(0.744)
Econo	-1.119**	-1.069**	-1.098**	-1.119**	-0.209	-0.247	-0.209	-0.208
	(0.514)	(0.516)	(0.502)	(0.517)	(0.415)	(0.419)	(0.417)	(0.421)
GPAX	1.063	0.919	1.055	1.046	-0.138	0.0169	-0.138	-0.138
	(1.031)	(1.056)	(1.029)	(1.022)	(0.762)	(0.825)	(0.762)	(0.762)
Skills_Reading	-0.410	-0.459	-0.437	-0.430	0.136	0.171	0.137	0.137
	(0.387)	(0.405)	(0.405)	(0.391)	(0.302)	(0.306)	(0.309)	(0.301)
Skills_Writing	0.173	0.178	0.185	0.154	-0.288	-0.278	-0.288	-0.287
	(0.393)	(0.394)	(0.394)	(0.394)	(0.327)	(0.327)	(0.327)	(0.334)
Skills_Math	-0.0959	-0.0979	-0.0918	-0.0448	-0.218	-0.222	-0.218	-0.220
	(0.319)	(0.317)	(0.318)	(0.333)	(0.285)	(0.286)	(0.284)	(0.287)
Skills_Digital	0.421	0.492	0.423	0.438	0.774**	0.704*	0.774**	0.774**
	(0.293)	(0.302)	(0.293)	(0.295)	(0.375)	(0.393)	(0.374)	(0.380)
X_Male	1.114*	1.057	1.084	1.072*	0.199	0.227	0.199	0.200
	(0.652)	(0.676)	(0.684)	(0.650)	(0.508)	(0.494)	(0.496)	(0.506)
Constant	1.058	1.470	1.110	1.166	-0.717	-1.228	-0.719	-0.722
	(2.790)	(2.926)	(2.771)	(2.803)	(2.278)	(2.464)	(2.284)	(2.291)
Model	Logistic	Logistic	Logistic	Logistic	Logistic	Logistic	Logistic	Logistic
Observations	119	119	119	119	119	119	119	119
Psudo R-squared	0.2963	0.2992	0.2969	0.3012	0.2090	0.2134	0.2090	0.2090

Other fixed-effect variables included in the regressions but not shown in the table: Participant group, Study Program, University, Year

Robust standard errors in parentheses

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

6. Conclusion and Policy Implications

In this paper, we explore domains where ChatGPT's expertise may not yet be fully developed, specifically focusing on tasks conducted in a non-English language (Thai) and math & data analysis tasks utilizing a less commonly used programming package (Stata). We recruited 121 participants who were undergraduate economics students (i.e., economics major), asking them to complete writing analysis tasks in Thai and math & data analysis tasks in Stata. Our findings indicate that, on average, participants achieved better scores and completed tasks more quickly using ChatGPT. However, a closer examination reveals that 34% and 42% of participants did not demonstrate improvement with ChatGPT for writing analysis and math & data analysis tasks, respectively. Further analysis showed that participants with higher econometrics grades—serving as a proxy for higher ability—tended to perform worse in writing analysis tasks when using ChatGPT. Additionally, participants with superior digital skills were found to perform better with ChatGPT across tasks.

Interestingly, our results showed that the *GPAX* variable was not significant in any of our models, whereas econometrics grades did. Our hypothesis is that econometrics grades might serve as a more accurate measure of participants' field-specific abilities (in this case, they are undergraduate economics students) compared to GPAX, which encompasses a broader range of subjects, including those not directly related to economics (e.g., General Education). Additionally, the *ChatExp* variable*,* a proxy of 'ChatGPT proficiency' determined by how oftern each participant uses ChatGPT per week, was never significant our models. On the other hand, the *digital skills* variable was significant. Our hypothesis is that general digital competencies may enable participants to more effectively engage with any digital or technical program, including ChatGPT, regardless of their frequency of ChatGPT usage per week.

From our survey, we asked the participants whether they feel that ChatGPT can improve or worsen the skills of the users. Their answers are shown in Table 8. Panels A through E outline the participants' perception of how ChatGPT would impact thinking, speaking, reading, writing, and math skills, respectively. Interestingly, the participants mostly answer "probably worse" meaning that they feel that ChatGPT could worsen the skills of the users if not used appropriately.

This research provides critical insights for educators, policymakers, business leaders, and workers, facilitating a deeper understanding of how generative AI technologies can either augment or impair people's performance. As a result, relevant parties are better equipped to make well-informed decisions, adapt their strategies, or develop suitable policies in response. Furthermore, this study makes a significant contribution to the field of education by providing insights on students' performance in light of generative AI's emergence. A key takeaway is the pivotal role of human skills in identifying and mitigating the limitations of these technologies, underscoring the complementary relationship between human expertise and artificial intelligence. This highlights the necessity for ongoing education and skill development in maximizing the benefits of AI while addressing its challenges.

Based on the findings of this research, there are few policy implications that can be drawn. Given the mixed performance improvements observed when using ChatGPT, particularly among higher-ability students in writing analysis tasks, education policies should focus on tailoring AI integration to individual student needs. Policymakers should consider implementing adaptive learning environments where AI tools like ChatGPT are used in conjunction with personalized instruction, ensuring that students with different skill levels benefit appropriately. For example, students with stronger econometrics backgrounds may need more targeted guidance on when and how to use ChatGPT effectively without hindering their analytical skills.

Table 8: How would ChatGPT impact skills?

Panel A: Thinking Skills

Variable	Obs	Percent
How would ChatGPT impact thinking skills?	121	100.00
1: Worse	13	10.74
2: Probably Worse	38	31.40
3: Not Sure	30	24.79
4: Probably Better	25	20.66
5: Better	15	12.40

Panel B: Speaking Skills

Variable	Obs	Percent
How would ChatGPT impact speaking skills?	121	100.00
1: Worse	7	5.79
2: Probably Worse	46	38.02
3: Not Sure	37	30.58
4: Probably Better	22	18.18
5: Better	9	7.44

Panel C: Reading Skills

Variable	Obs	Percent
How would ChatGPT impact reading skills?	121	100.00
1: Worse	21	17.36
2: Probably Worse	35	28.93
3: Not Sure	32	26.45
4: Probably Better	20	16.53
5: Better	13	10.74

Panel D: Writing Skills

Variable	Obs	Percent
How would ChatGPT impact writing skills?	121	100.00
1: Worse	23	19.01
2: Probably Worse	35	28.93
3: Not Sure	30	24.79
4: Probably Better	23	19.01
5: Better	10	8.26

Panel E: Math Skills

Variable	Obs	Percent
How would ChatGPT impact Math skills?	121	100.00
1: Worse	10	8.26
2: Probably Worse	27	22.31
3: Not Sure	43	35.54
4: Probably Better	27	22.31
5: Better	14	11.57

7. References

Brynjolfsson, E., Li, D., & Raymond, L. R. (2023). *Generative AI at Work* (No. w31161). National Bureau of Economic Research.

Chen, B., Zhang, Z., Langrené, N. & Zhu S. (2023). Unleashing the potential of prompt engineering in Large Language Models: a comprehensive review. arXiv preprint arXiv:2310.14735.

Cheng, L., Li, X., & Bing, L. (2023). Is GPT-4 a Good Data Analyst?. *arXiv preprint arXiv:2305.15038*.

Choi, J. H., & Schwarcz, D. (2024). AI assistance in legal analysis: An empirical study. Journal of Legal Education, 73.

Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., & Lakhani, K. R. (2023). Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. Harvard Business School Working Paper No. 24-013.

Fauzi, F., Tuhuteru, L., Sampe, F., Ausat, A. M. A., & Hatta, H. R. (2023). Analysing the role of ChatGPT in improving student productivity in higher education. *Journal on Education*, *5*(4), 14886-14891.

Goswami, A., & Dutta, S. (2015). Gender differences in technology usage—A literature review. *Open Journal of Business and Management*, 4(1), 51-59.

Kim, A., Muhn, M., & Nikolaev, V. (2023). From Transcripts to Insights: Uncovering Corporate Risks Using Generative AI. *Becker Friedman Institute Working Paper No.2-23-132, University of Chicago*.

Kim, D. G., & Moon, A. (2024). From helping hand to stumbling block: the ChatGPT paradox in competency experiment. *Applied Economics Letters*, 1-5.

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

Raman, R., Mandal, S., Das, P., Kaur, T., Sanjanasri, J. P., & Nedungadi, P. (2024). Exploring University Students' Adoption of ChatGPT Using the Diffusion of Innovation Theory and Sentiment Analysis With Gender Dimension. *Human Behavior and Emerging Technologies*, *2024*(1), 3085910.

Rutinowski, J., Franke, S., Endendyk, J., Dormuth, I., Roidl, M., & Pauly, M. (2024). The Self-Perception and Political Biases of ChatGPT. *Human Behavior and Emerging Technologies*, *2024*(1), 7115633.