The impact of generative AI on productivity in central banking: Evidence from the National Bank of Slovakia*

Aleš Maršál
National Bank of Slovakia

Patryk Perkowski
Yeshiva University

September 13, 2024

Preliminary and incomplete. Please do not cite or distribute without author's permission.

Abstract

We report the results of a field experiment at the National Bank of Slovakia that examined the productivity impacts of generative AI in central banking. We randomized employees of the central bank to complete workplace tasks with and without the use of generative AI. We find that access to generative AI leads to large productivity gains, both in terms of quality of work and also efficiency measured in time. We find that the positive productivity benefits of generative AI are found throughout the distribution of participants. Moreover, generative AI has positive productivity impacts not only for the majority of generalist tasks, but also specialist tasks which require subject-matter expertise. We conclude by discussing the implications of our findings for the use of generative AI in central banking, and skill-biased technical change.

Keywords: Generative AI, Large Language Models, Central Banking, Field Experiment, Worker Productivity
JEL Classification: J24, M15, E58, O33

*Aleš Maršál (ales.marsal@nbs.sk); Patryk Perkowski (patryk.perkowski@yu.edu). We are grateful to Pavel Gertler and Jan Klacso for their support in executing this field experiment. The views expressed herein are those of the author(s) and do not necessarily represent the views of the National Bank of Slovakia. IRB approval for this study was obtained from WCG IRB under protocol number 20241892. The experiment was pre-registered at https://osf.io/23jr6/.

1 Introduction

In this paper, we investigate the impacts of generative AI on productivity in central banking through a field experiment with the National Bank of Slovakia (NBS). In the spring of 2024, we recruited employees at NBS to participate in a study on technology use at the bank. The goal of the experiment was to understand the productivity impacts of generative AI in central banking, both across the distribution of workplace tasks and also the distribution of employees. To that end, we randomized the presence of ChatGPT while participants completed various workplace tasks.

We designed the experiment to estimate the impact of generative AI on productivity in not only generalist job tasks, but also specialist ones. In terms of the former, we used the task classification system from Autor (2013) and designed a series of seven types of general workplace tasks that are common to modern-day work. These task types included: (i) routine cognitive; (ii) non-routine cognitive; (iii) routine manual; (iv) nonroutine manual; (v) analytical; (vi) leadership; and (vii) interpersonal and communication tasks. In terms of the latter, we designed highly-specialized tasks that required subject-matter expertise. For example, one of the tasks for participants in monetary policy and research roles was to create a monetary policy report for a fictitious country. Our experimental design also allows us to estimate the distributional impacts of generative AI across the employee population. Each participant in our sample completed 16 tasks (=14 generalist tasks and two specialist ones), allowing us to estimate the within-person treatment effect of generative AI.

We collected two main types of productivity outcomes. The first outcome is the quality of responses submitted. A team of three evaluators coded up their evaluations of each of the task responses on a scale of zero to ten. We use the average of these three evaluations to measure the quality of work submitted. The second outcome is the amount of time taken to complete the various tasks, which we measure in minutes. We use both of these measures to understand how generative AI impacts productivity.

Our results indicate that generative AI led to large improvements in both our productivity measures. In terms of the quality of responses on the generalist tasks, we find an average quality improvement of between 33 and 44 percent depending on the econometric specification. In terms of time to complete the task, participants took on average 21 percent less time to complete the tasks when they had access to generative AI. The effect sizes (relative to the control mean) were about twice as large for

improvements in quality versus improvements in time.

Moreover, we find that the productivity benefits of generative AI are widespread across task types. Generative AI led to a large increase in productivity for six out of the seven generalist task types. The positive impact is not only limited to non-routine cognitive tasks, where generative AI increased performance by 76 percent for routine cognitive tasks and 51 percent for non-routine cognitive tasks. We also see large increases in performance for non-routine manual tasks, analytical tasks, leadership tasks, and communication tasks. The one exception is for routine manual tasks, where participants saw a decrease in performance when using generative AI.

Our results also indicate large productivity benefits throughout the distribution of participants. 94 percent of participants scored higher on our measure of quality when they had access to generative AI versus when they did not.

In our study, we find that generative AI has an "equalizing" effect on the performance distribution of subjects. Consistent with findings from previous studies, such as Dell'Acqua et al. (2023), generative AI significantly enhances the performance of lower-performing individuals, enabling them to close the gap with top performers. While high performers also improve with the use of generative AI, their relative gains are more modest. However, when it comes to time efficiency, the results deviate from this pattern. Although all participants experience time savings, the extent of the benefits varies significantly, leading to a wider dispersion in time saved across individuals. Hence, while generative AI may level the playing field in terms of output quality, it simultaneously amplifies disparities in task completion speed. To our knowledge, this is the first study to examine the distributional effects of generative AI on time efficiency, marking a novel contribution to the literature.

2 Experiment

2.1 Setting

Our field experiment was held at the National Bank of Slovakia, the country's central bank.

2.2 Sample

Our sample consists of 97 employees throughout the organizational hierarchy at NBS.

Table 1 contains summary statistics for the sample. The sample is highly educated– 92 percent have a master's degree or higher. The average age is 40 years old, and 35 percent of participants are female. Participants have been employees at NBS for on average 8 years, and 5 percent manage or lead a team. The average participant spends 31 percent of their time working on analytical tasks, 26 percent on non-routine cognitive tasks, 18 percent of their time on routine cognitive tasks, and 16 percent on leadership and management tasks.

Overall, the sample exhibits awareness of generative AI, though workplace rates of usage are lower. 95 percent have heard of generative AI before and 75 percent have used generative AI before, but only around half have used generative AI at work before.

2.3 Experimental design

The experiment proceeded in two stages, which we display in Figure 1.

2.3.1 Tasks

The first stage of the experiment focused on understanding the impact of generative AI on performance in general workplace tasks. We used the task classification system from Autor (2013) and designed a series of seven types of general workplace tasks that are common to modern-day work. These task types included: (i) routine cognitive; (ii) non-routine cognitive; (iii) routine manual; (iv) nonroutine manual; (v) analytical; (vi) leadership; and (vii) interpersonal and communication tasks. For each of these general task types, we created two specific tasks to allow us to vary the use of generative AI. For example, each participant would complete one of the general analytical tasks with the use of generative AI, and complete the other one without the use of generative AI. This led to a total of 14 tasks.

Our goal with the design of these general workplace tasks was two-fold. First, we wanted to have general tasks so they are applicable to large parts of the organization, and white collar work more generally. Second, as specified in our pre-analysis plan, we were interested in understanding the impact of generative AI depending on task type. To that end, we designed these tasks to be broadly

representative, to correspond to the respective categories in the Autor (2013) task taxonomy, and to be similar to work done at the organization.

In Appendix A.1, we list out the exact general tasks that participants had to complete. For example, one of the routine cognitive tasks asked participants to proofread and edit a paragraph for grammatical and spelling mistakes. One of the analytical tasks asked participants to calculate the percentage of the workforce that would be subject to a hypothetical new remote work policy. Given the experiment was run using laptops, we were unable to have participants do purely manual tasks. However, we created versions of manual tasks that were able to be completed online. For example, one of the non-routine manual tasks asked participants to walk through the steps they would take in order to fix a broken printer in the office. Overall, participants had one hour to complete all 14 general workplace tasks, such that each task averaged about four minutes.

We additionally included a second stage of the experiment that focused on more specialized tasks. In early work, we document that employees are less likely to use generative AI for more specialized tasks versus more general workplace tasks (Perkowski and Maršál, 2024). This may reflect the inability of these technologies to help with workplace tasks that require more skills, or it may reflect participant aversion to using generative AI for more advanced tasks. Indeed, informal conversations with participants revealed that even among generative AI users at work, the second stage of the experiment was the first time they used the technology for more specialized job tasks. Our interest here was to examine the productivity impacts of generative AI in more specialized tasks.

One difficulty in studying the impacts of generative AI on specialized workplace tasks is the heterogeneous positions occupied by participants in our sample. As Table 1 indicates, our participants were spread across numerous divisions at the bank. For feasibility purposes, we used the organizational structure of NBS to assign employees to respective blocks given their division at the bank. This led to a total of five blocks of employees, who perform similar tasks at the organization. These blocks were: (i) supervision (which includes supervision and financial stability, and supervision and financial consumer protection); (ii) monetary policy and research (which includes monetary policy and market operations, and research and statistics); (iii) financial markets (which includes payment systems and cash management, and risk management and settlement; (iv) information technology (which includes information technology, financial management, and operation activities), and (v) other (which included strategy and development, the Office of the Governor, human resources, internal audit, legal services, and the NBS foundation).

We then worked alongside industry professionals to design highly-specialized tasks for each of these blocks.

The specialized tasks were designed to require high level subject-matter expertise. For example, one of the tasks for participants in the monetary policy and research block was to create a monetary policy report for a fictitious country. Participants were given a dataset and asked to apply the Hodrick-Prescott (HP filter) to decompose GDP into cyclical and trend components, fit an ARIMA model to the HP trend component, forecast the trend component beyond the sample period, and develop a strategic plan for monetary policy adjustments over the next year.[1] In Appendix A.2, we list out the specialized tasks that participants had to complete during the second stage of the experiment, depending on their block. Participants completed two specialized tasks during this stage: one with the use of generative AI and one without. They had 30 minutes for each of the specialized tasks.

Following the experiment, participants completed a survey questionnaire describing their experience with the task and beliefs regarding generative AI. Overall, the experiment took around two hours. Participants were paid a base rate of $X, with a performance bonus.

2.3.2 Random assignment

Our experiment used a within-subjects design, while also permitting some randomization between-subjects. While all participants completed 14 tasks, we randomly allowed the use of generative AI for a subsample of these tasks. We then randomized access to generative AI across task categories within each subject. Some participants were eligible to use generative AI for the first set of the seven general tasks and one specialized workplace task, while others were eligible to use it for the second set and other specialized task. This design ensured that participants completed each of the task types from Autor (2013) both with and without generative AI.

Additionally, we collected each participant's chat transcripts with the large language model. We can use this data to examine the extent of uptake among participants both across all task types, and by specific task categories.

[1] The one exception to this is the specialized tasks for those in the other block. Given the heterogeneity in this group, we used general workplace tasks for this group.

2.4 Data

Our data come from three sources. First, we use the data collected from the tasks during the actual experiment itself. This includes the submission for the tasks plus the amount of time it took to complete the tasks. Second, we have detailed chat transcripts for each participant on their use of generative AI. This was collected directly through ChatGPT. Lastly, for each participant, we have detailed survey data collected over a month prior to the experiment itself. This includes exposure to generative AI, workplace characteristics, and beliefs regarding generative AI. [2]

Our data is at the participant-task level, and two key outcomes capture performance in our setting. The first captures a subjective evaluation of task performance by evaluators. For each of the 14 tasks, we devised a zero to ten rubric to measure performance using the rubrics in Appendix sections A.1 and A.2. Each general task response was evaluated by three individuals (a PhD student, a post-doctoral fellow, and an employee at the bank), and we took the average of the evaluations across all three evaluators. Meanwhile industry professionals evaluated the responses to the specialized task. All evaluators were blind to treatment assignment. Our first performance measure is this evaluation of task performance.

The second is an efficiency measure, which captures how long it took participants to complete each task. We use timestamps from submitting the tasks to generate the time it took participants to complete each task. Given that the general tasks were given as bundles of seven tasks, we cannot estimate efficiency for each general task. Instead, we have the amount of time it took to complete all seven tasks.

2.5 Empirical strategy

Our goal is to understand the impact of generative AI on productivity. Our experimental design enables us to estimate treatment effects at various levels.

First, we can estimate the impact of generative AI on productivity across all participants and tasks using the following equation:

$$y_{p,t} = \beta_0 + \beta_1 * GAI_{p,t} + \theta * P_p + \epsilon_{p,t} \tag{1}$$

[2] See Perkowski and Maršál (2024) for additional information on this survey.

In this equation, p indexes participants and t indexes tasks. $y_{p,t}$ is a performance measure for participant p on task m. $GAI_{p,t}$ is a binary indicator that equals one if participant p had access to generative AI for task t, and zero otherwise. $P_p$is a matrix of participant-level fixed effects and $\epsilon_{p,t}$ is the error term. We estimate equation 1 using robust standard errors clustered at the participant level. The coefficient of interest is β_1, which captures the average impact of generative AI on performance across all participants and tasks.

The benefit of equation 1 is that it allows us to estimate the impact of generative AI at the participant-level. Given that we have 14 observations per participant, half of which are with generative AI and half without, we can estimate the productivity impact of generative AI for each participant. We can thus plot the distribution of treatment effects to see whether some participants benefit more versus less with the use of generative AI. Given that we only have two observations per participant-task type (for example, each participant only completed two of analytical tasks), we are unable to estimate treatment effects at the participant-task-type level. Instead, equation 1 allows us look at the average and distributional impacts of generative AI across our participants and across all talks.

In order to examine how the impact of generative AI varies depending on the task type, we estimate the following equation:

$$y_{p,t} = \beta_0 + \beta_1 * GAI_{p,t} + \delta * X_p + \lambda * T_t + \epsilon_{p,t} \tag{2}$$

This equation is similar to equation 1 except for two changes. First, we replace the matrix of participant fixed effects (P_p) with a vector of participant-level controls (X_p). These controls include participant demographics, previous exposure to generative AI, beliefs in the productivity impact, among others. Second, we add a vector of task fixed effects (T_t), which contains binary indicators for each of the eight types of tasks in the study. We can then estimate interactions between $GAI_{p,t}$ and T_t to understand whether the impact of generative AI on analytical tasks is different than its impact on non-routine cognitive tasks, for example. We can also use this equation to estimate heterogeneous effects by participant attributes. We do so by interacting $GAI_{p,t}$ and participant attributes in X_m. In accordance with our pre-analysis plan, we are especially interested in participant heterogeneous treatment effects by gender, tenure, the percent of workload in English, and previous exposure to generative AI.

3 Results

We break out the results section into two primary subsections. In our first subsection, we examine the impact of generative AI on productivity in generalist central banking tasks through three types of analysis. We examine the overall impact of generative AI on performance in these tasks. We then estimate the impact of generative AI depending on the type of generalist task. Lastly, we examine the distributional impacts of generative AI and heterogeneous treatment effects depending on participants. In the second subsection, we focus on the impact of generative AI on productivity in specialist central banking tasks.[3]

3.1 Impact of generative AI on productivity in generalist central banking tasks

3.1.1 Overall impact

We begin by examining the overall impact of generative AI on productivity in generalist central banking tasks. In Figure 2, we plot the distribution of our two productivity measures for the generalist tasks by treatment condition. In Panel A, we display the total score across all seven tasks in a given round, while Panel B displays the total number of minutes taken to complete all seven tasks.

The results in Figure 2 indicate that generative AI led to large average improvements in both our productivity measures. In terms of the aggregate score on the task, participants scored on average 60 points when they had access to generative AI versus 41 points when they did not have access to generative AI. In terms of the time spent on the task, those who had access to generative AI spent on average 23 minutes to complete all seven tasks, versus 30 minutes for those who did not have access to generative AI. Thus, the graphical evidence in Figure 2 provide suggestive evidence that generative AI led to improvements in quality and time of completion,

Table 2 more formally examines the overall impact of generative AI on productivity using the estimation strategy detailed in section 2.5. Columns 1 and 2 use the score on the task as our measure of productivity, while columns 3 and 4 use the time spent on the task as our measure of task performance.

The results in Table 2 indicate that generative AI led to large improvements in productivity on

[3]These results are currently in progress and not in the current draft of the manuscript.

the generalist central banking tasks as measured by the score on the task. Using the within-subjects design in column 1, we find that participants scored 2.6 points higher on an average task when they had access to generative AI versus when they did not. This represents an average improvement of 44 percent (= 2.593/5.859) relative to when participants did not have access to generative AI. We find qualitatively similar results when we limit our sample to the first seven tasks and use the between-subjects design in column 2. Participants who had access to generative AI in the first seven tasks scored 2.2 points higher than those who did not have access in the first seven tasks, representing an average increase of 33 percent relative to the control mean.

We find that generative AI also led to similar improvements using our time measure of productivity. The results in columns 3 and 4 indicate that participants with access to generative AI took around six minutes fewer to complete the seven tasks versus those who did not have access to generative AI. This amounts to a productivity improvement of around 21 percent (= 6.314/30.155).

Overall, the results indicate that generative AI led to large improvements in productivity across the generalist tasks. The effect sizes (relative to the control mean) were about twice as large for improvements in quality versus improvements in time.

3.1.2 Impact by task type

We next explore the impact of generative AI on productivity across the various generalist tasks types. Our design featured the seven task generalist tasks types, and we can examine how generative AI impacted scores across the various task types. Given that we do not have the time spent on each individual task, we cannot use the time measure of productivity for this analysis. Instead, we focus on the score for each of the generalist task types.

Table 3 examines the impact of generative AI across the various generalist task categories. It displays the results of regressions of task score on a binary indicator for having access to generative AI, phase fixed effects, and participant controls. Each regression limits the sample to a specific type of generalist task, which is listed at the top of the table..

The results in Table 3 indicate that the productivity benefits of generative AI are widespread across task types. Generative AI led to a large increase in productivity for six out of the seven task types. The impact is not only limited to cognitive tasks, where generative AI increased performance by 76 percent

for routine cognitive tasks and 51 percent for non-routine cognitive tasks. We also see large increases in performance for non-routine manual tasks, analytical tasks, leadership tasks, and communication tasks. The one exception is for routine manual tasks, where participants saw a decrease in performance.

3.1.3 Distributional impact of generative AI

We next examine the distributional impacts of generative AI across participants. Given that we have 14 observations per participant for the generalist tasks, half of which were with generative AI and half of which were without, we are able to estimate a participant-specific treatment effect. To do so, we loop through each participant, and regress that participant's task score on a binary indicator for whether they had access to generative AI. The coefficient of interest captures the difference in average productivity across all 14 tasks, when the participant uses generative AI versus does not.

Figure 3 displays the distributional impact of generative AI on individual productivity across our participants. The results in Figure 3 illustrate that the productivity benefits of generative AI are large for the majority of the distribution. 94 percent of participants scored higher when they had access to generative AI versus when they did not.

3.2 Impact of generative AI on performance in specialized central banking tasks

4 Conclusion

References

Autor, David H, "The "task approach" to labor markets: an overview," *Journal of Labour Market Research*, 2013, *46* (3), 185–199.

Dell'Acqua, Fabrizio, Edward McFowland III, Ethan R. Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R. Lakhani, "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality," Working Paper No. 24-013, Harvard Business School Technology & Operations Mgt. Unit September 2023.

Perkowski, Patryk and Aleš Maršál, "Unlocking the productiviyty benefits of generative AI: Evidence from three central banks," *Working paper*, 2024.

Tables

Table 1: **Summary statistics**

	N	Mean	SD	Min	Max
By participant Education, less than high school	97	0.00	0.00	0.00	0.00
Education, high school or associate's degree	97	0.03	0.17	0.00	1.00
Education, university	97	0.02	0.14	0.00	1.00
Education, masters degree or higher	97	0.92	0.28	0.00	1.00
Age, in years	97	39.85	8.45	21.00	61.00
Female	97	0.35	0.48	0.00	1.00
Tenure at bank, in years	97	8.09	9.76	0.08	60.00
Manages or leads team	97	0.05	0.22	0.00	1.00
Has a laptop	97	0.98	0.14	0.00	1.00
Total weekly hours	97	46.74	11.51	8.50	60.00
Percent of work week, routine cognitive	97	0.18	0.12	0.00	0.60
Percent of work week, non-routine cognitive	97	0.26	0.16	0.00	0.81
Percent of work week, routine manual	97	0.03	0.06	0.00	0.37
Percent of work week, non-routine manual	97	0.02	0.04	0.00	0.25
Percent of work week, analytical	97	0.31	0.19	0.00	0.93
Percent of work week, leadership and management	97	0.16	0.10	0.00	0.50
Percent of work week, interpersonal and communication	97	0.03	0.07	0.00	0.43
Heard of LLM	97	0.95	0.22	0.00	1.00
Used LLM for either personal or work use	97	0.75	0.43	0.00	1.00
Used LLM for personal use	97	0.68	0.47	0.00	1.00
Used LLM for workplace use	97	0.53	0.50	0.00	1.00
Score on task	1356	7.23	3.17	0.00	10.00
AI treatment condition	1358	0.50	0.50	0.00	1.00
Time (minutes)	194	27.04	5.83	12.00	37.00

Notes: This table displays summary statistics of the sample.

Table 2: **Impact of generative AI on productivity in generalist central banking tasks**

	Score on task		Time to complete	
	(1)	(2)	(3)	(4)
AI treatment condition	2.593***	2.216***	-6.314***	-6.854***
	(0.154)	(0.224)	(0.656)	(0.986)
R2	0.325	0.315	0.301	0.370
Observations	1,356	677	194	97
Control mean	5.859	6.641	30.155	31.095
Fixed effects	Yes	No	Yes	No
Controls	No	Yes	No	Yes
p-value	0.000	0.000	0.000	0.000

Standard errors in parentheses

* p<0.10, ** p<0.05, *** p<0.010

Notes: This table examines the impact of generative AI on productivity in generalist central banking tasks. The first two columns use the score on the task as the measure of productivity, while the second two columns use the time to complete the task (in minutes). Column 1 displays the results of a regression of the task score, on a binary indicator for whether the participant is in the AI treatment condition, participant-level fixed effects, and task-type fixed effects. Column 2 conducts the between-subjects design by subsetting the sample to only the first seven tasks. It displays the results of a regression of task score on the AI indicator, fixed effects for task types, and participant prior exposure to generative AI. Column 3 displays the results of a regression of the time to complete the task on a binary indicator for whether the participant is in the AI treatment condition and a fixed effect for the phase of the experiment. Column 4 conducts the between-subjects design by subsetting the sample to only the first seven tasks. It displays the results of a regression of time on the AI indicator and participant prior exposure to generative AI. All regressions include robust standard errors clustered at the participant level.

Table 3: **Impact of generative AI on productivity, by generalist task types**

	Score on task, by task type						
	Routine Cognitive	Non-routine Cognitive	Routine Manual	Non-routine Manual	Analytical	Leadership	Communication
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
AI treatment condition	3.5***	2.9***	-1.3***	4.1***	2.7***	3.4***	3.0***
	(0.3)	(0.3)	(0.3)	(0.3)	(0.4)	(0.3)	(0.3)
R2	0.57	0.43	0.12	0.55	0.38	0.38	0.30
Observations	194	194	194	194	194	192	194
Control mean	4.6	5.6	8.1	4.7	5.7	5.8	6.6
Effect size, %	76	51	-15	87	47	60	45
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
p-value	0.00	0.00	0.00	0.00	0.00	0.00	0.00

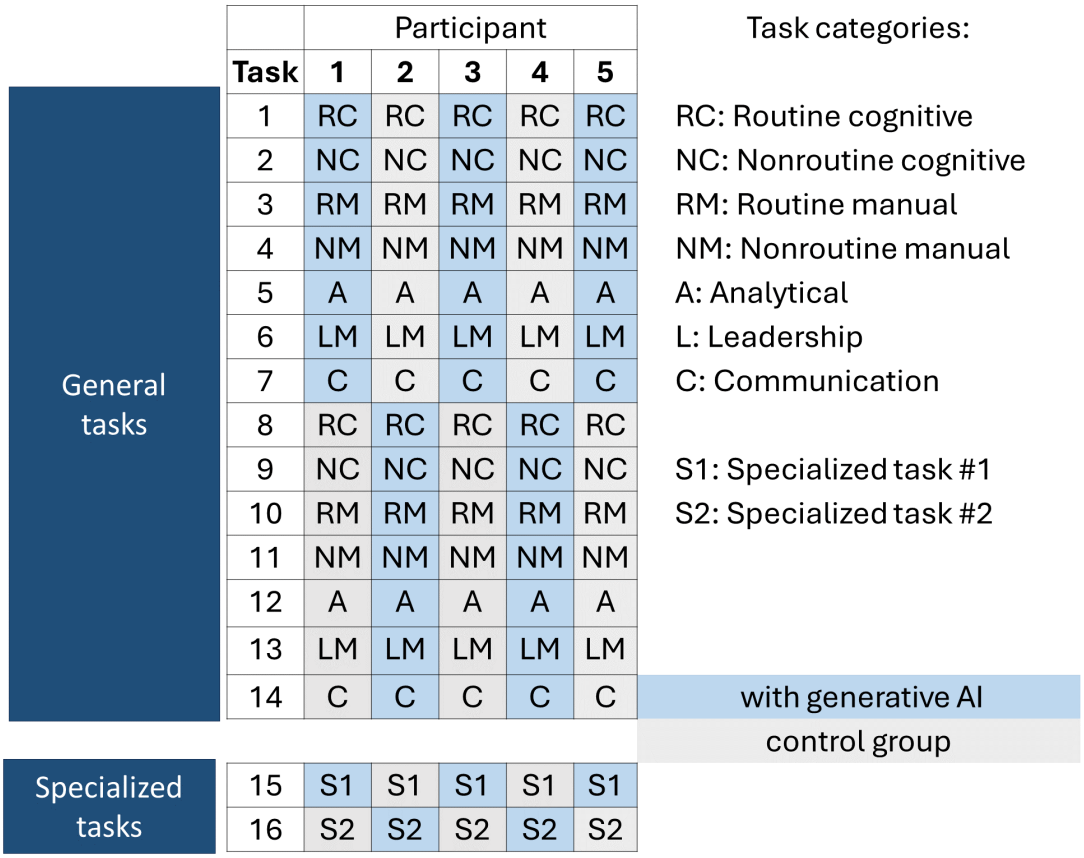
Standard errors in parentheses

* p<0.10, ** p<0.05, *** p<0.010

Notes: This table examines the impact of generative AI on productivity by generalist central banking tasks. It displays the results of a regression of task score on a binary indicator for having access to generative AI, phase fixed effects, and participant controls. Each regression limits the sample to a specific type of generalist task, which is listed at the top of the table. All regressions include robust standard errors clustered at the participant level.

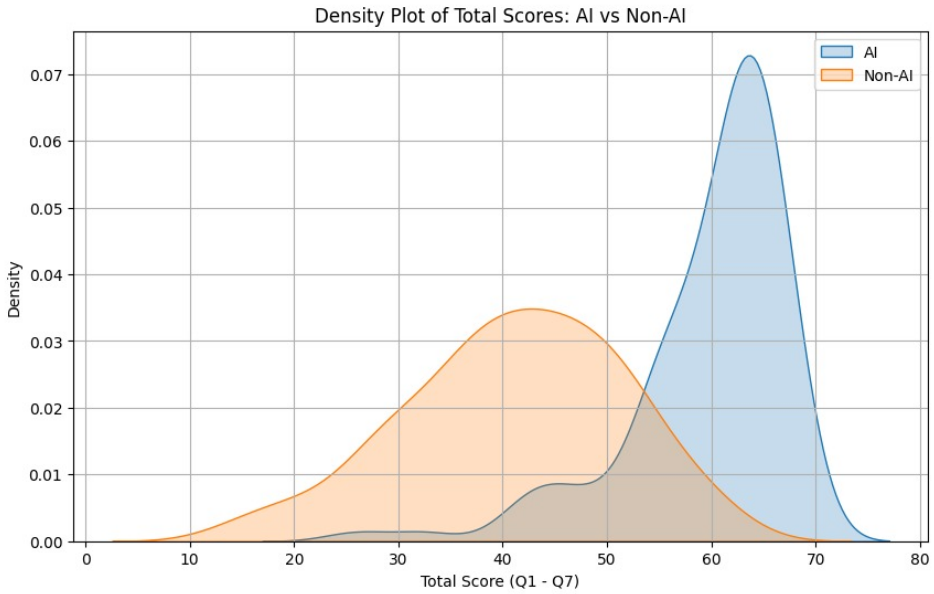
Figures

Figure 1: **Structure of experiment**

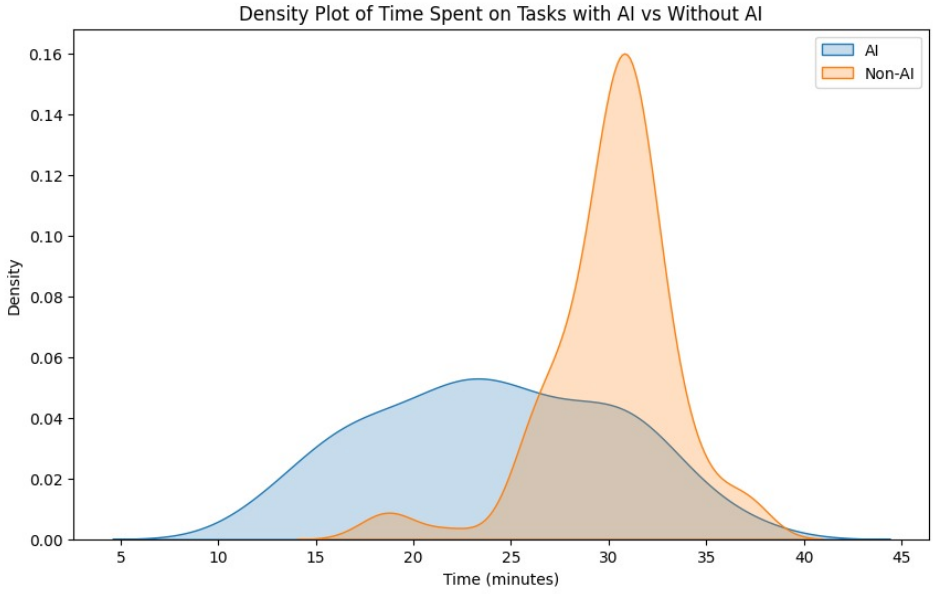


Notes: This figure displays the structure of the experiment

Figure 2: **Distribution of aggregate productivity measures, by generative AI treatment condition**

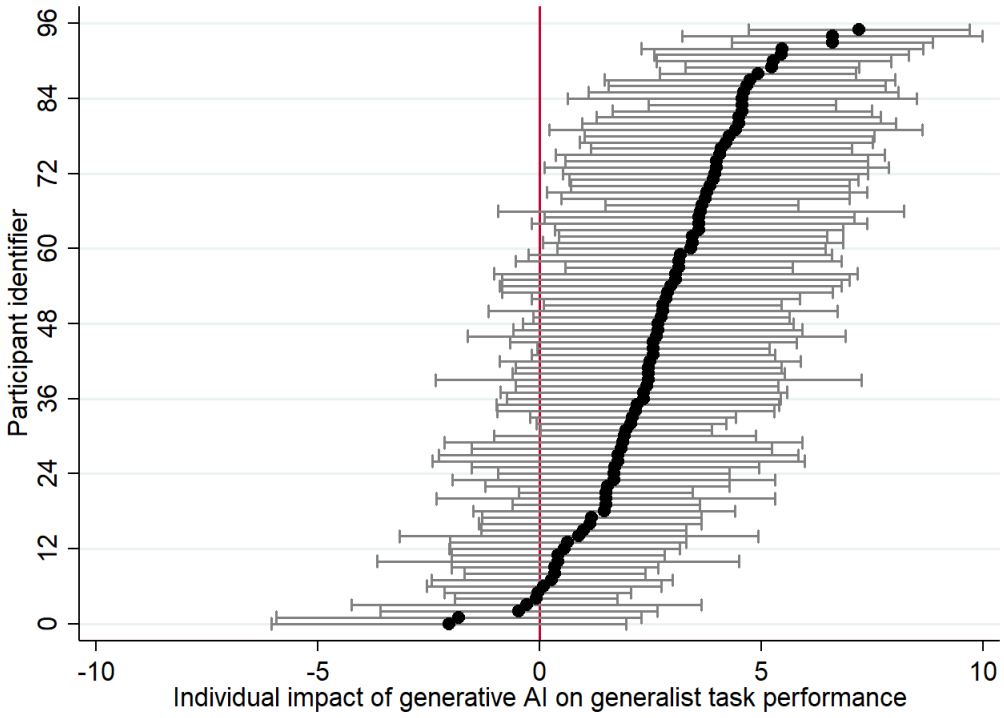


(a) Total task score



(b) Total time to complete the task

Figure 3: **Distributional impact of generative AI on productivity in generalist tasks, by participant type**



Notes: Tihs figure examines the distributional impact of generative AI on performance in generalist tasks for each participant. We loop through each participant, and regress that participant's task score on a binary indicator for whether they had access to generative AI. The coefficient displayed in the figure captures the difference in average productivity across all 14 tasks, when the participant uses generative AI versus does not. We sort the treatment effects from largest to smallest.

Appendix: For Online Publication Only

A Additional experimental materials

A.1 General workplace tasks

A.2 Specialized workplace tasks