

From Helping Hand to Stumbling Block: the ChatGPT Paradox in Competency Experiment*

Duk Gyoo Kim[†] Ahram Moon[‡]

January 23, 2024

Abstract

We ran a controlled laboratory experiment to examine whether ChatGPT’s aid can increase the participants’ performance in three different—reading and writing, mathematical problem-solving, and computational thinking—tasks. We find that the math score significantly *decreases* with ChatGPT’s assistance. This result is mainly because the low-ability subjects couldn’t discern the hallucinated answers with the correct ones, and it contests the general idea that ChatGPT can complement the workers with less expertise.

JEL Classification: C91, J24, O33, D83

Keywords: Laboratory experiment, ChatGPT, Labor productivity

1 Introduction

It is crucial to explore the possibilities and circumstances in which generative AI can be utilized in the labor market, not only for simple automation, but also for augmenting human capacities. Automation, underpinned by technologies ranging from software to robots and AI, progressively extends its scope from routine tasks to more complex and non-routine activities. Automation technologies replacing routine tasks lead to job displacement (Autor et al., 2003; Acemoglu and Autor, 2011; Goos et al., 2014) and create new jobs and higher labor productivity, triggering reinstatement effect (Acemoglu and Restrepo, 2019). Social intelligence and creativity are traditionally considered as bottlenecks to automation, but the advent of generative AI applications such as ChatGPT may ease these bottlenecks (Frey and Osborne, 2024). Thus, the impact of such AI on the labor market depends on its developmental and adoption pathways, and the quantitative changes in employment will rely on the relative magnitude of these effects. In essence, generative AI might lead to job displacement or reinstatement depending on whether it augments human capabilities or merely automates repetitive tasks.

Would ChatGPT’s aid increase the performance of skills relevant to the labor productivity? If so, would the high-ability workers get benefited more because ChatGPT supplements the skill, or the low-ability workers get benefited more because it complements the low ability?

*This work was partly supported by Institute of Information & Communications Technology Planning & Evaluation grant funded by the Ministry of Science and ICT (No.RS-2023-00215734). Approval for human subjects research was obtained from the Internal Review Board of Sungkyunkwan University (No. 2023-06-032).

[†]Department of Economics, Sungkyunkwan University. Email: kim.dukgyoo@skku.edu

[‡]Korea Information Society Development Institute. Email: ahmoon@kisdi.re.kr

To answer these questions, we conducted a controlled laboratory experiment examining the treatment effect of ChatGPT’s assistance on three—reading and writing, mathematical problem-solving, and computational thinking—tasks. Those three task domains represent core competencies for labor productivity. The treatment group was asked to use ChatGPT to perform the provided tasks, while the control group was not allowed to use it.

We found that the reading and writing task score in the treatment group is not statistically different from that in the control group, but the score distribution is more left-skewed with the same support, suggesting that the subjects with intermediate ability get benefited, but the high or low-ability subjects don’t. However, ChatGPT’s assistance significantly decreases the average score of the math task. This result is mainly because the subjects with low math ability couldn’t discern ChatGPT’s plausible (but incorrect) responses with the correct ones. That is, the subjects lacking the ability to check the validity of ChatGPT’s responses are likely to rely more on them, leading to lower performance. The average computational thinking task score of the treatment group is statistically insignificant compared to the control group. Overall, ChatGPT’s assistance was not beneficial, and sometimes harmful, for low-skilled participants. It can be paradoxical that those who need ChatGPT’s assistance the most should rely less on it. It contests the widespread belief that ChatGPT can augment the labor force with less expertise.

Our study contributes to the fast-growing literature on the economic applications of ChatGPT. Studies exploring rationality of machines and potentials for human-machine collaboration through generative AI indicate that the advancement of Large Language Models is showing capabilities for rational decision-making and reasoning (Webb et al., 2023; Kim et al., 2023). Furthermore, the applications to cognitive and knowledge-intensive tasks such as writing and counseling seem to improve overall productivity and reduce productivity disparities (Noy and Zhang, 2023; Brynjolfsson et al., 2023). Emerging research also identifies the functional limitations of generative AI. Chen et al. (2023) report that ChatGPT’s level of rationality varies by context, highlighting the limitations of it. Dell’Acqua et al. (2023) confirm that productivity can be hindered by GPT utilization patterns. Our study stands out in the existing literature by contemplating whether the rationality of ChatGPT can support fundamental human skills, rather than focusing on job-oriented tasks.

2 Experimental Design

Figure 1 summarizes the process of the experiment. The participants, Korean-speaking university students aged 19 or older with diverse academic disciplines, were randomly divided into one of two groups and asked to solve problems designed to assess their competence in two of the following three tasks: reading and writing (W-task hereinafter), mathematical problem-solving (M-task), and computational thinking (C-task). These tasks are fundamental skills applied to most occupational settings. The control group was allowed to use the Internet except for ChatGPT, while the treatment group was instructed to utilize ChatGPT actively. To give a common experience with ChatGPT, the treatment group participants copied and pasted the prepared three prompts into ChatGPT before performing the required task. To symmetrize the control group’s experimental procedure, the control group participants practiced advanced search functions by copying and pasting the prepared three queries into the Google search bar. To measure the participant’s competency, we consciously created test questions using existing competency assessment tools for college students and adults. For the C-task, we asked participants to complete natural language algorithms, pseudo codes, and logic flowcharts, to

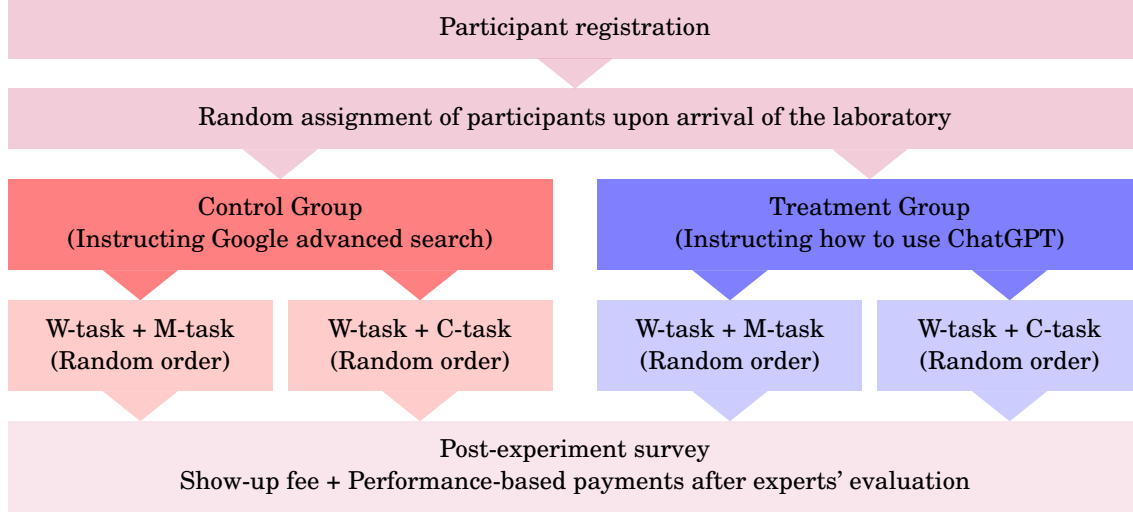


Figure 1: Experimental Design

ensure that participants without programming language experience could write their responses. Participants received performance-based payments which vary from 13,500KRW (about 10.24USD as of September 1, 2023) to 27,000KRW. The experiment was conducted in September and October 2023. See the [Online Appendix](#) for more detailed descriptions of the experiment and the test questions.

3 Results

This section reports the results of our experiment which involved a total of 349 participants, focusing on how much the task scores improved with ChatGPT's assistance. Figure 2 shows the average scores of three tasks obtained by participants in each control and treatment group.

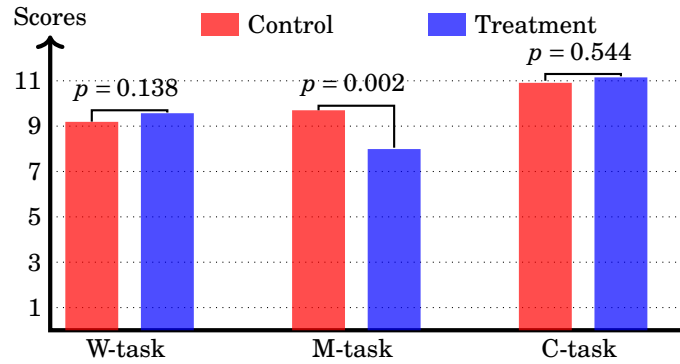


Figure 2: Average Scores of Each Task

In the W-task, the null hypothesis that the average score of the treatment group is the same as the control group cannot be rejected (t-test, p -value=0.138), but the null hypothesis that the score distribution of the control group and the treatment group is the same is rejected at the 5% significance level (Wilcoxon rank-

sum test, $p\text{-value}=0.0489$). It implies that using ChatGPT did not lead to an overall improvement in reading and writing scores, but a shift in the distribution. The first panel of Figure 3 shows the empirical cumulative distributions of the W-task scores for the control and treatment groups. In the interval between the bottom 40% and 95%, the distribution of the control group (dashed line) is to the left of the treatment group (solid line), suggesting that participants with competence in this interval achieved better scores with ChatGPT’s assistance.

Result 1. *Using ChatGPT weakly improves performance in the reading and writing task for participants with intermediate-level proficiency.*

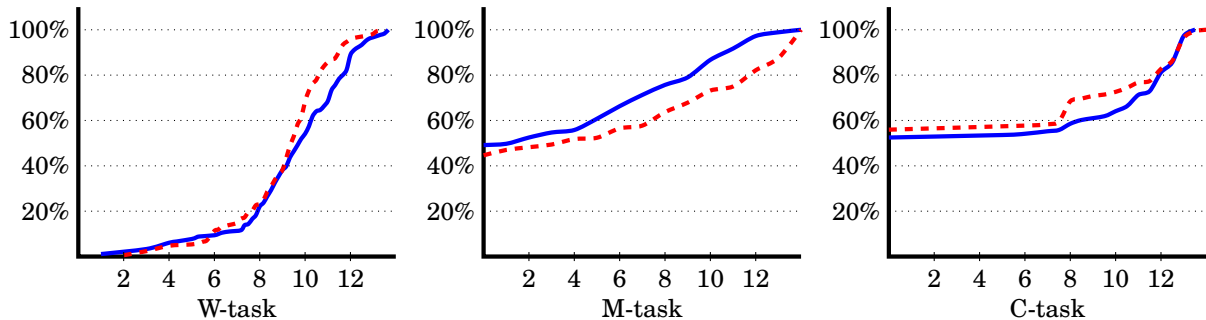


Figure 3: Empirical CDFs of Each Task

For the M-task, the average score of the treatment group is also significantly lower than the control group ($p\text{-value}=0.002$). This suggests that the average score *decreased* when the participants used ChatGPT to solve mathematical problems.

We claim this is because ChatGPT generates natural language responses that *seem* to fit the question, rather than analytically solving the problem, which often leads to incorrect answers. Participants with weaker mathematical problem-solving skills are less able to discern ChatGPT’s hallucinated answers from the correct ones, and this inability can adversely affect their performance. To support this claim, Figure 4 illustrates how ChatGPT can adversely affect the M-task. When one M-task question was typed on January 14, 2024, ChatGPT generated the correct sequence of steps but produced an erroneous answer. The wrong answer is deceptively persuasive, so low-skilled participants, believing that ChatGPT’s answers are convincing, lack the incentive to reevaluate them. ChatGPT provided a correct response (on the right of Figure 4) on the same date, but the fundamental issue remains: whether the users can discern the incorrect but seemingly plausible responses from the correct ones. It implies that the users with less expertise—who might need generative AI’s aid more—should not rely on it unless they become capable of evaluating the validity of its responses.

Result 2. *Using ChatGPT decreased participants’ performance in the mathematical problem-solving task.*

Regarding the C-task, no significant difference was found between the control and treatment groups. The null hypothesis that the data were drawn from the same population is not rejected ($p\text{-value}=0.482$), and the mean scores show a statistically insignificant difference ($p\text{-value}=0.544$). When assessing participants’ computational thinking skills, many of the questions required them to complete pseudo-code or natural language algorithms making their performance less dependent on previous programming languages exposure. These questions are not in a format that one can merely copy and paste into ChatGPT, so an ideal user would need to

ChatGPT 3.5 ▾



You

You paid 6,600 KRW for kiwis that cost 300 KRW each and apples that cost 700 KRW each. What is the maximum number of kiwis and apples that can be purchased together if more than one of each is purchased?



ChatGPT

To find the maximum number of kiwis (k) and apples (a) that can be purchased together, considering the given constraints, we can set up an equation based on the cost:

You paid 6,600 KRW for kiwis that cost 300 KRW each and apples that cost 700 KRW each.

The total cost (C) can be expressed as follows:

$$C = 300k + 700a$$

Given that $C = 6,600$, the equation becomes:

$$300k + 700a = 6,600$$

Now, we want to find the maximum values for k and a such that the equation is satisfied.

To simplify, we can divide the entire equation by 100:

$$3k + 7a = 66$$

Now, we need to find integer values for k and a that satisfy this equation. The maximum values for k and a occur when $k = 22$ and $a = 0$. This is because if $k = 23$, the left side of the equation would exceed 66.

So, the maximum number of kiwis and apples that can be purchased together, given the constraints and that more than one of each is purchased, is $k = 22$ and $a = 0$. In this case, only kiwis are purchased.



ChatGPT 3.5 ▾



You

You paid 6,600 KRW for kiwis that cost 300 KRW each and apples that cost 700 KRW each. What is the maximum number of kiwis and apples that can be purchased together if more than one of each is purchased?



ChatGPT

Let's denote the number of kiwis as x and the number of apples as y . The cost equation can be expressed as:

$$300x + 700y = 6600$$

Now, we want to find the maximum values of x and y that satisfy this equation. One way to approach this is to find the greatest common divisor (GCD) of 300 and 700, and then see if the total cost, 6600, is divisible by the GCD.

The GCD of 300 and 700 is 100. So, we divide both sides of the equation by 100:

$$3x + 7y = 66$$

Now, we need to find integer solutions for x and y that satisfy this equation and maximize their values. One solution is $x = 15$ and $y = 3$. This satisfies the equation, and it's a valid solution because both x and y are positive.

So, the maximum number of kiwis and apples that can be purchased together is 15 kiwis and 3 apples.



Figure 4: Different Answers to the Same Query
Used the default GPT-3.5. The same prompt was queried on the same day, Jan 14, 2024.

modify the question to make it easier for ChatGPT to process.¹ However, according to the post-experiment survey, only 6.8% of the treatment group participants answered they "modified and typed the question into ChatGPT, checked that ChatGPT's answer was correct, and submitted it". It is well known that proper prompts are necessary for proper ChatGPT's responses, and without mastering effective prompts, ChatGPT's aid doesn't significantly enhance competency.

Result 3. *Using ChatGPT without proper prompts does not improve performance in the computational thinking task.*

4 Concluding Remarks

Reading and writing, mathematical problem-solving, and computational thinking are core competencies for labor productivity. We conducted a controlled experiment to examine whether ChatGPT helps in those skills. Ultimately, this study is part of the broader examination about whether generative AI apps can positively affect labor productivity and a more flexible labor market.

We found using ChatGPT improved reading and writing performance of the participants with intermediate skills, consistent with [Noy and Zhang \(2023\)](#). However, in the mathematical problem-solving task, ChatGPT *hampered* participants' performance. This suggests that when ChatGPT's responses appear to be plausible but incorrect, using ChatGPT can be detrimental unless the user can validate ChatGPT's responses. Paradoxically, if a low-skilled person is incompetent to evaluate ChatGPT's outputs, she might consider not relying on them. Also, if one has already possessed adequate skills to perform the relevant tasks, the benefits of using ChatGPT may be negligible. In the computational thinking task, ChatGPT's assistance was limited. One clear takeaway is that similar challenges could arise in several other domains if the abilities to subjectively judge the quality of ChatGPT's responses and to properly compose prompts are not equipped beforehand. Overall, these results contest the widespread idea that generative AI applications' aid can positively affect labor productivity.

References

- Acemoglu, Daron and David Autor**, "Chapter 12 - Skills, Tasks and Technologies: Implications for Employment and Earnings," in David Card and Orley Ashenfelter, eds., *Handbook of Labor Economics*, Vol. 4, Elsevier, 2011, pp. 1043–1171.
- **and Pascual Restrepo**, "Automation and New Tasks: How Technology Displaces and Reinstates Labor," *Journal of Economic Perspectives*, May 2019, 33 (2), 3–30.
- Autor, David H., Frank Levy, and Richard J. Murnane**, "The Skill Content of Recent Technological Change: An Empirical Exploration*," *The Quarterly Journal of Economics*, 11 2003, 118 (4), 1279–1333.
- Brynjolfsson, Erik, Danielle Li, and Lindsey R. Raymond**, "Generative AI at Work," Working Paper 31161, National Bureau of Economic Research April 2023.
- Chen, Yiting, Tracy Xiao Liu, You Shan, and Songfa Zhong**, "The emergence of economic rationality of GPT," *Proceedings of the National Academy of Sciences*, 2023, 120 (51), e2316205120.

¹We ensure that ChatGPT generates correct answers to all questions when using proper prompts.

- Dell’Acqua, Fabrizio, Edward McFowland, Ethan R. Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Kraye, François Cadelon, and Karim R. Lakhani**, “Navigating the jagged technological frontier: field experimental evidence of the effects of AI on knowledge worker productivity and quality,” *Technology & Operations Mgt. Unit Working Paper 24-013*, Harvard Business School 2023.
- Frey, Carl Benedikt and Michael A. Osborne**, “Generative AI and the Future of Work: A Reappraisal(forthcoming),” *Brown Journal of World Affairs*, 2024.
- Goos, Maarten, Alan Manning, and Anna Salomons**, “Explaining Job Polarization: Routine-Biased Technological Change and Offshoring,” *American Economic Review*, August 2014, *104* (8), 2509–26.
- Kim, Jeongbin, Matthew Kovach, Kyu-Min Lee, Euncheol Shin, and Hector Tzavellas**, “Learning to be Homo Economicus: Can an LLM Learn Preferences from Choice Data?,” 2023. Working Paper.
- Noy, Shakked and Whitney Zhang**, “Experimental evidence on the productivity effects of generative artificial intelligence,” *Science*, 2023, *381* (6654), 187–192.
- Webb, Taylor, Keith J Holyoak, and Hongjing Lu**, “Emergent analogical reasoning in large language models,” *Nature Human Behaviour*, 2023, *7* (9), 1526–1541.