GenAI as an Exoskeleton: Experimental Evidence on Knowledge Workers Using GenAI on New Skills

Emma Wiles[1*], Lisa Krayer[2*], Mohamed Abbadi[2], Urvi Awasthi[2], Ryan Kennedy[2], Pamela Mishkin[3], Daniel Sack[2], Francois Candelon[2]

[1*]Boston University, 595 Commonwealth Ave, Boston, 02215, MA.
[2]Henderson Institute, BCG, 200 Pier Four Blvd, Boston, 02110, MA.
[3]Economic Impacts Research, OpenAI, 3180 18th Street, SF, 94110, CA.

Abstract

"Reskilling" often refers to the process by which workers acquire new capabilities or knowledge, enabling them to move jobs or industries as the demands of the market changes. This paper demonstrates that while generative artificial intelligence (GenAI) can act as an "exoskeleton,*" enhancing workers' capabilities while they attempt new skills, these gains are dependent on the continued use of the technology. When the "exoskeleton" is removed, little to no knowledge is retained independently, revealing that the newfound capabilities are temporary and reliant on the external support provided by GenAI. We run a randomized controlled trial on "reskilling" with GenAI by providing Boston Consulting Group (BCG) consultants with access and training in using ChatGPT to solve technical problems. We measure their performance on real data science tasks outside their skill sets, which cannot be independently solved by ChatGPT. Treated workers score 49, 20, and 18 percentage points higher than those in the control group on the three tasks and perform close to the level of real BCG data scientists on two of the three tasks. However, treated workers are no better at answering technical questions without the use of ChatGPT post-experiment, suggesting their demonstrated newfound technical capabilities do not imply knowledge acquisition. These results suggest that GenAI can be used to help workers reskill to meet the greater technical demands of the labor market but that the work of nontechnical workers using GenAI is not interchangeable with that of data scientists.[†]

[*] "Exoskeletons are wearable structures that support and assist movement, or augment the capabilities of the human body." [1]

[†]Corresponding authors: emma.b.wiles@gmail.com, krayer.lisa@bcg.com. The full details of the experiment, including the survey text, can be found in our online appendix. The details

1 Introduction

The rapid advances in generative artificial intelligence (GenAI) and its widespread deployment have sparked both excitement and concern about its potential impact on the workforce. These models' increasing capabilities in both automating and augmenting aspects of knowledge work raise questions about the impacts of technologies on the labor market—how jobs can be redesigned and how workers can adapt to the changing demands of employers. A key strategy for mitigating the negative effects of automation on workers is to "reskill" them into new and more in-demand skills [2, 3]. While recent studies have explored the effects of GenAI on worker performance in tasks within their existing skill sets [4, 5], there remains a gap in understanding how GenAI can be used to help workers acquire the new skills necessary to adapt to the changing demands of the labor market. Reskilling has been defined as workers gaining the knowledge and capabilities to take on new roles [6]. We differentiate between two types of skill acquisition—one based on workers capabilities in completing tasks and another where they acquire deeper knowledge, enabling them to complete tasks independently of technological aids. In general we will define skills in line with [7], or as "a worker's endowment of capabilities for performing various tasks." In this paper, we introduce the analogy of GenAI as an "exoskeleton," a tool that temporarily enhances non-technical workers' technical capabilities. Like an exoskeleton, GenAI enables non-technical workers to perform complex tasks outside their existing skill sets, but once the technology is removed, the enhanced capabilities disappear.

In this paper, we run a randomized controlled trial (RCT) involving 986 consultants at the BCG, a global management consulting firm. We show that providing nontechnical professionals with access to and training in GenAI chat-based tools (here, we use ChatGPT) can significantly enhance their ability to accurately complete data science tasks for problems that GenAI can perform and explain well but cannot independently solve. Participants are randomly assigned to either a treatment group, which receives access to and training in ChatGPT, or a control group, which receives training on using Stack Overflow and other resources commonly used by data scientists. Each participant is randomly assigned two of three 90-minute tasks. The coding task, statistical understanding task, and prediction task are all developed in collaboration with the Economic Impacts Research Team at OpenAI. These tasks are specifically designed to be challenging for ChatGPT to solve independently, requiring human input at multiple steps. Treated workers have access to ChatGPT's latest model, GPT-4, and are encouraged to use it while attempting to complete the tasks.

For the 487 participants who submitted output for these tasks, we measure the treatment and control group workers' performance by comparing their output to a benchmark set by the performance of data scientists (who perform the tasks without access to GenAI) who regularly perform these types of tasks. This approach allows us to directly test how much nontechnical workers armed with GenAI chatbots as

of our preregistration can be found here. We used Gitub Copilot, Claude, and ChatGPT for light copyediting and graphics creations. L. Krayer, M. Abbadi, U. Awasthi, R. Kennedy, D. Sack, and F. Candelon are or were employees of BCG. P. Mishkin is an employee of OpenAI. E. Wiles was a paid consultant for BCG during the running of the experiment.

their guide can do relative to workers in data science roles. While the expectations of data scientists vary from firm to firm, the tasks we study are representative of work that many data scientists commonly perform. However, it should be noted that we do not assess the impact of GenAI on workers' performance on their ability to complete end-to-end data science problems nor do we test more advanced techniques like deep learning.

Treated workers perform better on all three tasks than the control group and just as well as the data scientists on the coding task. Treated workers are more likely to submit answers for the three tasks and take less time to submit answers for the prediction and coding problems than do those workers in the control group. On scores normalized to 0, which is the average performance of data scientists, treated workers perform 49, 20, and 18 percentage points better than those in the control group on coding, statistics, and prediction problems, respectively. On the prediction problem, treated workers perform better than do control group workers, but there is a large gap between their performance and the benchmark set by the data scientists. We find that treated workers who begin the experiment with some coding skills perform just as well as do the data scientists on coding and statistics tasks.

Following the experiment, we find that workers in the treatment group are not able to answer technical questions on sections where they are not allowed to use ChatGPT. We also find that compared with those in the control group, treated workers exhibit greater overconfidence in the current capabilities of ChatGPT. In fact, following the experiment, treated workers perform worse than do control workers at estimating which types of problems ChatGPT can and cannot solve.

We contribute to the new literature on the effects of GenAI on worker productivity and performance. In one study, customer service agents given access to GenAI suggestions are able to resolve customer complaints more efficiently, and with no loss in quality, than are those without such access [5]. In another example, while workers with access to GenAI can complete tasks within the range of the model's ability faster and more accurately than can workers without such access, they perform worse on tasks outside the model's abilities [8]. We provide the first example that we are aware of that GenAI can also improve the productivity of workers on complex tasks *outside* of their skill sets.

Second, we contribute to the literature on job training and reskilling workers. Automation from AI is a serious concern among academics and policymakers, with reskilling workers cited as one of the primary strategies for workers to keep from being displaced [2, 3]. There is evidence that the earnings premium for technology-intensive college subjects declines faster than that for more general subjects [9], suggesting benefits for technical workers to be able to learn new skills. Given the evidence that GenAI is a general-purpose technology [10] and an effective teacher [11], it may allow workers to flexibly solve new types of problems as they emerge.

Finally, we contribute to the literature on the limitations of human-AI interactions. Previous studies have shown that people who are given access to AI often cannot judge the quality of AI outputs [12]. Employers who are given access to AI-written first drafts of job postings produce more generic job postings that are less likely to lead to a hire [13]. [14] finds that recruiters take AI's suggestions, even when such suggestions are not correct. Our finding that workers with training in ChatGPT are overconfident in the model's abilities complements these findings, and we show that exposure to ChatGPT worsens the ability of workers to predict the model boundaries.

2 Methods

2.1 Experimental design

We report the results from a large RCT run on consultants of BCG, a global managerial consulting firm, to test whether nontechnical knowledge workers can perform data science work with the help of ChatGPT-4, preregistered on March 13, 2024. The experiment took place in late March and early April of 2024. In the recruitment phase, all current BCG consultants were sent an email inviting them to participate in a study on upskilling and GenAI. Those who registered were surveyed on their demographics, programming and ChatGPT skills, technology openness, trust in ChatGPT, and learning orientation [15–17]. Simultaneously, BCG data scientists were invited to participate in a similar exercise, where they would simply complete the tasks used in the experiment without using AI. The output from the 44 data scientists who participated served as the benchmark for the "typical performance of a data scientist."

After the registration survey, workers were randomly assigned to either a treatment group or a control group. Random assignment was stratified across gender, location, role (i.e., associate or consultant), coding skills, college degree (i.e., bachelor's, master's, or Ph.D.), and experience with ChatGPT for coding. The initial sample of 986 workers was split equally into treatment and control groups, of which 573 workers began the survey. Our analysis sample consisted of 487 participants who completed the two tasks.

The experiment contained four phases (Appendix Figure A1):

1. **A pre-experiment survey** consisting of questions on participants' subjective coding skills and GenAI usage and a series of questions for which they had to predict whether or not GPT-4 could correctly answer them.
2. **A 15-20 minute tailored training** to each experimental condition. The treatment group received training on ChatGPT prompting for technical problems, while the control group received training on leveraging Google and commonly used websites such as Stack Overflow and Khan Academy. Both trainings involved a combination of videos and interactive practice.
3. **Three data science tasks** designed to be more technical than participants' current roles. These problems were outside the expected skill set of knowledge workers and often representative of those tasks often undertaken by data scientists.
4. **A postexperiment survey** that included a series of technical questions related to the tasks that all participants had to answer without the help of ChatGPT.

2.2 Task descriptions

The first task was a coding task testing workers' ability to write code in Python. Workers were required to write and submit Python code to clean and answer questions about two datasets. These workers were likely to know the basics of data cleaning using tools like Microsoft Excel and Alteryx for their roles as consultants. However, more than 60% reported either having never coded before or knowing only the basics of coding.

The second task measured statistical understanding, the ability to fact-check GenAI's statistical analysis recommendations, and the ability to interpret the output of machine learning models. In this task, workers were guided through a series of potential GenAI outputs to create a model that predicted whether a couple would take out a mortgage based on historical data. They were then asked a series of questions on how they would respond to various situations while solving machine learning problems, such determining the cause of poor model performance.

The third task was a prediction task, which required workers to create a predictive model using historical data on international football games to develop an investment strategy. Their goal was to assess the "predictability" of their model; i.e., they needed to assess how reliable their model was for investment decisions. This task involved testing participants' ability to understand predictive modeling and correctly identify when and how to apply machine learning models.

The tasks were created by BCG data scientists in collaboration with the Economic Impacts Research Team at OpenAI. They were designed such that ChatGPT is a useful guide, but is unable to correctly answer any of the questions independently. For all three tasks, if the participant let ChatGPT answer the question on its own, then the answer would be incorrect. Each task was intended to take 90 minutes to complete, and thus, to avoid fatigue, we assigned each participant two of the three tasks randomly, with the task order being randomized.

2.3 Analysis sample and main outcomes

For most of the experimental results, we included workers who submitted answers for both their first and second tasks. This sample included 487 workers across the treatment and control groups. In the first panel of Table 1, we show attrition from the registration survey to the analysis sample. While there was some attrition at each stage, it was not significantly different between the treatment and control groups. We show that those who completed the experiment and the attritors did not differ between the treatment and control groups, as shown in Appendix Table B2. Nevertheless, in the case of differential attrition, we report Lee Bounds [18] on our main results in Appendix Table B15 and find that all the main results retain their significance.

Our first set of outcomes is related to how the workers performed on the three data science tasks. We measured workers' likelihood of finishing each task, how long it took them to complete each task, and how they performed in terms of correctness and their process of completing each task.

For the coding task, there was a conservatively defined correct answer, which received a "1," while any other answer received a '0.' To obtain the correct answer, a worker must take 10 correct steps, and thus, we also measured the percentage of the correct steps taken by the worker. The statistical understanding task included multiple-choice questions with both correct and incorrect answers and open-ended questions. For the prediction task, participants' output is a vector of probabilities that

Table 1 Building the experimental sample

Flow from initial allocation into analysis sample				
	Total (N)	*Treatment (N)*	*Control (N)*	*P-value*
Total workers allocated	983	493	493	
↪ began survey	573	298	275	0.33
↪ completed first task	511	270	241	0.19
↪ completed second task	487	260	227	0.13
Effect of treatment on submitting each task				
↪ completed coding task	300	167	133	0.01
↪ completed statistics task	329	175	154	0.10
↪ completed predictions task	298	162	136	0.04

Notes: This table reports the means and standard errors of various pretreatment covariates for the treatment and control groups. The first panel describes the flow of the sample from the allocation to the sample we use for our main experimental analysis. The complete allocated sample is described in the first line, with each following line defined cumulatively. Each worker was assigned two tasks, and the following lines compare the number of workers who submitted any work for each of the tasks. Those who completed both tasks made up the main experimental sample. The second panel provides the sample size for each treatment group and each task and shows treatment effects on whether or not workers completed each task.

were compared to the benchmarks set by the data scientists' output (see Section A.3 for details). We used GPT-4 Turbo API to automatically grade the parts of the three tasks that did not have a binary right or wrong answer, for example, the free response portion of the statistics task. This automation was rigorously validated by a data scientist.

Our second set of outcomes tested whether workers appeared to retain knowledge without ChatGPT. We measured this through questions in the postexperiment survey, in which neither group was allowed to use ChatGPT to answer. Here, we hoped to test whether any new abilities that workers gained while using ChatGPT in a time-pressured environment could be retained without the tool.

Third, we tested whether experience using ChatGPT helped workers better gauge the bounds of its abilities. In the pre- and postexperiment surveys, we provided workers with a series of problems and asked them "How likely is GPT-4 to solve this problem correctly?" [19]. We hypothesized that after completing these tasks, workers in the treatment group would be better at forecasting which types of problems ChatGPT can reliably solve compared to those in the control group.

3 Results

3.1 Treated workers performed better on data science tasks

Figure 1 shows that treated workers were able to perform all three data science problems better than were their counterparts in the control group. We normalize the performance for each task so that 0 is equivalent to the average performance of the data scientists' benchmark.

On the coding task, treated workers performed 49 percentage points better than did workers in the control group. Furthermore, the benchmark set by the data scientists was within the 95% confidence interval (CI) of the treated workers' scores. In the 90 minutes allotted for the coding task, no worker in the control group achieved a perfect score, while five workers in the treatment group achieved a perfect score. Because of this low level of variation, we make the primary outcome for the coding task the percentage of steps that the worker took correctly, i.e., how many of the correct steps he or she took in the allotted time. On average, control group workers were 63 percentage points away from the performance of data scientists, while treatment group workers were only 14 percentage points away. Not only did workers in the treatment group perform much better, but the 95% CI around their scores includes the benchmark set by the data scientists.

On the statistical understanding problem, the treatment group performed only 12 percentage points lower than the benchmark, although we can reject with 95% confidence that they performed as well as the data scientists. In the control group, workers' performance was on average 32 percentage points lower than that of the data scientists.

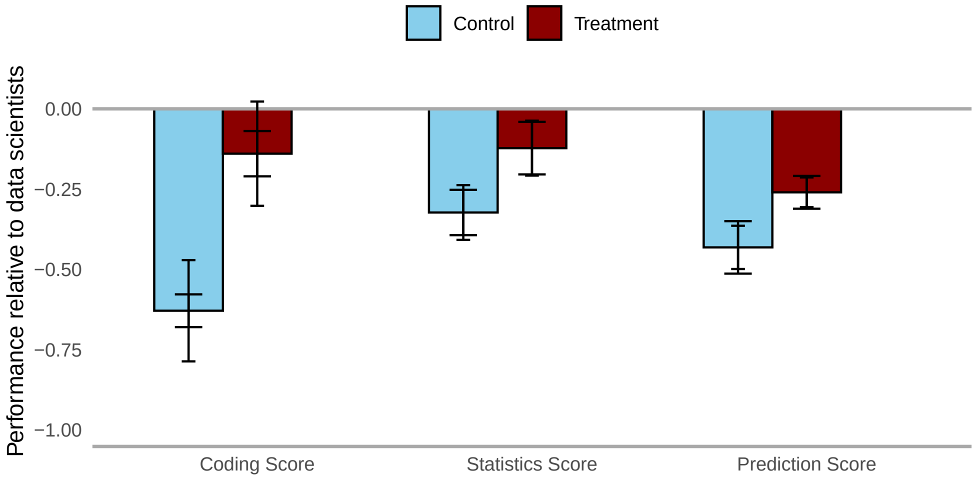
On the prediction problem, compared with control group workers, treated workers obtain results 17 percentage points closer to those of data scientists. We use the outputs provided by the data scientists as the benchmarks since there are multiple valid predictive models and no single ground truth. Each worker's score for the prediction problem is the minimum absolute error between his or her vector and the benchmark vectors. For ease of interpretation, we multiply this error term by -1 so that a score of 0 implies that workers submitted a correct vector and a score of -1 denotes a vector that is orthogonal to any baseline answer. The control group had an average score of -0.43. The treatment group performed better, with an average score of -0.26, but we can reject with 95% confidence that they performed as well as did the data scientists.

3.2 Workers with coding experience performed as well as did the data scientists on two of the three tasks

Treated workers who reported being comfortable with coding prior to the experiment performed almost as well as did the data scientists on two of the three tasks, as shown in Figure 2. In the control group, we see that prior coding ability is associated with better performance in coding and statistics tasks, while on the prediction task, control group workers of all coding backgrounds performed similarly. The treatment effects are largest for those with the least coding experience on the statistics and coding problems.

Among those with prior coding experience, the control group's average score on the coding task was 56 percentage points away from the data scientists' benchmark, while the treated workers scored a mere 5 percentage points less than the data scientists. On the statistics task, the control group's competent coders performed 21 percentage points away from the benchmark, while the treated workers scored 5 percentage points away. On the prediction task, competent coders in the treatment group performed 27 percentage points better than did those in the control group but remained 20 percentage points away from the data scientists' benchmark.

Fig. 1 Effect of AI treatment on workers' ability to solve data science problems



Notes: This plot reports the effect of the treatment on the workers' performance across the 3 tasks. The x-axis is the mean score for each treatment group on each set of problems, where 0 is the average for data scientists. The first outcome is the percentage of correct steps they take in answering the coding question. The second outcome is the sum of the workers' scores on each statistics question, divided by the total number of possible points. The third outcome is the score they received on the prediction problem, which is 1 minus their mean absolute error. The wider bars indicate the 95% CIs of the mean of the treatment group relative to the mean in the control group. The narrower bars indicate the 95% CIs of the treatment and control means compared to the data scientists' benchmark of 0. All standard errors are Huber–White robust. The sample includes all experimental participants who submitted a portion of each task for grading. The text of the problems can be found in Appendix Section A. Regression details for comparing treatment and control can be found in Appendix Table B3 and those for comparing the treatment arms to the benchmark in Appendix Table B9.

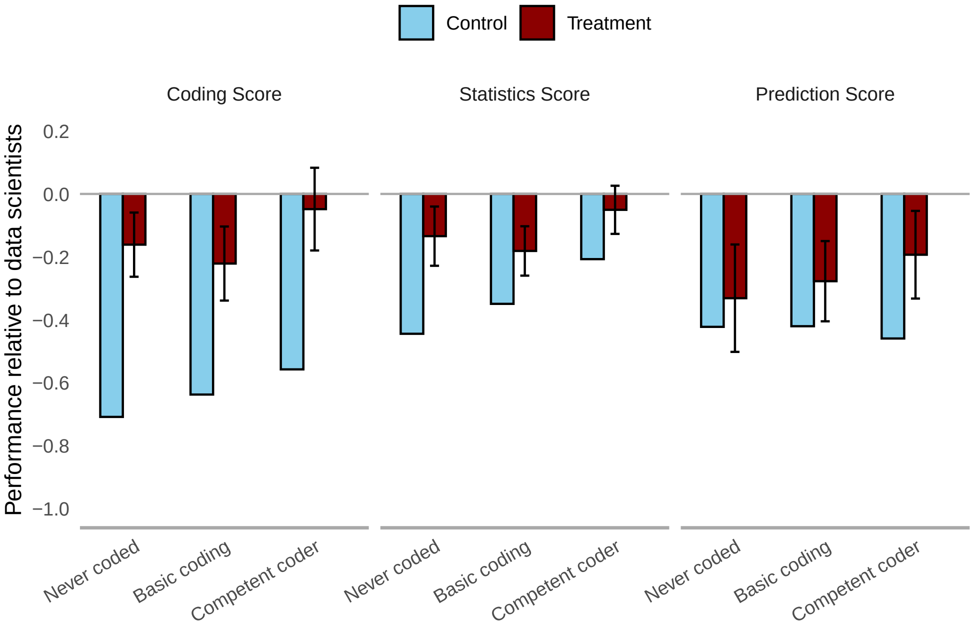
3.3 No impact on workers' ability to answer technical problems without the help of ChatGPT

Despite evidence that the use of AI improved workers' performance on the data science problems, after ChatGPT is taken away, they are no longer likely to be able to answer questions about probabilities, machine learning, or coding. In the postexperiment survey, workers are asked five questions on topics related to their tasks. Workers in both groups are instructed not to use ChatGPT to answer these questions. Figure 3 shows that the treated group performs no better than does the control group on these questions. To ensure robustness, we examine the heterogeneous treatment effects based on whether the worker performed a task directly related to the type of question. Our analysis suggests no direct effects, as illustrated in Table B10 in the Appendix.

3.4 Treated workers exhibit overconfidence in AI's current capabilities

Workers in the treatment group perform worse at estimating whether something is within ChatGPT's capabilities after the experiment. Before and after the experiment, we pose seven questions (or prompts) and ask the workers to report their opinions on

Fig. 2 Effect of AI treatment on workers' ability to solve data science problems according to their prior coding experience

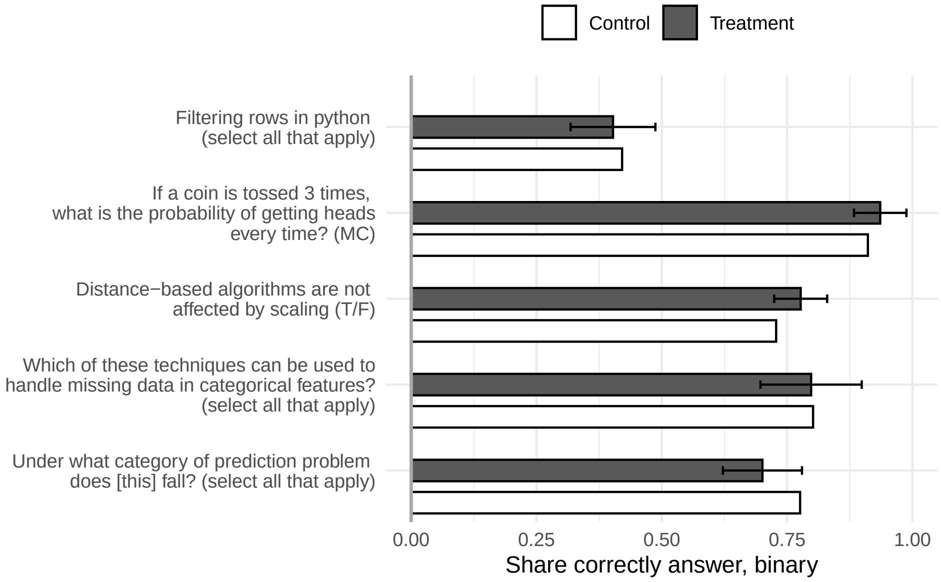


Notes: This plot reports the effect of the treatment on the workers' scores on each data science task according to their self-reported prior experience with coding. The x-axis is the mean score for each treatment group on each set of problems, where 0 is the benchmark set by the data scientists and -1 means that the worker received no points on the task. The first outcome is the percentage of correct steps that they take in answering the coding question. The second outcome is the sum of the workers' scores on each statistics question, divided by the total number of possible points. The third outcome is the score they received on the prediction problem, which is -1 times their mean absolute error. Huber–White robust standard errors are based on the difference between the means of the treatment and control groups, and we plot the 95% CIs around each estimate. Because the data scientists are all at least competent coders, we cannot compare the performance of each worker in the treatment group to that of the data scientists interacted with coding experience. The sample includes all experimental participants who submitted a portion of each task for grading. The text of the problems can be found in Appendix Section A. Regression details can be found in Table B11.

the likelihood of ChatGPT correctly answering the questions [19]. ChatGPT is unable to answer 5 of the questions given before the experiment and 4 of the questions given after the experiment.

Prior to the experiment, workers in each group have comparable levels of confidence in ChatGPT's ability to correctly answer similar questions. After the experiment, however, workers in the treatment group become significantly more optimistic and more incorrect about ChatGPT's capabilities, as shown in Figure 4. Workers in both groups are optimistic about ChatGPT's capabilities—the base rate for each question is between 65 and 78%. For all four of the questions that ChatGPT cannot answer, treated workers report a 5 to 10 percentage point greater likelihood that it can correctly answer each question. Surprisingly, the only two questions where the treatment and

Fig. 3 Effects of AI treatment on post-experiment data science knowledge without the use of GenAI



Notes: This analysis looks at the effect of treatment on workers' ability to correctly answer data science questions after the conclusion of the experiment. The x-axis is the mean probability of getting the correct answer for each treatment group. The y-axis has the text of each question, with the format of the answer. A 95% CI is plotted around each estimate. The text of the questions can be found in the "postsurvey" section of the Online Appendix. Regression details can be found in Appendix Table B7.

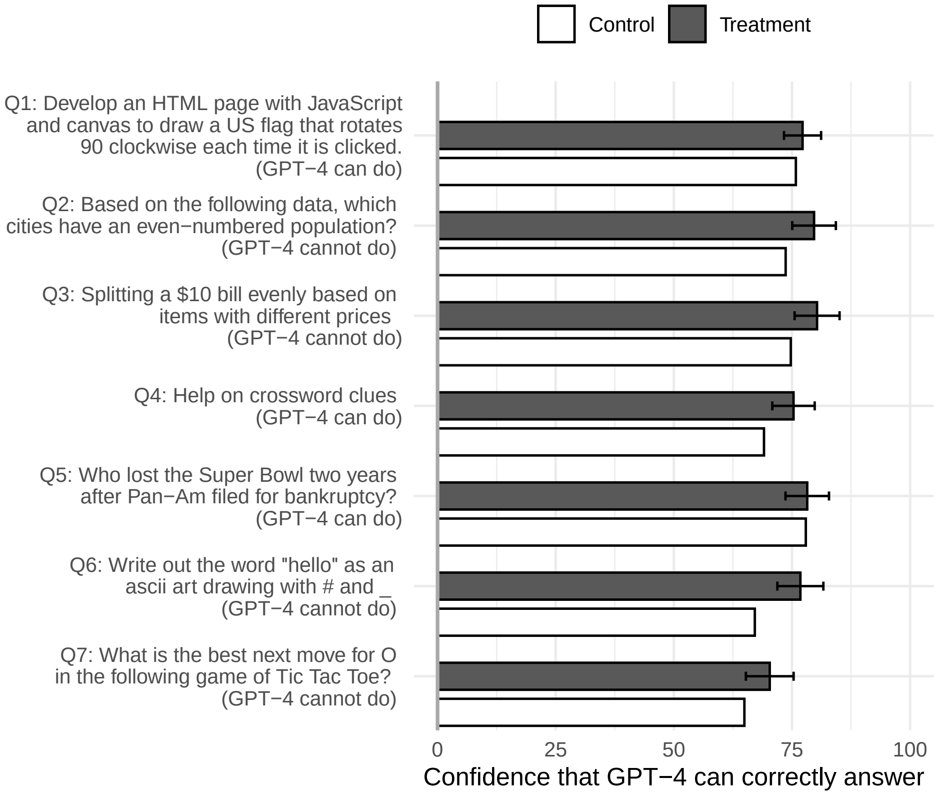
control groups are equally optimistic are two of the three questions that ChatGPT can correctly answer.

Given fears about overreliance on GenAI causing harm, we may be concerned that workers who are overconfident in ChatGPT perform worse or exhibit smaller treatment effects as they are less likely to effectively check the system's work. In Appendix Table B12, we interact the treatment effects with the pre-experiment measure of overconfidence. We define overconfidence based on their answers to the pre-experiment survey, where we ask workers to estimate whether the ChatGPT can solve various problems. Most of the workers exhibit overconfidence. For the 5 problems that ChatGPT cannot solve, 80% of the workers believe that ChatGPT can solve at least 3 problems. Therefore, we label workers as overconfident if they believe that ChatGPT can solve at least 4 of the 5 problems that it cannot solve. We find that overconfident workers have slightly weaker treatment effects than workers who are less overconfident; however, the former workers still exhibit large positive treatment effects.

4 Discussion

We run an RCT in the setting of a global management consulting firm to understand to what extent GenAI can be used to help nontechnical knowledge workers perform tasks, or "reskill," outside their current skill set, in this case, in performing data science tasks. We find that workers given access to and training in ChatGPT significantly outperform

Fig. 4 Effect of AI treatment on workers' predictions about the capabilities of ChatGPT



Notes: This analysis looks at the effect of treatment on workers' confidence in AI's capabilities. The x-axis is the difference in workers' confidence out of 100 that ChatGPT can answer the question correctly. A summary of the text of each question is on the y axis, with whether or not ChatGPT can obtain the correct answer. A 95% CI is plotted around each estimate. The text of the questions can be found in the "postsurvey" section of the Online Appendix [19]. Regression details can be found in Appendix Table B8.

those in the control group, with the largest effects seen on the coding task. Treated workers perform better on three different margins: treated workers are more likely to complete tasks, complete tasks in less time, and obtain higher scores conditional on completing tasks. In addition, workers perform nearly at the level of data scientists on the coding and statistics tasks. Despite these gains in the treated workers technical capabilities, we do not observe any difference in the levels of knowledge retention between the treated and control groups. These results suggest that demonstrating new capabilities is not equivalent to gaining new knowledge. These findings highlight the potential for AI itself to help workers adapt to the changing skill demands of the labor market. We provide empirical evidence of how AI-enabled reskilling can be used to help workers adapt to avoid possible job displacement from AI and automation [2, 3].

There is a large stream of literature that shows that GenAI increases productivity for tasks within knowledge workers' skill sets[5, 8, 20]. Here, we provide evidence that it can also be used to help workers perform new tasks outside their skill set. These results have implications for both workers and managers.

For knowledge workers, the results suggest that using GenAI can expand the range of jobs available to them. For hiring managers, the results suggest that they consider a wider pool of applicants who are proficient in GenAI. Managing this likely necessitates a reorganization of teams. For instance, teams may benefit from including someone with a deep understanding of data science to oversee and evaluate the work produced by less technical workers.

Our results also point to some important limitations of GenAI use for work outside one's skill set. While treated workers can complete the data science tasks with the aid of ChatGPT, they do not demonstrate any greater retention of data science knowledge. This finding suggests that there may be limits to the depth of genuine skill acquisition, at least in the short term. Prior research indicates that practice and repetition [21, 22] are crucial for skill acquisition. Although it is outside the scope of this study, the impacts on long-term learning are likely different after workers practice and are exposed long term to GenAI.

Moreover, we find that exposure to ChatGPT induces overconfidence in AI's abilities, with treated workers being more likely to believe that ChatGPT can solve problems that it actually cannot. This finding echoes those from other recent studies on human susceptibility to AI errors and overreliance on AI assistance[12, 13] and suggests that organizations think carefully about monitoring when workers are using GenAI to complete tasks outside their skill sets. Despite this finding, we do not see evidence that workers who are overconfident in GenAI's capabilities prior to the experiment perform any worse on the tasks than do other workers. While this suggests that workers' overconfidence does not cause any changes in their performance, in settings where the consequences of errors are more significant, organizations should implement quality control measures such as evaluating the output from GenAI-assisted workers.

We believe that this paper provides the first piece of evidence that GenAI can be used to widen the scope of work that workers can perform. However, there are some important limitations to these results. The tasks we study are representative of work that data scientists regularly perform, such as making predictions, cleaning and building data, and evaluating ML model outputs, but do not include more advanced techniques, such as deep learning, which are often part of real-world data scientists' jobs. Moreover, due to the limited scope of the experiment, we do not assess the impact of GenAI on performance in a more typical end-to-end data science problem.

Despite these limitations, our findings highlight the potential of GenAI as a tool for "reskilling." Fully realizing this potential while mitigating the risks of overconfidence and overreliance require ongoing research to aid in responsible implementation. Future research should address whether longer-term exposure to and practice with GenAI can help workers gain not only new capabilities but also a deeper learning of skills.

References

[1] Iberdrola: What are exoskeletons and how can they help us overcome our human limitations? Accessed: 2024-09-03 (2024). https://www.iberdrola.com/innovation/what-are-exoskeletons

[2] Acemoglu, D., Restrepo, P.: The race between man and machine: Implications of technology for growth, factor shares, and employment. American economic review **108**(6), 1488–1542 (2018)

[3] Djankov, S., Saliola, F.: The changing nature of work. Journal of International Affairs **72**(1), 57–74 (2018)

[4] Noy, S., Zhang, W.: Experimental evidence on the productivity effects of generative artificial intelligence. Available at SSRN 4375283 (2023)

[5] Brynjolfsson, E., Li, D., Raymond, L.R.: Generative ai at work. Technical report, National Bureau of Economic Research (2023)

[6] Li, L.: Reskilling and upskilling the future-ready workforce for industry 4.0 and beyond. Information Systems Frontiers, 1–16 (2022)

[7] Acemoglu, D., Autor, D.: Skills, tasks and technologies: Implications for employment and earnings. In: Handbook of Labor Economics vol. 4, pp. 1043–1171. Elsevier, ??? (2011)

[8] Dell'Acqua, F., McFowland, E., Mollick, E.R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Lakhani, K.R.: Navigating the jagged technological frontier: Field experimental evidence of the effects of ai on knowledge worker productivity and quality. Technical report, Harvard Business School Technology & Operations Mgt. Unit Working Paper (2023)

[9] Deming, D.J., Noray, K.: Earnings dynamics, changing job skills, and stem careers. The Quarterly Journal of Economics **135**(4), 1965–2005 (2020)

[10] Eloundou, T., Manning, S., Mishkin, P., Rock, D.: Gpts are gpts: An early look at the labor market impact potential of large language models. arXiv preprint arXiv:2303.10130 (2023)

[11] Mollick, E.R., Mollick, L.: New Modes of Learning Enabled by AI Chatbots: Three Methods and Assignments. Available at SSRN: https://ssrn.com/abstract=4300783 or http://dx.doi.org/10.2139/ssrn.4300783 (2022)

[12] Agarwal, N., Moehring, A., Rajpurkar, P., Salz, T.: Combining human expertise with artificial intelligence: Experimental evidence from radiology. Technical report, National Bureau of Economic Research (2023)

[13] Wiles, Emma, Horton, John: More, but worse: The impact of ai writing assistance on the supply and quality of job posts (2024)

[14] Dell'Acqua, F.: Falling asleep at the wheel: Human/ai collaboration in a field experiment on hr recruiters. Technical report, Working paper (2022)

[15] Agarwal, R., Prasad, J.: A conceptual and operational definition of personal innovativeness in the domain of information technology. Information Systems Research **9**(2), 204–215 (1998)

[16] Miron, E., Erez, M., Naveh, E.: Do personal characteristics and cultural values that promote innovation, quality, and efficiency compete or complement each other? Journal of Organizational Behavior **25**(2), 175–199 (2004)

[17] Jha, S., Bhattacharyya, S.S.: Learning orientation and performance orientation: Scale development and its relationship with performance. Global Business Review **14**(1), 43–54 (2013)

[18] Lee, D.S.: Training, wages, and sample selection: Estimating sharp bounds on treatment effects. Review of Economic Studies **76**(3), 1071–1102 (2009)

[19] Carlini, N.: A GPT-4 Capability Forecasting Challenge. https://nicholas.carlini.com/writing/llm-forecast/ (2023)

[20] Peng, S., Kalliamvakou, E., Cihon, P., Demire, M.: The impact of ai on developer productivity: Evidence from github copilot. arXiv preprint arXiv:2302.06590 (2023)

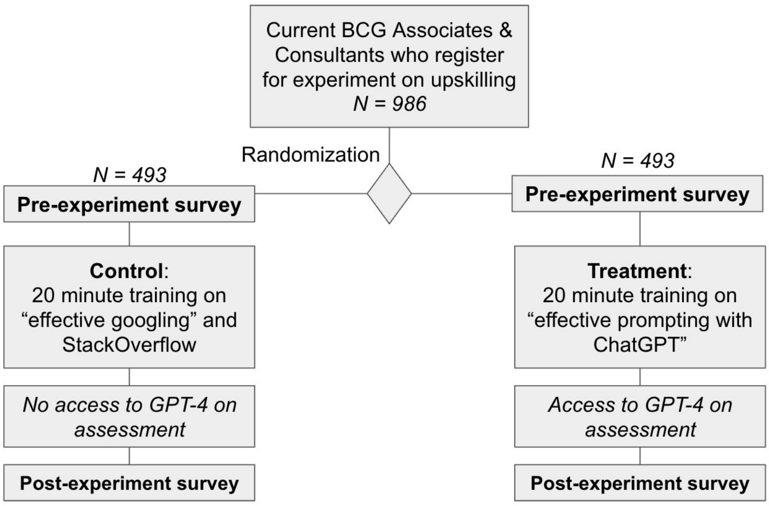
[21] Anderson, J.R.: Cognitive Skills and Their Acquisition. Psychology Press, ??? (2013)

[22] DeKeyser, R.: Skill acquisition theory. In: Theories in Second Language Acquisition, pp. 83–104. Routledge, ??? (2020)

Appendix A Experimental design and task overview

A.1 Experimental design

Fig. A1 Overview of the experimental design


Current BCG Associates & Consultants who register for experiment on upskilling
N = 986
Randomization
N = 493
Pre-experiment survey
Control: 20 minute training on "effective googling" and StackOverflow
No access to GPT-4 on assessment
Post-experiment survey
N = 493
Pre-experiment survey
Treatment: 20 minute training on "effective prompting with ChatGPT"
Access to GPT-4 on assessment
Post-experiment survey

Notes: The registration survey is in Appendix Section **??**. The pre-experiment survey and training information can be found in Appendix Section **??**. The text of all three tasks can be found in Appendix Section A.2. The postexperiment survey can be found in Appendix Section **??**.

A.2 Task description

For the purpose of this study, we identify three types of tasks to test nontechnical workers' ability to perform data science-related work. The three tasks are independent of each other and test different skills, including the ability to 1) write Python code to perform data cleaning; 2) identify and correct modeling errors, apply statistical metrics, interpret statistical plots, and calculate probabilities; and 3) build and assess a predictive model. The details of each task are as follows:

Coding task

The coding task is designed to test the ability of a worker to write functioning Python code. Workers are asked to clean, merge, and filter two datasets. This type of task is required in the early stages of any data science exercise. We provide participants with detailed guidelines and tips to help them accomplish the task (see Appendix A.3 for details) to isolate the skill of writing functioning code from that of knowing the appropriate data cleaning steps. Specifically, participants are prompted to write Python code to read two data files that contain information about customers, order numbers, and products. They are then provided clear data quality and cleaning guidelines and tips for handling missing, junk, and duplicate data and asked to identify the five customers with the highest average by cost in a certain period. This exercise involved the following steps:

- Reading the data/creating a data frame
- Splitting combined string features into separate columns
- Handling feature types for filtering
- Detecting junk values
- Removing null and junk values
- Joining data
- Calculating basic metrics

Before beginning the task, we provide participants with written and video instructions on how to download and install Colab, Google's integrated development environment (IDE), to ensure that it is functioning properly (see Section **??** in the Appendix for details). We choose Colab as the IDE because it is relatively simple to set up for nontechnical workers and allows us to access and trace the code for evaluation after task completion.

We divide the task into ten distinct steps, as described in the coding task rubric (Online Appendix Section A). For phase analysis, we combine steps 1-3 as the "data cleaning" phase, steps 4-5 as the "data correction" phase and steps 6-10 as the "data merging and filtering" phase.

Heterogeneous treatment effects on the coding task, the robustness of the results and the details of the phase analysis can be found in Appendix Sections B.2.1, B.2.2 and B.2.3, respectively.

Statistical understanding task

In the statistical knowledge task, we assess participants' ability to understand and apply statistical knowledge. Workers are guided through a series of potential statistical analysis recommendations and other erroneous GenAI outputs including 1) an erroneous analysis plan for creating a model that predicts whether a couple will take out a mortgage based on historical data, 2) analyzing and interpreting residual plots, 3) understanding the application of statistical metrics such as empirical risk to graphical outputs and 4) calculating situational probability. The responses to these questions include a mix of open-ended and multiple-choice answers. The questions are designed to encourage engagement and interpretation rather than simply finding and copying answers. Participants are explicitly instructed not to copy and paste the questions or images into ChatGPT or Google.

The questions are framed around the following three scenarios:

1. Given a data sample that contains financial and demographic information about partners who purchase a home, we ask participants to identify the steps needed to explore and preprocess the data, split the data, train and evaluate the model and then interpret various residual plots.
2. Given a certain distribution of data points, participants are asked how to classify additional data points, identify empirical risks and name potential classifiers that can create certain boundaries.
3. Participants are given a scenario involving a delivery truck that went missing, where the manager has to decide which route to choose based on situational probability.

In addition to the overall average treatment effects, we address the following questions:

- the type of statistical knowledge they test (i.e., debugging a series of steps in a proposed model by ChatGPT, analyzing and interpreting plots, and calculating situational probability)
- whether the responses are open ended or multiple choice, and
- whether the questions are visual or text based.

Section A in the Online Appendix provides the grading rubric for the statistical knowledge task.

Heterogeneous treatment effects on the statistical knowledge task, the robustness of the results and the details of treatment effects based on the type of questions can be found in Sections B.2.1, B.2.2 and B.2.4, respectively.

Prediction task

The prediction task aims to test participants' ability to conduct quantitative analysis and build a predictive model based on historical data. We provide a dataset containing information about soccer matches, including the date, home-team name, away-team name, tournament type, whether the game is at a neutral location, and the number of goals scored by the home and away teams. Participants are asked to quantify the predictability of each match in the dataset. Finally, they are asked to identify the most surprising match based on their predictability metric.

The deliverables of the problem-solving task include the participants' textual description of the methodology they employ to solve the problem and a tabular format of their results for the quantified predictability of each match. Unlike the coding and statistical knowledge tasks, this task is more "open ended", requiring participants to choose their method to build the model and identify an appropriate outcome to measure predictability. The rubric for grading the prediction task is detailed in Section **??** of the Appendix.

Heterogeneous treatment effects on the prediction task, the robustness of the results and the details on the phase analysis can be found in Sections B.2.1, B.2.2 and B.2.5, respectively.

A.3 Task grading methodology

Each task is graded with quantifiable measures of the correctness of the answers and approach, depending on the hypothesis. Each task is graded on both the correctness of the answer and the steps the participant uses to solve the problem. Below, we describe the main outcomes for the correctness of the answer and for the process scores.

Coding task

There is one distinct correct answer for the coding assignment. Correctness is a binary measure where 0 is incorrect and 1 is correct. Second, we compare the output from the workers and data scientists to a rubric we create with 10 steps, where each step is necessary to obtain the correct score. As with the statistics task, we grade this against the rubric (shown in Online Appendix Section A). The rubric scores are weighted correctness scores such that the final score is determined by a weighted sum across all answers as follows:

$$\text{Total correctness} = \sum_{i=1}^{n} (\text{Correctness of answer}_i \times \text{Complexity weight}_i) \tag{A1}$$

where n is the total number of distinct questions and the correctness of the answer, and the complexity weighting is defined as the level of complexity of the question. Similarly, this score is between $(0, 1)$, where 1 means that the worker takes all of the correct steps.

Statistical understanding task

Each question in the statistics task is graded against the rubric (shown in Online Appendix Section A). The rubric scores are weighted correctness scores such that the final score is determined by a weighted sum across all answers as follows:

$$\text{Total correctness} = \sum_{i=1}^{n} (\text{Correctness of answer}_i \times \text{Complexity weight}_i) \tag{A2}$$

where n is the total number of distinct questions and the correctness of the answer, and the complexity weighting is defined as the level of complexity of the question. The complexity weightings are determined by asking several lead data scientists with more than 5 years of experience to rank the complexity of each question and average it across their answers. This correctness measure is bounded by $(0, 1)$, where 1 is a perfect score.

Prediction task

The problem-solving task is designed to have numerous possible answers, some of which are better than others. We use the answers submitted by the data scientists as the baseline/benchmark to grade the workers' results. Specifically, participants submit a predictability score for each match. We normalize participants' predictability scores for each match, score_i, and calculate a loss score for the answers submitted by the workers compared to the data science benchmarks, DS score_i. For each participant, we create a loss score defined as follows:

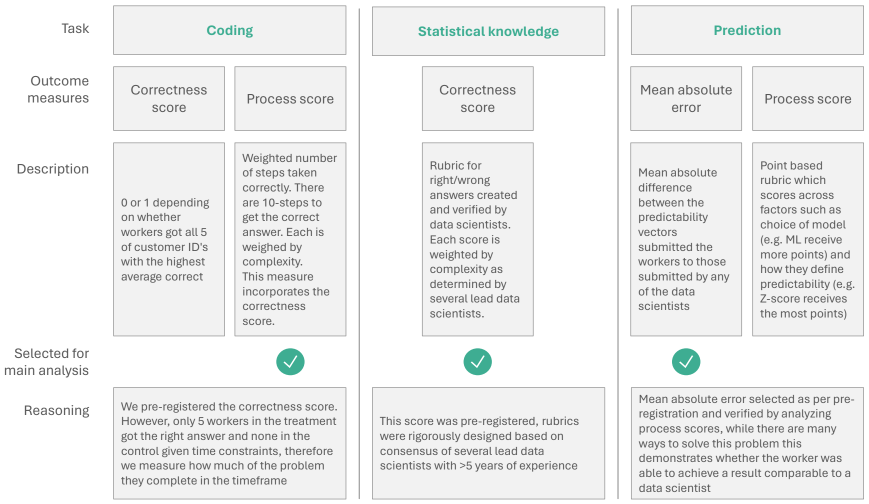
$$\text{Loss Score} = \frac{1}{n}\sum_{i=0}^{n} |\text{score}_i - \text{DS score}_i| \tag{A3}$$

where n is the number of soccer matches in the dataset.

The final score we give workers on the prediction problem is 1 minus the loss score so that the score is between $(0, 1)$, where 1 is a perfect score.

Outcome selection

Fig. A2 Outcome description and selection for the main analysis



Task	Coding		Statistical knowledge	Prediction	
Outcome measures	Correctness score	Process score	Correctness score	Mean absolute error	Process score
Description	0 or 1 depending on whether workers got all 5 of customer ID's with the highest average correct	Weighted number of steps taken correctly. There are 10-steps to get the correct answer. Each is weighed by complexity. This measure incorporates the correctness score.	Rubric for right/wrong answers created and verified by data scientists. Each score is weighted by complexity as determined by several lead data scientists.	Mean absolute difference between the predictability vectors submitted the workers to those submitted by any of the data scientists	Point based rubric which scores across factors such as choice of model (e.g. ML receive more points) and how they define predictability (e.g. Z-score receives the most points)
Selected for main analysis		✓	✓	✓	
Reasoning	We pre-registered the correctness score. However, only 5 workers in the treatment got the right answer and none in the control given time constraints, therefore we measure how much of the problem they complete in the timeframe		This score was pre-registered, rubrics were rigorously designed based on consensus of several lead data scientists with >5 years of experience	Mean absolute error selected as per pre-registration and verified by analyzing process scores, while there are many ways to solve this problem this demonstrates whether the worker was able to achieve a result comparable to a data scientist	

Notes: This figure illustrates the different outcomes for each task, those selected for our main analysis and the reasoning behind the selection.

A.4 Estimation of treatment effects

Across each of these metrics, we employ Equation (4) to estimate the average treatment effects based on ordinary least squares regression, where y_i is the dependent variable (e.g., representing a quantifiable measure of output quality in the coding task and efficiency of code), and T_{GPT} is the ChatGPT treatment dummy. Finally, X_i is a set of covariates collected in the survey—office location, gender, tenure at BCG, and whether the worker is a native English speaker.

$$y_i = \beta_0 + \beta_{\text{GPT}} T_{\text{GPT}} + \gamma X_i + \varepsilon_i \tag{A4}$$

Appendix B Appendix tables and figures

B.1 Balance check

Table B1 Comparison of worker covariates by treatment assignment

Flow from initial allocation into analysis sample	*Total (N)*	*Treatment (N)*	*Control (N)*	*P-value*
Total workers allocated	983	493	493	
↪ began survey	573	298	275	0.33
↪ completed first task	511	270	241	0.19
↪ completed second task	487	260	227	0.13

Pre-allocation attributes of final sample: N = 487	Treatment mean: $\bar{X}_{TRT}$	Control mean: $\bar{X}_{CTL}$	P-value
Female	0.369	0.37	0.985
Bachelors Degree	0.238	0.291	0.192
Masters Degree	0.677	0.604	0.092
Doctorate	0.085	0.106	0.428
Consultant	0.515	0.493	0.361
Low Tenure	0.535	0.542	0.893
Native English	0.471	0.423	0.343
Office in Africa	0.019	0.018	0.896
Office in Asia Pacific	0.135	0.115	0.505
Office in Central or South America	0.019	0.004	0.14
Office in Europe or Middle East	0.492	0.52	0.546
Office in North America	0.335	0.344	0.835
Code at Work	0.277	0.285	0.875
Some Python Familiarity	0.335	0.331	0.922
Familiar with ChatGPT for Coding	0.413	0.381	0.518
Use ChatGPT Daily for Work	0.394	0.438	0.371

Notes: This table reports the means and standard errors of various pretreatment covariates for the treatment and control groups. The first panel describes the flow of the sample from the allocation to the sample we use for our main experimental analysis. The complete allocated sample is described in the first line, with each following line defined cumulatively. Each worker is assigned two tasks, and the following lines compare the number of workers who submit any work for each of the tasks. Those who complete both tasks make up the main experimental sample. The second panel looks at the pre-allocation characteristics of the job seekers in the sample we use for our analysis, N = 487. We report the fraction of workers on their self-reported i) gender, ii) highest degree achieved, and iii) office location. The reported p values are for two-sided t-tests of the null hypothesis of no difference in means across groups.

Table B2 Comparing those who submit something for both tasks (primary analysis sample) to attritors

		Sample Mean	Attritor Mean	P-value
Control	Female	0.38	0.48	0.18
Treatment		0.37	0.43	0.42
Control	Ofice in Europe or Middle East	0.52	0.43	0.19
Treatment		0.50	0.57	0.38
Control	Native English speaker	0.50	0.49	0.86
Treatment		0.41	0.48	0.41
Control	New hire (<1 year)	0.51	0.40	0.11
Treatment		0.54	0.55	0.98
Control	Proficient coder or better	0.31	0.22	0.11
Treatment		0.32	0.20	0.09
Control	Never coded	0.30	0.37	0.35
Treatment		0.31	0.32	0.93
Control	At most 1 coding language	0.93	0.95	0.75
Treatment		0.96	1.00	0.00
Control	PhD	0.11	0.00	0.00
Treatment		0.09	0.05	0.26
Control	Uses ChatGPT daily for work	0.38	0.48	0.16
Treatment		0.45	0.50	0.52
Control	Familiar with prompt engineering	0.68	0.58	0.15
Treatment		0.67	0.68	0.91
Control	Never code for work	0.59	0.68	0.28
Treatment		0.61	0.69	0.42

Notes: This table reports the mean of various pre-experiment covariates amongst the primary analysis sample with those who attrit. The primary analysis sample is made up of workers who submitted anything to be graded for both of their two tasks. The sample here is made up of all workers who started the pre-experiment survey, N = 573. We run a Welch Two Sample t-test on each covariate for attritors and non-attritors, within each treatment group.

B.2 Treatment effects of AI

Table B3 Effects of AI to workers performance on data science tasks, relative to data scientists

	Dependent variable:		
	Coding Task Score	Stats Task Score	Prediction Task Score
	(1)	(2)	(3)
GenAI Treatment Assigned (Trt)	0.490^{***}	0.201^{***}	0.172^{***}
	(0.036)	(0.026)	(0.042)
Mean Y in Control Group	-0.63	-0.32	-0.43
Observations	300	330	298
R^2	0.360	0.151	0.056

Notes: This table analyzes the effect of the treatment on the consultants ability to correctly answer questions. Each outcome is normalized relative to the performance of BCG data scientists. The first outcome is the percentage of correct steps they take in answering the coding question. The second outcome is the sum of the consultant's score on each statistics question, divided by the total number of possible points. The third outcome is the score they got on the prediction problem, which is -1 times their mean absolute error from the correct answer. Exact definition of grading for each problem can be found in Appendix Section A. All standard errors are huber white robust.

Table B4 Effects of AI to whether or not they submit any answer on each task

	Dependent variable:		
	Stats Submitted	Prediction Submitted	Coding Submitted
	(1)	(2)	(3)
GenAI Treatment Assigned (Trt)	0.062^{*}	0.018	0.069^{*}
	(0.033)	(0.035)	(0.040)
Mean Y in Control Group	0.85	0.87	0.78
Observations	369	364	369
R^2	0.024	0.006	0.015

Notes: This table analyzes the effect of the treatment on the consultants submitting any answer to each question. Text of problems can be found in Appendix Section A.

Table B5 Effects of AI to whether or not they get submit any answer on each task

	Dependent variable:	
	Task 1 Submitted	Task 2 Submitted
	(1)	(2)
GenAI Treatment Assigned (Trt)	0.026	0.043
	(0.026)	(0.030)
Mean Y in Control Group	0.88	0.83
Observations	573	573
R^2	0.014	0.016

Notes: This table analyzes the effect of the treatment on the consultants submitting any answer to each question. Text of problems can be found in Appendix Section A.

Table B6 Effects of AI on number of minutes to complete each task

	Dependent variable:		
	Mins on Stats	Mins on Prediction	Mins on Coding
	(1)	(2)	(3)
GenAI Treatment Assigned (Trt)	2.312	-14.450^{***}	-8.884^{***}
	(2.407)	(2.946)	(2.521)
Mean Y in Control Group	63.34	68.48	78.48
Observations	327	318	303
R^2	0.021	0.094	0.058

Notes: This table analyzes the effect of the treatment on the length of time it took for consultants to finish each task, conditional on completion. The outcome in Column (1) is the number of minutes they spent on the Statistics task. The outcome in Column (2) is the number of minutes spent on the Problem Solving and Prediction task. And the outcome in Column (3) is the number of minutes spent on the Coding task. Consultants were randomly assigned two of the three tasks, and given 90 minutes maximum to complete each.

Table B7 Effects of AI treatment to post experiment data science knowledge without use of AI

	Dependent variable:				
	Data science or coding question				
	(1)	(2)	(3)	(4)	(5)
GenAI Treatment Assigned (Trt)	0.001	0.025	0.049^{*}	−0.004	-0.075^{*}
	(0.039)	(0.027)	(0.027)	(0.052)	(0.040)
Mean Y in Control Group	0.35	0.91	0.73	0.80	0.78
Observations	573	399	418	253	408
R^2	0.016	0.017	0.014	0.018	0.050

Notes: This table analyzes the effect of the treatment on the consultants' ability to answer data science and coding questions, after the conclusion of the experiment. Text of questions can be found in the Online Appendix.

Table B8 Effects of AI treatment to post experiment questions about GPT-4's capabilities

	Dependent variable:						
	"Can GPT-4 answer [this question] correctly?"						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
GenAI Treatment Assigned (Trt)	1.424	6.070**	5.574**	6.293***	0.287	9.653***	5.370**
	(2.001)	(2.352)	(2.427)	(2.286)	(2.348)	(2.480)	(2.576)
Mean Y in Control Group	75.82	73.66	74.75	69.08	77.93	67.14	64.93
Observations	454	475	473	464	465	431	451
R^2	0.043	0.023	0.028	0.026	0.005	0.086	0.013

Notes: This table analyzes the effect of the treatment on the consultants' confidence in GPT-4's ability to get the right answer on various questions, after the conclusion of the experiment. For each question, the consultant gave a percentage confidence in GPT-4's ability to answer the question correctly. The question in Columns 1 and 4, and 5 GPT-4 usually get correct. Questions 2,3,6 and7, GPT-4 almost never get correct. Text of questions can be found in the Online Appendix.

Table B9 Performance of workers in treatment and control groups on the 3 tasks compared to data scientists'

	Dependent variable:		
	Coding Task Score	Stats Task Score	Prediction Task Score
	(1)	(2)	(3)
Control Assigned	-0.629***	-0.323***	0.432***
	(0.0809)	(0.0437)	(0.0345)
GenAI Treatment Assigned (Trt)	-0.140*	-0.123***	0.260***
	(0.0833)	(0.0439)	(0.0238)
Mean Y for Data Scientists	0	0	0
Observations	331	357	342
R^2	0.363	0.185	0.152

Notes:
This table illustrates the performance of workers in the control and treatment arms compared to data scientists'. The first outcome is the percentage of correct steps they took on the coding problem, weighted by difficulty. The second outcome is the points the worker got on the statistics question, divided by the maximum points. The third outcome is their mean absolute error of the distance from their answer to the data scientist's answer. Exact definition of grading for each problem can be found in Appendix Section A. All standard errors are Huber White robust SEs.

B.2.1 Heterogeneous treatment effects of AI

Table B10 Effects of AI treatment to post experiment data science knowledge without use of AI

	Dependent variable:				
	Data science or coding question				
	(1)	(2)	(3)	(4)	(5)
GenAI Treatment Assigned (Trt)	−0.040	0.035	−0.011	0.085	0.059
	(0.071)	(0.047)	(0.043)	(0.075)	(0.072)
Assigned Coding Task	−0.042				
	(0.064)				
Assigned Coding Task x Trt	0.040				
	(0.088)				
Assigned Stats Task		0.009			
		(0.045)			
Assigned Stats Task x Trt		−0.012			
		(0.057)			
Assigned Prediction Task			−0.012	0.037	0.142^{**}
			(0.039)	(0.074)	(0.062)
Assigned Prediction Task x Trt			0.090^{*}	−0.152	-0.207^{**}
			(0.055)	(0.101)	(0.086)
Constant	0.434^{***}	0.905^{***}	0.736^{***}	0.780^{***}	0.684^{***}
	(0.052)	(0.037)	(0.030)	(0.059)	(0.054)
Observations	505	399	418	253	408
R^2	0.001	0.003	0.018	0.011	0.025

Notes: This table analyzes the effect of the treatment on the consultants ability to answer data science and coding questions, after the conclusion of the experiment. The first problem is about coding, the second is about statistics, and the final three are about machine learning and prediction. For each problem the treatment is interacted with an indicator for whichever task gave the worker experience in that topic. Text of questions can be found in the Online Appendix. Standard errors are huber white robust.

Table B11 Effects of AI to workers performance on data science tasks, relative to data scientists

	Dependent variable:		
	Coding Task Score	Stats Task Score	Prediction Task Score
	(1)	(2)	(3)
GenAI Treatment Assigned (Trt)	0.547^{***}	0.311^{***}	0.091
	(0.052)	(0.048)	(0.087)
Coding basics	0.071^{*}	0.095^{**}	0.002
	(0.043)	(0.042)	(0.082)
Competent coder	0.151^{***}	0.238^{***}	−0.037
	(0.049)	(0.041)	(0.092)
Coding basics x Trt	−0.131	-0.143^{**}	0.052
	(0.080)	(0.063)	(0.109)
Competent coder x Trt	−0.038	-0.154^{**}	0.175
	(0.085)	(0.062)	(0.112)
Mean Y in Control Group	-0.63	-0.32	-0.43
Observations	300	330	298
R^2	0.387	0.243	0.070

Notes: This table analyzes the effect of the treatment on the consultants ability to correctly answer questions, by their pre-experiment coding knowledge. The omitted variable is "No prior coding experience." Workers who report having basic coding ability are classified as "Coding basics" and those who report knowing how to code are classifed as "competent coders." Each outcome is normalized relative to the performance of BCG data scientists. The first outcome is the percentage of correct steps they take in answering the coding question. The second outcome is the sum of the consultant's score on each statistics question, divided by the total number of possible points. The third outcome is the score they got on the prediction problem, which is 1- their mean absolute error. Exact definition of grading for each problem can be found in Appendix Section A. All standard errors are huber white robust.

Table B12 Effects of AI to workers performance on data science tasks, relative to data scientists

	Dependent variable:		
	Coding Task Score	Stats Task Score	Prediction Task Score
	(1)	(2)	(3)
GenAI Treatment Assigned (Trt)	0.503***	0.183***	0.128***
	(0.041)	(0.030)	(0.047)
Knowledge of GPT's strengths	0.017	0.003	−0.128
	(0.052)	(0.051)	(0.085)
Knowledge of GPT's strengths x Trt	−0.062	0.080	0.195**
	(0.088)	(0.063)	(0.098)
Mean Y in Control Group	-0.63	-0.32	-0.43
Observations	300	330	298
R^2	0.361	0.160	0.069

Notes: This table analyzes the effect of the treatment on the consultants ability to correctly answer questions, by their pre-experiment coding knowledge. "Knowledge of GPT's strengths" is 1 if the consultant got 4 out of 7, or more of questions correct on the pre-experiment survey asking about their guesses of whether or not GPT-4 can correctly answer a question. The omitted variable is consultants who got fewer tha 4 out of 7 correct. Each outcome is normalized relative to the performance of BCG data scientists. The first outcome is the percentage of correct steps they take in answering the coding question. The second outcome is the sum of the consultant's score on each statistics question, divided by the total number of possible points. The third outcome is the score they got on the prediction problem, which is -1 times their mean absolute error. Exact definition of grading for each problem can be found in Appendix Section A. All standard errors are huber white robust.

Table B13 Effects of AI to workers performance on data science tasks, relative to data scientists

	Dependent variable:		
	Coding Task Score	Stats Task Score	Prediction Task Score
	(1)	(2)	(3)
GenAI Treatment Assigned (Trt)	0.512^{***}	0.250^{***}	0.214^{***}
	(0.048)	(0.038)	(0.061)
Overconfident	0.026	0.041	0.088
	(0.042)	(0.037)	(0.068)
Overconfident x Trt	−0.046	-0.094^{*}	−0.087
	(0.072)	(0.052)	(0.083)
Mean Y in Control Group	-0.63	-0.32	-0.43
Observations	300	330	298
R^2	0.361	0.159	0.063

Notes: This table analyzes the effect of the treatment on the consultants ability to correctly answer questions, by their pre-experiment estimates of GPT-4's current capabilities. "Overconfident" is 1 if the consultant believed that GPT-4 could correctly answer at least three of the five questions which GPT-4 could not do. Each outcome is normalized relative to the performance of BCG data scientists. The first outcome is the percentage of correct steps they take in answering the coding question. The second outcome is the sum of the consultant's score on each statistics question, divided by the total number of possible points. The third outcome is the score they got on the prediction problem, which is -1 times their mean absolute error. Exact definition of grading for each problem can be found in Appendix Section A. All standard errors are huber white robust.

Table B14 Effects of AI to workers performance on data science tasks, relative to data scientists

	Dependent variable:		
	Coding Task Score	Stats Task Score	Prediction Task Score
	(1)	(2)	(3)
GenAI Treatment Assigned (Trt)	0.468^{***}	0.184^{***}	0.193^{***}
	(0.041)	(0.030)	(0.049)
Quant Degree	0.015	0.019	0.061
	(0.052)	(0.042)	(0.075)
Quant Degree x Trt	0.090	0.089	−0.070
	(0.081)	(0.057)	(0.093)
Mean Y in Control Group	-0.63	-0.32	-0.43
Observations	300	330	298
R^2	0.367	0.167	0.059

Notes: This table analyzes the effect of the treatment on the consultants ability to correctly answer questions, by their pre-experiment estimates of GPT-4's current capabilities. "Quant Degree" is 1 if the worker reports having any formal degree in Statistics, Economics, Mathematics or Data Science. Each outcome is normalized relative to the performance of BCG data scientists. The first outcome is the percentage of correct steps they take in answering the coding question. The second outcome is the sum of the consultant's score on each statistics question, divided by the total number of possible points. The third outcome is the score they got on the prediction problem, which is -1 times their mean absolute error. Exact definition of grading for each problem can be found in Appendix Section A. All standard errors are huber white robust.

Table B15 Lee Bounds on Treatment Effects for Main Results

	Treatment effect	**Lee Lower Bound**	**Lee Upper Bound**
Statistics Task Score	0.200***	0.194**	0.201***
	(0.026)	(0.043)	(0.043)
Prediction Task Score	0.182***	0.157***	0.201***
	(0.041)	(0.046)	(0.070)
Coding Task Score	0.491***	0.471***	0.556***
	(0.038)	(0.052)	(0.054)

Notes: This table analyzes the effect of the treatment on the consultants ability to correctly answer questions. The first outcome is the sum of the consultant's score on each statistics question, divided by the total number of possible points. The second outcome is the percentage of correct steps they take in answering the coding question. The third outcome is the score they got on the prediction problem, which is their mean absolute error multiplied by -1 for ease of interpretation. Exact definition of grading can be found in Appendix Section A. All regressions include no controls.

B.2.2 Robustness of the results

Table B16 Effects of AI to workers performance on coding task, relative to data scientists

	Dependent variable:				
	Coding Score				
	(1)	(2)	(3)	(4)	(5)
GenAI Treatment Assigned (Trt)	0.490***	0.491***	0.487***	0.559***	0.564***
	(0.038)	(0.038)	(0.038)	(0.069)	(0.069)
Coding basics				0.102	0.074
				(0.071)	(0.072)
Competent coder				0.179**	0.129
				(0.072)	(0.080)
Europe		0.053	0.072*	0.075*	0.088**
		(0.041)	(0.041)	(0.041)	(0.042)
Female		0.027	0.041	0.052	0.042
		(0.040)	(0.041)	(0.040)	(0.041)
Low tenure		−0.042	−0.063	−0.056	−0.080*
		(0.039)	(0.040)	(0.038)	(0.041)
Native English speaker		0.051	0.023	0.055	0.025
		(0.041)	(0.042)	(0.041)	(0.042)
Masters degree			−0.072		−0.074
			(0.046)		(0.046)
PhD			−0.125*		−0.117
			(0.072)		(0.072)
Degree in quantitative field			0.048		0.040
			(0.045)		(0.045)
STEM			0.063		0.045
			(0.040)		(0.043)
Minimal experience with prediction			−0.102**		−0.106**
			(0.051)		(0.051)
Some experience with prediction			0.046		0.0002
			(0.050)		(0.056)
Data visualization experience			0.034		0.023
			(0.049)		(0.049)
Coding basics x Trt				−0.162*	−0.160*
				(0.094)	(0.095)
Competent coder x Trt				−0.042	−0.056
				(0.094)	(0.094)
Constant	−0.629***	−0.668***	−0.653***	−0.781***	−0.690***
	(0.028)	(0.050)	(0.082)	(0.071)	(0.093)
Observations	300	300	300	300	300
R^2	0.360	0.369	0.409	0.403	0.425

Notes: This table analyzes the effect of the treatment on the worker's score on the coding task relative to the data scientists benchmark, where 0 is a perfect score. The first specification is the effect of treatment on their score. The second specification adds controls for gender, office location, native english status, and prior coding ability. The third specification is the same as column (2) but with additional controls for different measures of the workers prior technical experience. The fourth specification is the same as column (2), where the treatment is interacted with the workers prior coding experience. The fifth specification is the same as column (3), where the treatment is interacted with the workers prior coding experience.

Table B17 Effects of AI to workers performance on statistics task, relative to data scientists

	Dependent variable:				
	Statistics Score				
	(1)	(2)	(3)	(4)	(5)
GenAI Treatment Assigned (Trt)	0.201^{***}	0.201^{***}	0.206^{***}	0.308^{***}	0.309^{***}
	(0.026)	(0.026)	(0.026)	(0.045)	(0.045)
Coding basics				0.094^{**}	0.088^{*}
				(0.047)	(0.048)
Competent coder				0.234^{***}	0.210^{***}
				(0.045)	(0.050)
Europe		0.028	0.027	0.035	0.030
		(0.029)	(0.029)	(0.028)	(0.029)
Female		−0.033	−0.010	−0.014	−0.011
		(0.027)	(0.028)	(0.026)	(0.027)
Low tenure		0.036	0.051^{*}	0.032	0.042
		(0.027)	(0.028)	(0.026)	(0.028)
Native English speaker		0.038	0.033	0.038	0.027
		(0.029)	(0.030)	(0.028)	(0.030)
Masters degree			−0.027		−0.033
			(0.032)		(0.032)
PhD			0.004		−0.015
			(0.052)		(0.051)
Degree in quantitative field			0.041		0.032
			(0.032)		(0.031)
STEM			0.0003		−0.024
			(0.028)		(0.029)
Minimal experience with prediction			0.048		0.036
			(0.036)		(0.035)
Some experience with prediction			0.121^{***}		0.066^{*}
			(0.034)		(0.037)
Data visualization experience			0.044		0.020
			(0.035)		(0.035)
Coding basics x Trt				-0.137^{**}	-0.130^{**}
				(0.063)	(0.063)
Competent coder x Trt				-0.147^{**}	-0.153^{**}
				(0.061)	(0.062)
Constant	-0.323^{***}	-0.362^{***}	-0.461^{***}	-0.491^{***}	-0.499^{***}
	(0.019)	(0.036)	(0.061)	(0.046)	(0.065)
Observations	330	330	330	330	330
R^2	0.151	0.165	0.224	0.254	0.273

Notes: This table analyzes the effect of the treatment on the worker's score on the statistcis task relative to the data scientists benchmark, where 0 is a perfect score. The first specification is the effect of treatment on their score. The second specification adds controls for gender, office location, native english status, and prior coding ability. The third specification is the same as column (2) but with additional controls for different measures of the workers prior technical experience. The fourth specification is the same as column (2), where the treatment is interacted with the workers prior coding experience. The fifth specification is the same as column (3), where the treatment is interacted with the workers prior coding experience.

Table B18 Effects of AI to workers performance on a prediction task, relative to data scientists

	Dependent variable:				
	Mean Absolute Error				
	(1)	(2)	(3)	(4)	(5)
GenAI Treatment Assigned (Trt)	0.172***	0.182***	0.168***	0.103	0.086
	(0.041)	(0.041)	(0.042)	(0.078)	(0.078)
Coding basics				0.014	0.043
				(0.073)	(0.076)
Competent coder				−0.026	0.023
				(0.080)	(0.096)
Europe		0.020	0.016	0.025	0.019
		(0.046)	(0.046)	(0.046)	(0.047)
Female		0.023	0.011	0.028	0.008
		(0.043)	(0.044)	(0.043)	(0.044)
Low tenure		−0.008	−0.026	−0.011	−0.028
		(0.042)	(0.044)	(0.042)	(0.044)
Native English speaker		0.084*	0.101**	0.088*	0.104**
		(0.047)	(0.049)	(0.047)	(0.049)
Masters degree			0.088*		0.101**
			(0.050)		(0.050)
PhD			−0.084		−0.076
			(0.089)		(0.089)
Degree in quantitative field			0.002		0.012
			(0.048)		(0.049)
STEM			−0.016		−0.046
			(0.045)		(0.048)
Minimal experience with prediction			−0.099*		−0.094*
			(0.054)		(0.054)
Some experience with prediction			−0.037		−0.079
			(0.057)		(0.065)
Data visualization experience			0.050		0.030
			(0.052)		(0.053)
Coding basics x Trt				0.051	0.058
				(0.101)	(0.101)
Competent coder x Trt				0.173	0.176*
				(0.107)	(0.107)
Constant	−0.432***	−0.491***	−0.516***	−0.494***	−0.515***
	(0.030)	(0.055)	(0.087)	(0.077)	(0.098)
Observations	298	298	298	298	298
R^2	0.056	0.068	0.101	0.083	0.120

Notes: This table analyzes the effect of the treatment on the worker's score on the prediction problem, -1 times their mean absolute error from the correct answer, where 0 is a perfect score. The first specification is the effect of treatment on their score. The second specification adds controls for gender, office location, native english status, and prior coding ability. The third specification is the same as column (2) but with additional controls for different measures of the workers prior technical experience. The fourth specification is the same as column (2), where the treatment is interacted with the workers prior coding experience. The fifth specification is the same as column (3), where the treatment is interacted with the workers prior coding experience.

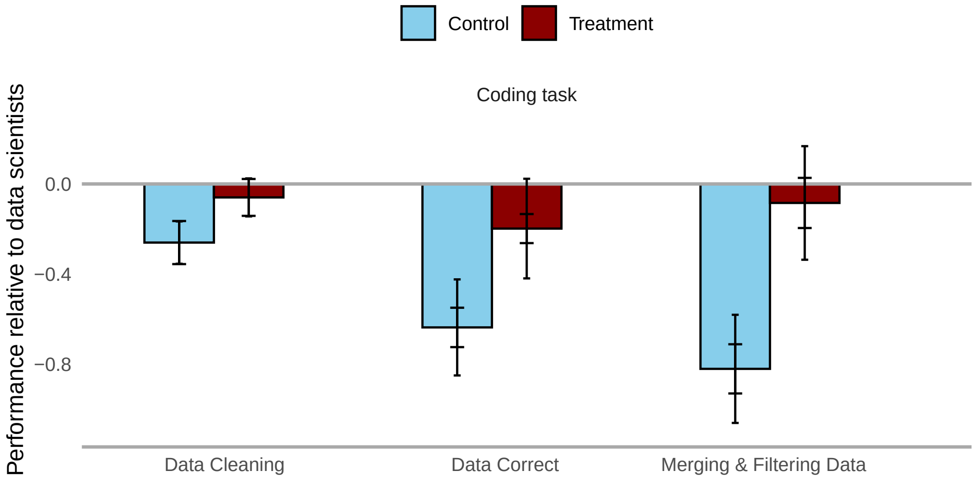
B.2.3 Treatment effects on the coding task – Details

Table B19 Effect of AI treatment on workers' coding performance, relative to data scientists

	Dependent variable:		
	Data Cleaning	Data Correct	Merge & Filter
	(1)	(2)	(3)
GenAI Treatment Assigned (Trt)	0.204^{***}	0.440^{***}	0.738^{***}
	(0.032)	(0.059)	(0.061)
Mean Y in Control Group	-0.26	-0.64	-0.82
Observations	300	300	300
R^2	0.129	0.172	0.343

Notes: This table reports the effect of the treatment on the consultants performance on the coding task. The y-axis is the worker's binary grade on each set of actions in python, where 0 is a perfect score and -1 is the lowest possible score. The first outcome is whether or not the worker correctly cleaned the data. The second outcome is whether or not they correctly handle nulls and duplicates. The third outcome is whether or not they correctly merge and then filter a dataset. The sample includes all experimental participants who submitted something for grading on the prediction task. Text of problems can be found in Appendix Section efsec:tasks. Exact definition of grading for each problem can be found in Appendix Section A. All regressions include controls for gender, location, native english status, and low tenure.

Fig. B3 Effect of AI treatment on workers' coding performance relative to that of data scientists



Notes: This plot reports the effect of the treatment on workers' performance on the coding task. The y-axis is the worker's binary grade on each set of actions in Python, where 0 is a perfect score and -1 is the lowest possible score. The first outcome is whether or not the worker correctly cleans the data. The second outcome is whether or not the worker correctly handles nulls and duplicates. The third outcome is whether or not the worker correctly merges and then filters a dataset. A 95% CI is plotted around the treatment group's mean as compared to that of the control group. The benchmark set by the data scientists is also plotted. All regressions include controls for gender, location, native English speaker status, and low tenure. The sample includes all experimental participants who submit a portion of each coding task for grading. The text of the problems can be found in Appendix Section A.2. Regression details can be found in Appendix Table B19.

B.2.4 Treatment effects on the statistics task – Details

Table B20 Effect of AI treatment on workers' statistics performance, relative to data scientists

	Dependent variable:		
	Identifying Errors	Visual ML Questions	Probability Questions
	(1)	(2)	(3)
GenAI Treatment Assigned (Trt)	0.389^{***}	0.187^{***}	0.068^{**}
	(0.080)	(0.029)	(0.032)
Mean Y in Control Group	-0.42	-0.38	-0.22
Observations	333	329	329
R^2	0.108	0.127	0.035

Notes: This table reports the effect of the treatment on the consultants grades on each step of the statistics task. Each outcome is the worker's grade on each set of problems, where 0 is a perfect score and -1 is the lowest possible score. The first outcome is the workers' score on a problem where they must identify errors in statistical reasoning. The second outcome is their score on questions where they look at data plotted and decide what types of classifier to use on it. The third outcome is their score on probability questions. The sample includes all experimental participants who submitted something for grading on the statistics task. Text of problems can be found in Appendix Section efsec:tasks. Exact definition of grading for each problem can be found in Appendix Section A. All regressions include controls for gender, location, native english status, and low tenure.

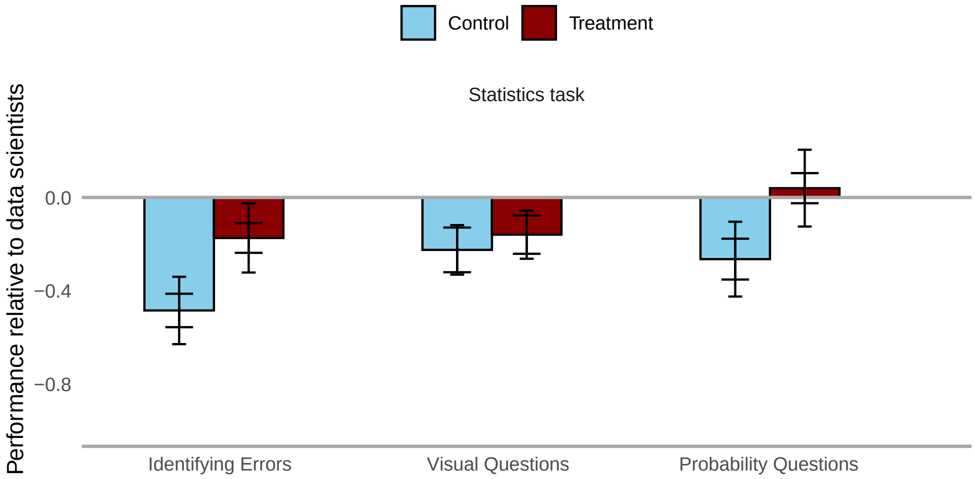
B.2.5 Treatment effects on the prediction task – Details

Table B21 Effect of AI treatment on workers' prediction scores, relative to data scientists

	Dependent variable:	
	Mean Absolute Error	Process Score
	(1)	(2)
GenAI Treatment Assigned (Trt)	0.182^{***}	0.127^{***}
	(0.041)	(0.045)
Mean Y in Control Group	-0.43	-0.38
Observations	298	281
R^2	0.068	0.036

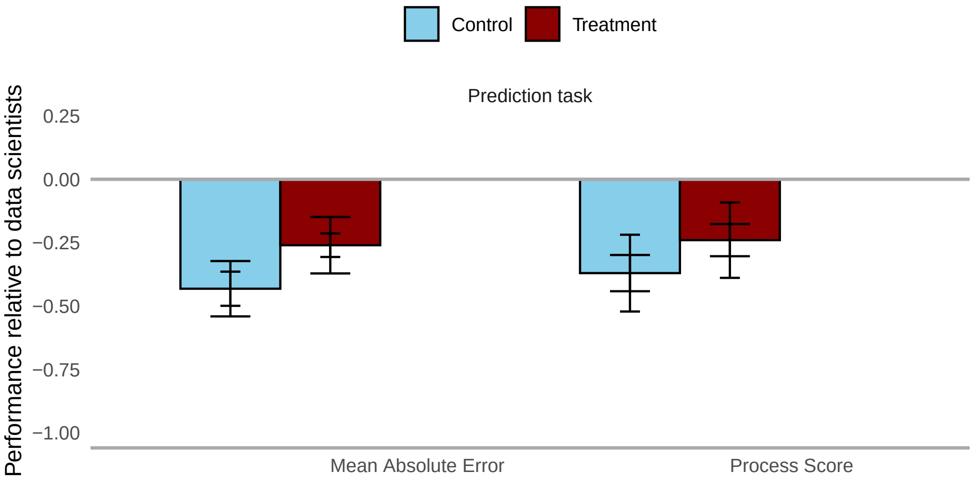
Notes: This table reports the effect of the treatment on the consultants performance on the prediction task. Each outcome is the worker's grade on each set of problems, where 0 is a perfect score and -1 is the lowest possible score. The first outcome is the score they got on the prediction problem, which is -1 times their mean absolute error. The second outcome is the percentage of "correct" steps they took. The sample includes all experimental participants who submitted something for grading on the prediction task. Text of problems can be found in Appendix Section efsec:tasks. Exact definition of grading for each problem can be found in Appendix Section A. All regressions include controls for gender, location, native english status, and low tenure.

Fig. B4 Effect of AI treatment on workers' statistical performance relative to that of data scientists



Notes: This plot reports the effect of the treatment on workers' grades on each step of the statistics task. The y-axis is the workers' grade on each set of problems, where 0 is a perfect score and -1 is the lowest possible score. The first outcome is the workers' score on a problem where they must identify errors in statistical reasoning. The second outcome is their score on questions where they look at data plotted and decide which types of classifier to use on it. The third outcome is their score on probability questions. A 95% CI is plotted around the treatment group's mean as compared to that of the control group. The benchmark set by the data scientists is also plotted. All regressions include controls for gender, location, native English speaker status, and low tenure. The sample includes all experimental participants who submit a portion of the statistics task for grading. The text of the problems can be found in Appendix Section A.2. Regression details can be found in Appendix Table B20.

Fig. B5 Effect of AI treatment on workers' prediction performance relative to that of data scientists



Notes: This plot reports the effect of the treatment on workers' performance on the prediction task. The y-axis is the worker's grade on each set of problems, where 0 is a perfect score and -1 is the lowest possible score. The first outcome is the score that the worker receives on the prediction problem, which is -1 times his or her mean absolute error. The second outcome is the percentage of "correct" steps the worker takes. A 95% CI is plotted around the treatment group's mean as compared to that of the control group. The benchmark set by the data scientists is also plotted. All regressions include controls for gender, location, native English speaker status, and low tenure. The sample includes all experimental participants who submit a portion of the statistics task for grading. The text of the problems can be found in Appendix Section A.2. Regression details can be found in Appendix Table B21.