No Great Equalizer: Experimental Evidence on AI in the UK Labor Market

Matthias Haslberger
University of St. Gallen

Jane Gingrich
University of Oxford

Jasmine Bhatia
Birkbeck, University of London

October 6, 2023

Draft - Comments Welcome!

Abstract

Generative artificial intelligence is already transforming how people work. There is an emerging consensus in early studies that it reduces inequalities in performance within specific occupational groups; however, the question of whether these results generalize to the labor market at large remains open. We conducted a pre-registered online experiment with a representative sample of the UK working-age population. We randomly assigned participants to treatments that encouraged or discouraged the use of ChatGPT and then asked them to complete a set of tasks of varying complexity and ambiguity. We find that exposure to ChatGPT increased productivity in all tasks, with greater benefits observed in more complex and less ambiguous tasks. ChatGPT did reduce performance inequality *within* occupational groups in most cases, but not *between* educational or occupational groups. Inequalities between younger and older workers even increased. This study indicates that generative AI has the potential to improve worker performance in a wide array of tasks, but the impact on aggregate inequalities is likely to depend on task-specific features and workers' characteristics.

We are grateful for the feedback and support we received from Noah Bacine, Patrick Emmenegger, and Lukas Paleckis. All remaining errors are our own. This project has been funded in part by the Canadian Institute for Advanced Research and the Schweizerisches Staatssekretariat für Bildung, Forschung und Innovation in the framework of the GOVPET research project.

Corresponding author: Matthias Haslberger (matthias.haslberger@unisg.ch)

1 Introduction

An emerging body of work on the labor market implications of generative AI suggests two clear consequences: it can increase productivity in complex tasks, but that these gains mostly accrue to those with lower or medium-levels of productivity (Noy and Zhang, 2023; Dell'Acqua et al., 2023; Peng et al., 2023; Brynjolfsson, Li and Raymond, 2023). When scaled up, these arguments suggest both an asymmetrically disruptive effect of generative AI on workers - it will impact more educated and professional workers – and an equalizing aggregate effect - it will compress differences among workers. Indeed, these claims have combined into an emerging public narrative about generative AI as disproportionately affecting white collar jobs (Cain Miller and Cox, 2023).

Several early studies have focused on how generative AI increases individual-level productivity for specific tasks and among specific occupational groups: college educated professionals specializing in writing tasks (Noy and Zhang, 2023), call centers employees (Brynjolfsson, Li and Raymond, 2023), software developers (Peng et al., 2023), and consultants (Dell'Acqua et al., 2023). However, we know less about how these findings generalize across different types of tasks and a broader range of skill or occupational groups. We think there are two reasons for caution in accepting the emerging narrative of an equalizing effect: there is substantial uncertainty about for which tasks generative AI enhances performance, and there may be skill disparities (or other factors) that shape workers' ability to adopt and best utilize technology in ways that reduce aggregate compression effects.

First, in what Autor (2022, 19) calls "Polyani's revenge", "*computers now know more than they can tell us*," at least in some domains. While generative AI has shrunk the set of tasks where workers' tacit knowledge or creativity is critical to performance, it also generates new errors (Frey and Osborne, 2023), resulting in uncertainty about the limits of its capabilities (Autor, 2022). It is also increasingly likely that there is not a linear relationship between task complexity and gains from generative AI, which Dell'Acqua et al. (2023) have characterized as a "jagged frontier" of AI capabilities. The implications of these changes for aggregate differences across workers remain uncertain.

Second, while generative AI challenges the existing paradigm of routine-biased technological change (RBTC), in which technology is most prone to replacing middle-skilled workers whose job comprises mostly routine tasks (Autor, Levy and Murnane, 2003; Acemoglu and Autor, 2011), the interaction between new technologies and skills is still uncertain. Existing work provides convincing evidence for a productivity-enhancing and equalising effect of AI within occupations. However, performance improvements for individuals in these groups could still leave substantial aggregate inequalities across occupational or skill groups.

To examine these possibilities, we provide, to our knowledge, the first analysis of the effects of generative

AI from a general population sample of British workers, encompassing the full range of occupational levels, skill sets, and demographics. We compare a randomly selected treatment group encouraged to use ChatGPT to complete three tasks of varying complexity and ambiguity with a control group instructed not to use any generative AI tools.

As expected, we find that exposure to ChatGPT significantly improved the efficiency and quality of responses. However, there were differences across tasks, with the greatest gains in more complex and less ambiguous tasks. Second, in most cases - though not all - we document a slight reduction in performance inequality within occupations, in line with existing evidence. However, exposure to ChatGPT did not result in a convergence in performance between different educational or occupational groups, and older workers benefited less than younger workers from ChatGPT exposure. Thus, while AI appears to compress the returns to occupation-specific skills, it reproduces existing inequalities between broader skill groups.

2 Motivation

Research examining the effects of generative AI on workers in non-routine occupations rests on two linked arguments: generative AI has the potential to perform more complex tasks and to reduce the complementarity between existing skills and technology, compressing differences in performance among groups of workers previously insulated from automation (Felten, Raj and Seamans, 2023). This work has prompted commentary expecting that generative AI might reduce labor market inequalities, and initial evidence seems to bear out this expectation (Noy and Zhang, 2023; Dell'Acqua et al., 2023; Peng et al., 2023; Brynjolfsson, Li and Raymond, 2023).

While early studies show support for compression effects, their focus on specific tasks or subsets of workers leaves critical questions about the aggregate effects of generative AI unanswered. Below, we schematize why relying on narrow samples can lead us to overestimate compression effects, and then outline expectations for our study.

Task-Based Dynamics Task-based theories suggest that certain tasks are more readily automatable than others. This is in part because some tasks draw on tacit knowledge that was historically difficult to specify and because humans were thought to possess unique cognitive capacities (e.g. creativity, judgment) that could not be technologically replicated (Autor, 2015). The advent of generative AI potentially alters this logic.

Generative AI models have been described as a general purpose technology which contains "sparks of artificial general intelligence" (Agrawal, Gans and Goldfarb, 2019; Bubeck et al., 2023; Eloundou et al.,

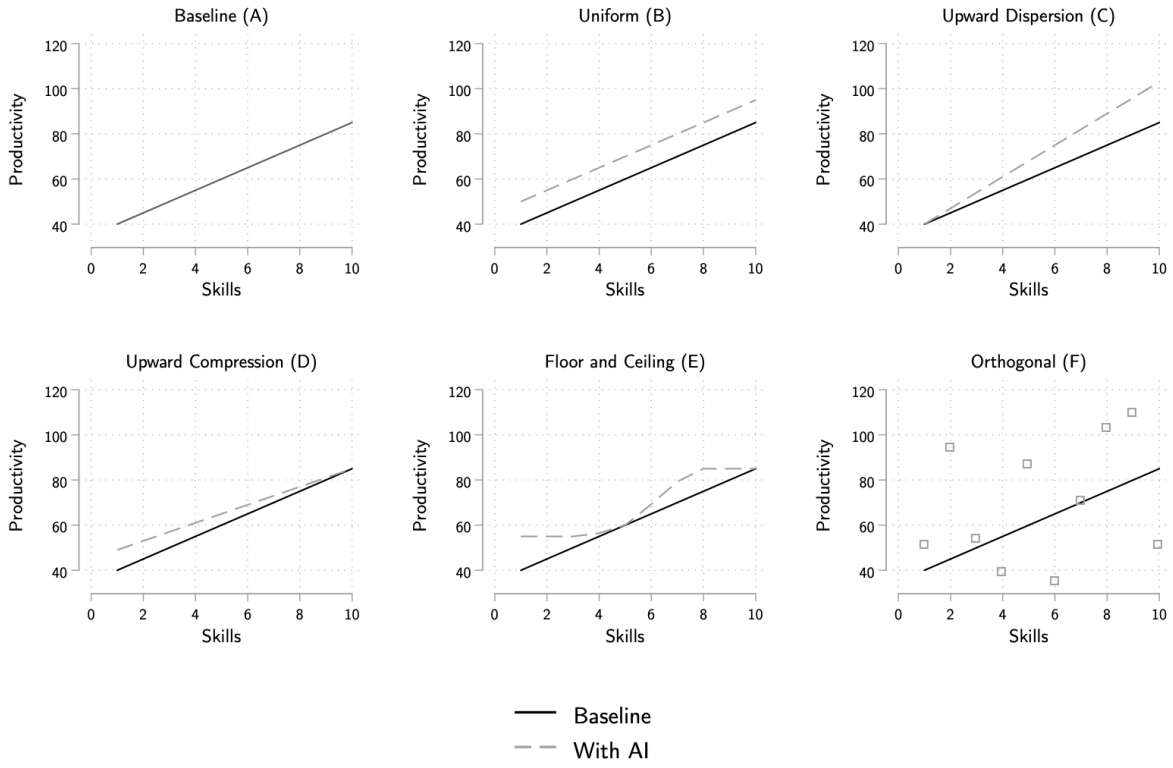
2023). This work suggests that AI can increasingly perform tasks previously insulated from automation, such as non-routine cognitive tasks that require abstract reasoning, prediction, or judgment (Eloundou et al., 2023; Felten, Raj and Seamans, 2023). Viewed from this perspective, generative AI should reduce variation between tasks by increasing floor-level performance in complex tasks and thus increasing mean overall performance in these tasks. Collectively, these dynamics should reduce the performance differential between more and less complex tasks. Put differently, as task complexity rises, the gap between AI users and non-users should increase.

However, as scholars of task complexity have long argued, complexity is itself multi-dimensional (Wood, 1986). Generative AI technologies may perform well in some areas and poorly in others (Frey and Osborne, 2023). Indeed, Dell'Acqua et al. (2023) speak of a "jagged frontier" of AI capabilities, and the quality of ChatGPT and other generative AI may already be changing over time (Chen, Zaharia and Zou, 2023; del Rio-Chanona, Laurentsyeva and Wachs, 2023). Generative AI can ironically struggle with certain complex but verifiable tasks, such as calculating mathematical proofs. If the AI itself introduces errors, then it may induce new variation among complex tasks. In other words, it could lead to a compression pattern in some tasks and not others. Testing this possibility requires examining multiple types of tasks with different levels of complexity.

Skill-Based Dynamics A second argument is that generative AI changes the relationship between productivity and underlying skills – as measured either by baseline performance or by existing qualifications such as higher education. Historically, the relationship between skills and technology-induced performance has varied. Some technologies, such as the loom, replaced large subsets of skilled workers, thereby reducing or severing the link between individual skills and productivity; others, like information technology, complemented skilled workers and enhanced the value of and demand for cognitive skills (Acemoglu and Autor, 2011). Figure 1 schematizes the potential variation in how skills and generative AI technology interact as compared to a hypothetical non-AI assisted baseline (A): uniformly productivity enhancing, maintaining existing skill gradients (B); complementing existing skills and thus increasing skill gradients (C); compressing skill differences (D); exerting floor and ceiling effects that compress skill differences non-linearly (E); or creating new divides orthogonal to existing skills (F).

As outlined above, the broad early consensus suggests that generative AI technologies produce compressing patterns (D). Noy and Zhang (2023) show that for college educated workers doing mid-level professional writing tasks, there is an increase in overall productivity and a reduction in inequality between workers for those using ChatGPT. Brynjolfsson, Li and Raymond (2023) find a similar effect examining productivity among call center workers. Those with higher baseline levels of productivity benefit less from AI use,

Figure 1: Simulated Skill Based Effects



Note: The figure shows the logic of six different scenarios linking baseline skills to performance.

whereas lower productivity workers gain more. Dell'Acqua et al. (2023) likewise find a compressing effect among highly skilled management consultants.

However, in generating insights from a narrow population group, these results do not rule out alternative scenarios outlined in Figure 1. For instance, they could demonstrate floor and ceiling effects (E) for top-level performers, where AI compresses performance for highly skilled workers but maintains or widens performance gaps between groups with a broader range of abilities. Alternatively, it could be the case that AI produces new disparities in productivity levels that are orthogonal to skills (F) but cannot easily be compressed (e.g. age). There is a further possibility that when we consider task and skill differences jointly there may be areas with more uniform increases (B), and others where the highly skilled more rapidly develop an understanding of when to deploy generative AI for a given task in ways that compensate for otherwise lower gains (C).

To explore these possibilities, we developed a survey experiment with a general population sample and a set of tasks with varying levels of complexity and ambiguity[1]. We go beyond existing studies by examining respondents across different levels of formal education and across occupational groups. This allows us to test hypotheses regarding potential differences between demographic groups. We pre-registered the following expectations.

H1a: Respondents in the treatment group (encouraged to use ChatGPT) will outperform respondents in the control group (discouraged from using ChatGPT).

H1b: Respondents in the treatment group will take less time to complete the tasks than respondents in the control group.

H1c: The performance differential between the treatment and control groups will increase in accordance with the complexity of the tasks, with task 1 being the least complex and task 3 being the most complex.

H2: Exposure to ChatGPT a) reduces/ b) increases within-task variation in performance compared to the control group.

H3: a) Older/ b) Female/ c) Less educated/ d) Non-professional respondents who are exposed to AI are less likely to use it, and benefit less from using it.

[1]We define complexity in terms of the difficulty for a human of average cognitive abilities to perform the tasks without technological help. We define ambiguity as the degree to which the task was open-ended (e.g. the extent to which there was a 'correct' answer to the task(s)). We did not pre-register a hypothesis on ambiguity; rather, this emerged as a finding in our results and merits further research.

3 Methods

To test these hypotheses, we conducted a pre-registered online experiment for which we recruited 1041 respondents through the survey company YouGov.[2] In contrast to other early studies of the effects of ChatGPT (Noy and Zhang, 2023; Dell'Acqua et al., 2023), our sample is largely representative of the UK working-age population. The sample was split into a treatment group (N = 504) which was encouraged to use ChatGPT and a control group (N = 537) which was instructed to complete the survey without using ChatGPT.[3] Fieldwork for the survey took place from 19th - 28th July 2023, following a pre-test in early July 2023. The median time for completion of the survey was approximately 28 minutes. Participants were informed before undertaking the survey that they would be compensated at twice the normal YouGov rate for a survey of this length. To increase participant effort, we furthermore offered a performance-based reward amounting to 50% of the participation reward to the top 10% of respondents in each group. After some self-assessment questions, participants were asked to complete three short text-based tasks of increasing complexity.[4]

In the first task, participants were presented with an email addressing a hypothetical workplace dispute and were asked to make any improvements they deemed appropriate. The email included deliberate grammatical and spelling errors and was written in a harsh, unprofessional tone. The evaluation criteria consider whether the errors have been corrected and whether the tone of the email has been made more constructive. In the second task, respondents evaluated the persuasiveness of two texts of about 500 words, each presenting opposing views on the merits of a universal basic income. The answers were graded based on whether they provided a well-reasoned and comprehensive assessment. In the final task we asked participants to answer three short questions about a complex text.[5] The grading criteria evaluate whether respondents correctly distilled the relevant information from the text. Throughout the paper, we refer to the three tasks as the email task, assessment task, and comprehension task, respectively. The three tasks increase in complexity, and are of a sufficiently general nature so that not only knowledge workers are likely to be familiar with them.[6] While not covering the entire range of use cases for AI, our design allows us to investigate the impact of AI use for tasks of varying complexity in a representative sample of the working age population.

We obtain two measures of respondents' performance on the tasks. We evaluate their answers on a five-point scale according to pre-specified criteria and record the time they take to complete the tasks. We used

[2]The pre-analysis plan can be viewed here.

[3]Additional details about the survey are provided in Appendix A. Despite their best efforts, YouGov could not at the time provide a fully representative sample. For example, men are overrepresented in the sample.

[4]YouGov provided demographic information, including age, sex, education, occupation, income, and other variables.

[5]The text and questions are adapted from a publicly-available LSAT practice test (Law School Admission Council, 2007).

[6]Nevertheless, with increasing complexity, an increasing share of respondents reports that they rarely or never perform similar tasks, as we show in Figure D3. However, there are no systematic differences between treatment groups

an innovative AI-supported approach to evaluate the answers to the tasks. Inspired by Gilardi, Alizadeh and Kubli (2023), we used ChatGPT (GPT-4) to grade the answers based on a detailed grading scheme (see Appendix G). Two of the authors each coded a random subset of 100 answers to assess the reliability of the ChatGPT grading. Average inter-coder agreement between ChatGPT and human coders was similar to human – human inter-coder agreement (see Table G1). Furthermore, the use of ChatGPT led to substantial time savings, illustrating the potential of LLMs to drastically increase the efficiency of coding text data for quantitative analysis. We moreover recorded the time (in seconds) respondents spent on each of the tasks. Since some respondents spent a large amount of time on a single task, we top-coded the time measure for each task at 900 seconds to exclude a small number of respondents who likely became distracted. Our main results are robust to choosing a less stringent cut-off and to excluding those respondents. Finally, respondents in the treatment group may have refrained from using ChatGPT, while respondents in the control group may have defied instructions and used AI. Section C details the methods we used to encourage and monitor compliance, which reassure us that respondents largely complied with instructions.

4 Results

4.1 ChatGPT Increases Productivity Across Tasks

We first show evidence that exposure to ChatGPT increases overall productivity in line with hypotheses 1a and 1b. We find mixed support for the effect on between-group differences posited in H1c. Finally, we find tentative support for H2a over H2b with regard to within-task differences in performance.

4.1.1 Overall and Between-Task Differences

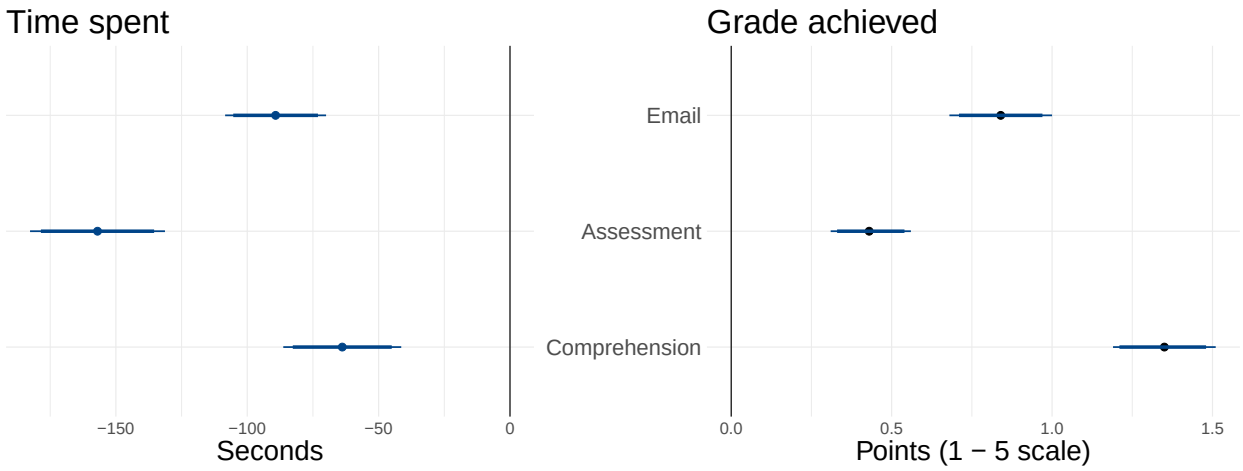
We find clear evidence that ChatGPT helps respondents give better answers to the tasks in a shorter amount of time. In Figure 2, we plot the intention-to-treat (ITT) effects.[7] The difference in time spent and grade achieved between treatment and control group is significant and sizeable for all three tasks. Predicted time savings range from 0.32 to 0.69 standard deviations (SD), while performance increases amount to between 0.41 and 0.92 SD. This provides clear evidence for hypotheses 1a and 1b: respondents with access to ChatGPT tend to perform text-based tasks of varying complexity both faster and better. This is in line with evidence from existing studies (Noy and Zhang, 2023; Dell'Acqua et al., 2023; Brynjolfsson, Li and Raymond, 2023). In their longer tasks (average duration close to 30 minutes), Noy and Zhang (2023) found a larger reduction

[7]Since we do not have full control over compliance with the instructions, treatments effects represent the ITT. However, in Section C we approximate the average treatment effect (ATE) by distinguishing between self-reported AI users and non-users, with even stronger results. The ITT presented here thus constitutes a lower-bound of the true effect, while the ATE represents an upper-bound.

in time spent and a smaller increase in grades achieved, but in general the size of the treatment effects expressed in terms of standard deviations is similar.

However, we find mixed evidence regarding a general relationship between task complexity and benefits of ChatGPT use (H1c). While the return to AI use is greatest in the comprehension task with 1.4 points (0.92 SD), it boosts scores in the assessment task by only 0.4 points (0.41 SD), while in the - arguably - least complicated email task, the average score in the treatment group is 0.8 points (0.61 SD) higher than in the control group.[8] The pattern in time spent offers a tentative explanation for this finding: the greater the difference in time spent, the smaller the difference in the grades achieved. Thus, it appears that ChatGPT users face a trade-off between time savings and performance improvements. Whether individuals and organizations will funnel the productivity gains from AI into higher output at constant quality and hours, constant output and quality with lower hours, or constant output and hours at higher quality, will be crucial to observe as the rollout proceeds.[9]

Figure 2: Coefficient plot of treatment effects



Note: The figure shows estimated treatment effects based on linear probability models, with 90% and 95% confidence intervals (thick and thin lines). N = 1,041. Full model output in Table E2.

4.1.2 Within-Task Differences

Existing studies based on a within-person design in narrow occupational groups find that less able participants benefit most from using ChatGPT, implying reduced within-task variation (H2a) (Noy and Zhang, 2023; Dell'Acqua et al., 2023). However, it is not obvious that the same pattern holds in the population at

[8] The difficulty of the tasks is of course somewhat subjective. Nevertheless, our judgment is validated by the average scores in the control group which follow the predicted pattern, see Table D1. Furthermore, respondents were on average most familiar with the email task and least familiar with the comprehension task, see Figure D3.

[9] We expand on the relationship between time spent and performance in Appendix F.

large. For example, AI might benefit better educated or otherwise privileged workers who can leverage the technology more effectively in what has been described as a "winner-takes-most" dynamic (H2b) (Agrawal, Gans and Goldfarb, 2019; Lane and Saint-Martin, 2021).

Table 1 suggests that H2a is closer to the truth, although there is no consistent pattern. It shows the coefficient of variation (CV) as a measure of the relative variability of the grades.[10] Within-group variation in grades is significantly lower in the treatment group for the email ($p < 0.1$) and comprehension task ($p < 0.05$), and significantly higher for the assessment task ($p < 0.01$). Thus, while average answer quality increases with AI exposure, variation in answer quality relative to the mean exhibits no clear tendency in a full population sample. Even within broad occupational categories, we see no unambiguous pattern (see Table E1). This qualifies the findings in other recent studies which document a compression of within-task performance in occupation-specific samples. Across the full breadth of the skill distribution, AI is no great equalizer.

The smaller improvement and the greater variation in answer quality in the treatment group in the assessment task point to a difference in the nature of this task that is independent of its overall complexity. The assessment task required respondents to make a reasoned judgment; there was no right or wrong answer (as we stated explicitly in the instructions). This is a task of medium complexity for humans, as evidenced by the average answer quality in the control group. However, ChatGPT struggles with ambiguity and conveying the instructions for the task to ChatGPT was not straightforward. Hence, while the assessment task is easier than the comprehension task without ChatGPT, it may be easier to prompt ChatGPT to solve the comprehension task than to assist with the assessment task. This illustrates that it may be difficult to effectively employ artificial intelligence in situations where it is necessary to navigate ambiguity, and where therefore prompting the AI requires higher skills. This indicates a certain degree of skill complementarity in using AI for tasks which are characterized by ambiguity, which may counterbalance the compressing effect of AI in other tasks.[11] We did not anticipate this qualitative difference between the tasks when pre-registering our hypotheses. Undoubtedly, delineating the borders of the "jagged frontier" (Dell'Acqua et al., 2023) of AI capabilities and systematizing the determinants of susceptibility and complementarity will greatly facilitate further research into the consequences of AI. In this endeavour, it is imperative for social scientists to be guided by the emerging literature on the technical characteristics and limitations of LLMs (McCoy et al., 2023; Berglund et al., 2023; Chen, Zaharia and Zou, 2023; OpenAI, 2023). For now, we observe that the hierarchy of task difficulty may differ for humans and AI tools, and posit that AI may be less useful for more ambiguous tasks.

[10] Figure D4 additionally shows density plots.
[11] The positive, albeit insignificant, interaction coefficient in Panel D of Table E3 supports this conjecture.

Table 1: Within-group variation in grades

Task	CV Treated	CV Control	F-statistic	p-value
Email	36.227	38.344	1.120	0.098
Assessment	41.525	35.552	1.364	0.000
Comprehension	46.412	50.028	1.162	0.044

Note: CV: coefficient of variation.

4.2 Group Differences in AI Use and Productivity

One of the main concerns surrounding the rise of artificial intelligence is that some demographic groups could struggle to adapt to the attendant changes in work organization or even be made redundant. A large literature on the economic and political consequences of globalization and automation convincingly shows that where structural change leads to concentrated losses, such as the decline of factory jobs for middle-skilled white men in the U.S., affected groups may stage a backlash (Autor, Dorn and Hanson, 2013; Autor et al., 2020). It is therefore crucial to anticipate potential cleavages that may appear in the wake of AI adoption. To put it bluntly, who will be the middle-skilled white men of the AI revolution?

4.2.1 The "Usual Suspects" Are More Likely to Use AI

We hypothesized that older, female, and less educated respondents, as well as those working in non-professional occupations would be less likely to use AI and, conditional on using it, benefit less from the technology. This is because these groups are generally perceived to be less exposed to cutting-edge technologies and therefore less tech-savvy. Bearing in mind that our sample includes only respondents who stated that they have a ChatGPT account, we still find the expected differences in self-reported *frequency* of use in Table 2. First of all, it must be noted that even in our positively selected sample, frequent users are still a minority. In all key categories no more than a quarter to a third of respondents stated that they use ChatGPT even once a week. Accounting furthermore for purely "recreational" use of AI technologies, it is clear that workplace adoption of generative AI is still in its early stages. Respondents under 50, men, and people working in professional and managerial occupations are approximately 10 percentage points more likely to report using ChatGPT at least once per week. The difference between university educated and non-university educated respondents, while in the expected direction, is less pronounced. We therefore find evidence that patterns of use correspond to prevalent stereotypes regarding novel digital technologies.

4.2.2 Who Benefits Most from Using AI?

The key question for policymakers and businesses, however, is what happens when people with different levels of affinity towards AI technology are made to use it? If the descriptive patterns are indicative of AI

Table 2: Frequent users of ChatGPT by demographic characteristics

	ChatGPT Use (%)	
Group	Infrequent	Frequent
15 - 35 years old	68	32
36 - 50 years old	69	31
51 - 80 years old	78	23
Female	76	24
Male	67	33
University degree	69	31
Non-university degree	73	27
Professional/managerial occupation	67	33
Other occupation	76	24

Note: Respondents are classified as frequent users if they report using ChatGPT at least once per week.

skills, AI might benefit already privileged groups and exacerbate existing labor market inequalities (Lane and Saint-Martin, 2021; Albanesi et al., 2023). Thus, to detect whether returns to AI exposure vary by age, sex, education, or occupation, we interact our treatment dummy with the respective demographic indicators. Full model results for the figures are presented in Table E3. Note that the interaction coefficients can be interpreted as the difference in differences of the treatment effect on the respective groups.

Differences in the frequency of use notwithstanding, the benefits of being exposed to AI appear to be fairly evenly distributed, with one major exception: age. All age groups perform significantly better on all three tasks if they are encouraged to use ChatGPT. Yet, across all three tasks, we find that younger users benefit more from exposure to ChatGPT than older users - even as they take the shortest amount of time (Figure 3). While the performance advantage of people aged 35 or less over those 51 and older is small or nonexistent in the control group, it ranges between 0.3 and 0.7 points in the treatment group. The difference in differences is statistically significant at the 10% level for the email task and at the 5% level for the comprehension task. We do not find heterogeneous treatment effects on response time: while older respondents take longer, AI exposure leads to a similar reduction in response time across age groups. Importantly, this is not due to educational expansion: the results are unchanged if we control for degree status. Thus, workers in all age groups become more productive with generative AI, but younger workers appear significantly better placed to take advantage of the technology. It is further important to bear in mind that our sample likely includes comparatively tech-savvy older workers and our results therefore probably understate the advantage to younger workers. As AI penetrates into workplaces, policies that enable older workers to adapt will therefore increase in importance.

By contrast, we find only minimal sex differences in treatment effects, despite the greater familiarity of men with ChatGPT. Rather, exposure to the treatment improves the performance of women and men by approximately the same amount. Men do complete the tasks faster with AI, although the only statistically

significant difference in differences is for the comprehension task. Thus, men do not appear to gain a material performance advantage from their greater familiarity with AI technology. Overall, we find no indication that there is anything about artificial intelligence that inherently favours women or men.

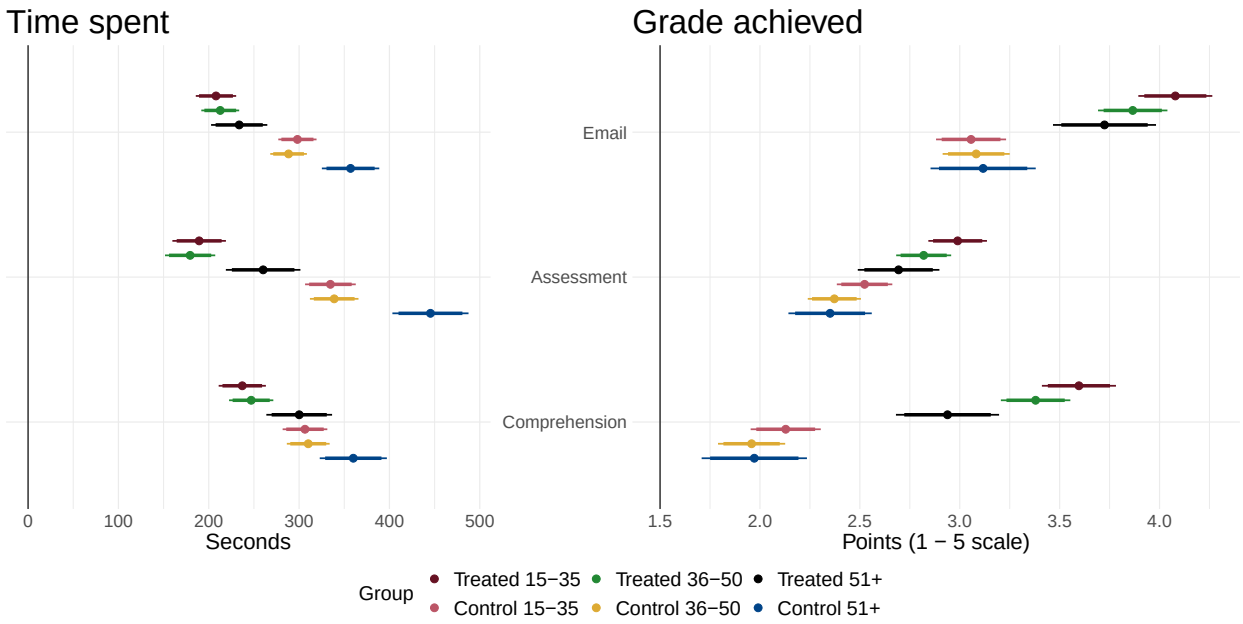
We also find no evidence that more educated workers or those working in professional or managerial occupations will monopolise the benefits of generative AI - but nor does it erase existing educational group differences. While respondents with a university degree achieve higher grades on all three tasks (albeit only significantly so in the most complex comprehension task), the treatment improves scores and reduces completion times in both groups by approximately the same amount. The difference in differences is statistically insignificant in all cases. Similarly, while professionals and managers in the control group perform significantly better in the assessment task (10% level) and comprehension task (5% level), in the treatment group there is only a statistically insignificant reduction in the performance differential. There are no significant occupational differences in time spent, nor differential treatment effects. Thus, greater initial familiarity with generative AI does not necessarily translate into greater performance improvements along educational or occupational lines.These results are more in line with the uniform productivity effect (Panel B of Figure 1) than the compression effects found in other studies.

Yet, this does not mean that the proliferation of AI will have no effect on gender, educational, or occupational inequalities in the labor market: we do find AI to be more useful for some tasks than others, and to the extent that women and men, high and low educated, or professionals and non-professionals perform on average different types of tasks (Grundke, Marcolin and Squicciarini, 2019; Haslberger, 2021; Cortes, Jaimovich and Siu, 2023), AI may narrow or exacerbate inequalities through compositional effects. Answering who selects into occupations particularly exposed to AI is however beyond the scope of this study.

5 Discussion

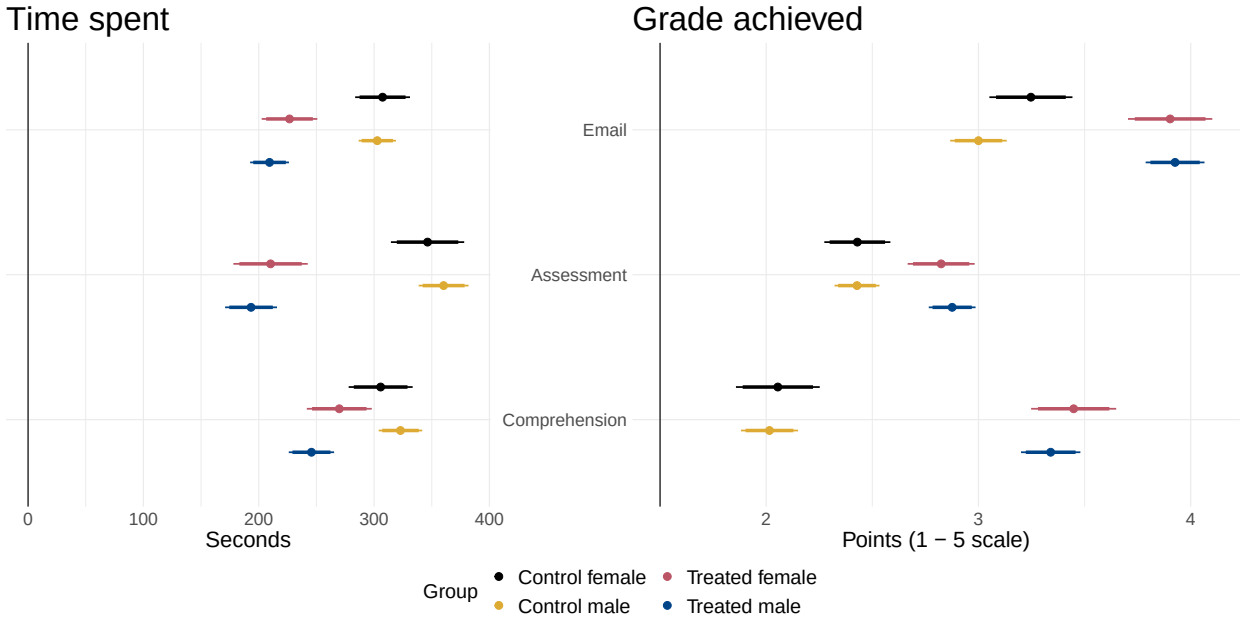
We conducted an innovative survey experiment in which a sample representative of the UK working age population completed three text-based tasks of varying complexity and ambiguity, with respondents randomised into treatments that encouraged or discouraged the use of ChatGPT. We found that the productivity gains from exposure to AI are substantial both in terms of time savings and quality improvements, regardless of task complexity. Moreover, and crucially, our findings suggest that generative AI tools are sufficiently intuitive for broad segments of the working age population to benefit from integrating them into their workflows. ChatGPT and similar technologies appear to make workers uniformly more productive in text-related tasks of varying complexity, regardless of their sex, education, or occupational background. The exception are older workers, who may struggle to adapt to the new technological possibilities. These results paint a

Figure 3: Younger workers benefit more from AI exposure



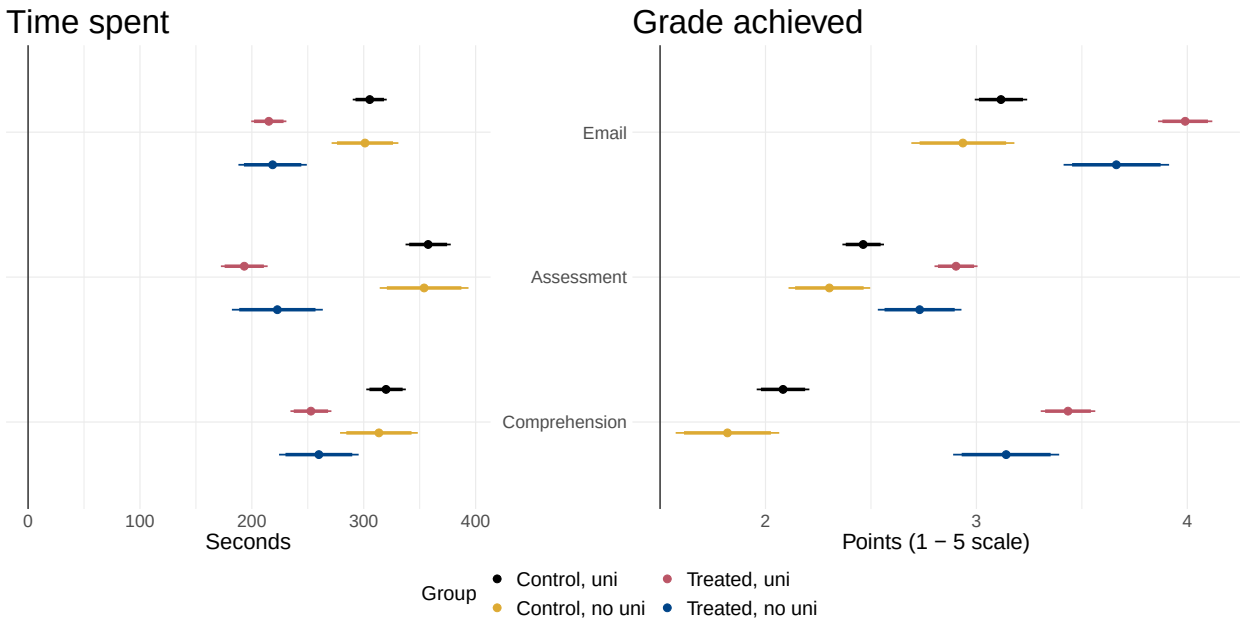
Note: The figure shows predicted values calculated from linear probability models interacting the treatment dummy with a categorical age indicator, with 90% and 95% confidence intervals (thick and thin lines). N = 1,041. Full model output in Table E3.

Figure 4: No sex differences in the effect of AI



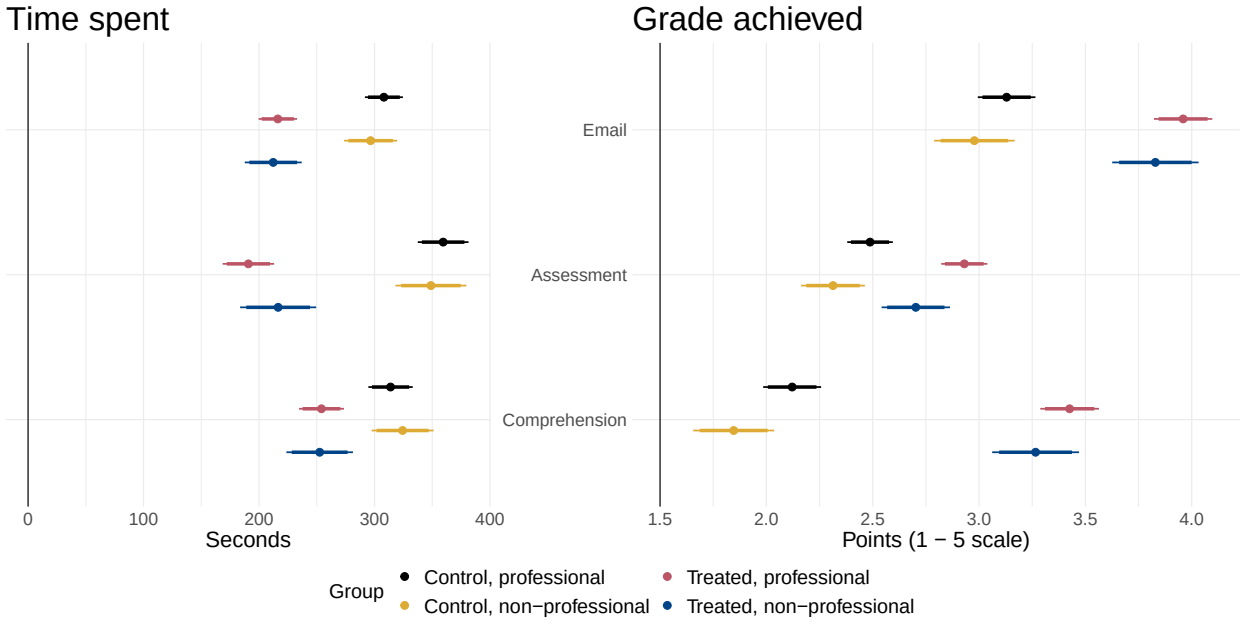
Note: The figure shows predicted values calculated from linear probability models interacting the treatment dummy with a dummy for female respondents, with 90% and 95% confidence intervals (thick and thin lines). N = 1,041. Full model output in Table E3.

Figure 5: Respondents with a university degree benefit as much as those without



Note: The figure shows predicted values calculated from linear probability models interacting the treatment dummy with a dummy for respondents with a university degree, with 90% and 95% confidence intervals (thick and thin lines). N = 1,027. Full model output in Table E3.

Figure 6: Professionals do not monopolise the benefits of AI



Note: The figure shows predicted values calculated from linear probability models interacting the treatment dummy with a dummy for respondents who work in a professional or managerial occupation, with 90% and 95% confidence intervals (thick and thin lines). N = 1,041. Full model output in Table E3.

nuanced picture of the prospective impacts of AI on the labor market at large.

All in all, our results caution against expectations that AI will be an equalizing force. In contrast to studies of specific occupations, we find little evidence that AI reduces aggregate inequalities in productivity across different socio-economic or other demographic groups. Moreover, older individuals appear to be particularly hesitant to use generative AI even when they are encouraged to do so and are aware of the benefits. New inequalities between groups appear most likely to arise from compositional effects. The centrality of text-based tasks such as the ones employed in our study varies across occupations, as does the representation of different demographic groups. It is therefore crucial for further research to determine more systematically in which kinds of tasks generative AI can be most effectively deployed.

Our results provide tentative evidence for such a taxonomy. We identify two axes along which tasks may be placed: complexity and ambiguity. AI appears to be most useful in tasks that are complex yet exhibit low ambiguity. However, this is an emergent perspective that requires further thought and empirical investigation – we did not design the present survey with this taxonomy in mind. Moreover, generative AI is a rapidly evolving technology and future iterations may change its set of capabilities. Yet, mapping and systematizing the "jagged frontier" (Dell'Acqua et al., 2023) of AI capabilities is a crucial requirement for studying its labor market impacts.

Our study contributes to an emerging literature on the labor market consequences of generative AI. It complements important existing work by Noy and Zhang (2023), Dell'Acqua et al. (2023), and Peng et al. (2023) by zooming out and studying the impact of generative AI on a representative sample of the working age population. The time and methodological constraints inherent to our online survey design contained certain trade-offs. We did not investigate whether AI training affects people's competence in using AI, as Dell'Acqua et al. (2023) suggest. Furthermore, unlike Noy and Zhang (2023) and Dell'Acqua et al. (2023), we could not collect a baseline measure of task-specific skills which would have allowed for within-individual comparisons.

While our study provides strong evidence that generative AI enables workers of all types to be significantly more productive in a range of text-related tasks, the impact on aggregate employment will depend on the regulatory environment, the reorganization of tasks and workplaces, changes in demand for the products and services made more affordable by AI, and not least on the acceptance by the broader public of services and decisions rendered by algorithms (Raviv, 2023). Nonetheless, our findings challenge the emerging consensus about the equalizing effects of generative AI, and underscore the importance of considering its effects across the labor market at large.

References

Acemoglu, Daron and David Autor. 2011. Skills, tasks and technologies: Implications for employment and earnings. In Handbook of Labor Economics. Vol. 4b Elsevier B.V. pp. 1043–1171. arXiv: 1011.1669v3 ISSN: 15734463.

Agrawal, Ajay, Joshua S. Gans and Avi Goldfarb. 2019. "Artificial Intelligence: The Ambiguous Labor Market Impact of Automating Prediction." Journal of Economic Perspectives 33(2):31–50.
URL: *https://pubs.aeaweb.org/doi/10.1257/jep.33.2.31*

Albanesi, Stefania, Antonio Dias da Silva, Juan F. Jimeno, Ana Lamo and Alena Wabitsch. 2023. "New Technologies and Jobs in Europe." NBER Working Paper Series (31357).

Autor, David. 2015. "Why Are There Still So Many Jobs? The History and Future of Workplace Automation." Journal of Economic Perspectives 29(3):3–30. ISBN: 9781589017009.
URL: *http://pubs.aeaweb.org/doi/10.1257/jep.29.3.3*

Autor, David. 2022. "The Labor Market Impacts of Technological Change: From Unbridled Enthusiasm to Qualified Optimism to Vast Uncertainty." NBER Working Paper Series (30074).

Autor, David, David Dorn and Gordon H. Hanson. 2013. "The China Syndrome: Local Labor Market Effects of Import Competition in the United States." American Economic Review 103(6):2121–2168. arXiv: 1011.1669v3 ISBN: 0002-8282.

Autor, David, David Dorn, Gordon Hanson and Kaveh Majlesi. 2020. "Importing political polarization? The electoral consequences of rising trade exposure." American Economic Review 110(10):3139–3183.

Autor, David, Frank Levy and Richard J. Murnane. 2003. "The Skill Content of Recent Technological Change: An Empirical Exploration." The Quarterly Journal of Economics 118(4):1279–1333.

Berglund, Lukas, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak and Owain Evans. 2023. "The Reversal Curse: LLMs trained on "A is B" fail to learn "B is A".". arXiv:2309.12288 [cs].
URL: *http://arxiv.org/abs/2309.12288*

Brynjolfsson, Erik, Danielle Li and Lindsey R. Raymond. 2023. "Generative AI at Work." NBER Working Paper Series (31161).

Bubeck, Sébastien, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro and Yi

Zhang. 2023. "Sparks of Artificial General Intelligence: Early experiments with GPT-4.". arXiv:2303.12712 [cs].
URL: *http://arxiv.org/abs/2303.12712*

Cain Miller, Claire and Courtney Cox. 2023. "In Reversal Because of A.I., Office Jobs Are Now More at Risk." The New York Times .
URL: *https://www.nytimes.com/2023/08/24/upshot/artificial-intelligence-jobs.html*

Chen, Lingjiao, Matei Zaharia and James Zou. 2023. "How is ChatGPT's behavior changing over time?". arXiv:2307.09009 [cs].
URL: *http://arxiv.org/abs/2307.09009*

Cortes, Guido Matias, Nir Jaimovich and Henry E. Siu. 2023. "The Growing Importance of Social Tasks in High-Paying Occupations: Implications for Sorting." Journal of Human Resources 58(5):1429–1451.
URL: *http://jhr.uwpress.org/lookup/doi/10.3368/jhr.58.5.0121-11455R1*

del Rio-Chanona, Maria, Nadzeya Laurentsyeva and Johannes Wachs. 2023. "Are Large Language Models a Threat to Digital Public Goods? Evidence from Activity on Stack Overflow.". arXiv:2307.07367 [cs].
URL: *http://arxiv.org/abs/2307.07367*

Dell'Acqua, Fabrizio, Edward McFowland Iii, Ethan Mollick, Hila Lifshitz-Assaf, Katherine C Kellogg, Saran Rajendran, Lisa Krayer, François Candelon and Karim R Lakhani. 2023. "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality." SSRN Electronic Journal .

Eloundou, Tyna, Sam Manning, Pamela Mishkin and Daniel Rock. 2023. "GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models.". arXiv:2303.10130 [cs, econ, q-fin].
URL: *http://arxiv.org/abs/2303.10130*

Felten, Edward W., Manav Raj and Robert Seamans. 2023. "Occupational Heterogeneity in Exposure to Generative AI." SSRN Electronic Journal .
URL: *https://www.ssrn.com/abstract=4414065*

Frey, Carl Benedikt and Michael Osborne. 2023. "Generative AI and the Future of Work: A Reappraisal." Brown Journal of World Affairs .

Gilardi, Fabrizio, Meysam Alizadeh and Maël Kubli. 2023. "ChatGPT outperforms crowd workers for text-annotation tasks." Proceedings of the National Academy of Sciences 120(30):e2305016120.
URL: *https://pnas.org/doi/10.1073/pnas.2305016120*

Grundke, Robert, Luca Marcolin and Mariagrazia Squicciarini. 2019. Narrowing the gender wage gap in the digital era: the role of skills. In Taking Stock: Data and Evidence on Gender Equality in Digital Access, Skills, and Leadership, ed. Araba Sey and Nancy Hafkin. Macau: United Nations University.

Haslberger, Matthias. 2021. "Rethinking the measurement of occupational task content." The Economic and Labour Relations Review 33(1):178–199.
URL: *https://doi.org/10.1177/10353046211037095*

Lane, Marguerita and Anne Saint-Martin. 2021. "The impact of Artificial Intelligence on the labour market: What do we know so far?" OECD Social, Employment and Migration Working Papers (256).
URL: *https://dx.doi.org/10.1787/7c895724-en*

Law School Admission Council. 2007. "The Official LSAT Preptest.". Form 8LSN75.

McCoy, R. Thomas, Shunyu Yao, Dan Friedman, Matthew Hardy and Thomas L. Griffiths. 2023. "Embers of Autoregression: Understanding Large Language Models Through the Problem They are Trained to Solve.". arXiv:2309.13638 [cs].
URL: *http://arxiv.org/abs/2309.13638*

Noy, Shakked and Whitney Zhang. 2023. "Experimental evidence on the productivity effects of generative artificial intelligence." Science (381):187–192.

OpenAI. 2023. "GPT-4 Technical Report.". arXiv:2303.08774 [cs].
URL: *http://arxiv.org/abs/2303.08774*

Peng, Sida, Eirini Kalliamvakou, Peter Cihon and Mert Demirer. 2023. "The Impact of AI on Developer Productivity: Evidence from GitHub Copilot.". arXiv:2302.06590 [cs].
URL: *http://arxiv.org/abs/2302.06590*

Raviv, Shir. 2023. "When Do Citizens Resist The Use of Algorithmic Decision-making in Public Policy? Theory and Evidence." SSRN Electronic Journal .
URL: *https://www.ssrn.com/abstract=4328400*

Wood, Robert E. 1986. "Task complexity: Definition of the construct." Organizational Behavior and Human Decision Processes 37(1):60–82.
URL: *https://linkinghub.elsevier.com/retrieve/pii/0749597886900440*

Appendix: For Online Publication

Contents

A Additional Details About the Survey

To ensure the feasibility of the study, YouGov asked 34,211 people between 14th April and 4th May 2023 whether they have a ChatGPT account. The study sample is recruited from the 5,350 respondents who stated that they have an account (free or paid). Below we list the feasibility question and the distribution of answers.

- ChatGPT is an AI-based computer program that can generate human-like text. Before taking this survey, had you ever used ChatGPT?
  - Yes, and I have a paid account (n=865)
  - Yes, and I have a free account (n=4485)
  - Yes, but I used someone else's account (n=774)
  - No, I have never used ChatGPT (n=28087)

While we acknowledge that this introduces some selection issues, we employ quotas for age, gender, income, and region to ensure a broadly representative sample. Given the user pool at the time, we did not manage to obtain a fully representative sample. As Table A1 shows, the sample is more male and more educated than our target population, the UK working age population. However, Table A2 shows good balance between the treatment and control groups. Moreover, any residual bias is likely to be conservative, as participants in the control group already have an account and could use ChatGPT with little extra effort.

Table A1: Sample characteristics

Characteristic	Count/Mean (%/SD) Treatment	Count/Mean (%/SD) Control
Sex: Female	165 (32.7%)	170 (31.7%)
Sex: Male	339 (67.3%)	367 (68.3%)
Uni: Yes	393 (79.1%)	421 (79.4%)
Uni: No	104 (20.9%)	109 (20.6%)
Professional: Yes	346 (68.7%)	355 (66.1%)
Professional: No	158 (31.4%)	182 (33.9%)
Age	40.5 (11.2)	39.8 (11.3)
HH Income	10.9 (3.2)	10.6 (3.4)
Skills Average	1.9 (0.5)	2.0 (0.6)

We pre-tested a version of the survey which included a control group in which we did not prime participants about AI at all. Since answer quality in this group was generally lower and a substantial share of respondents used ChatGPT, we did not include this control group in the final version of the survey.

Table A2: Balance Tests

Variable name	Type	Std. Mean Dif.
age	C	0.0616
sex: Male	B	-0.0108
degree	B	-0.0036
degree: NA	B	0.0009
prof_man	B	0.0254
hh_income	C	0.1017*
hh_income: NA	B	-0.0116
skills_avg	C	-0.0460
skills_avg: NA	B	0.0010
education: GCSEs / O-Levels or none	B	0.0046
education: A-Levels or equivalent	B	-0.0010
education: Undergraduate	B	-0.0216
education: Postgraduate	B	0.0180

Note: B = binary variable; C = continuous variable. Sample sizes: Control: 537; Treatment: 504.

B Deviations from the Pre-Analysis Plan

Some changes were made compared to the procedures outlined in the pre-analysis plan (PAP). Here we detail these changes and the reason why they were necessary.

- **Small changes to the grading scheme:** Some very minor changes to the grading scheme were necessary to improve the workflow. The full grading scheme as it was used in the final analysis is attached in Appendix G.
- **Changes to categories of moderator variables:** We split the age variable into three categories, rather than two as specified in the PAP.
- **Verification of compliance:** Instead of manually inspecting answers to identify misclassifications of ChatGPT use, we verified that compliers in the control group do not perform significantly different from non-compliers in the treatment group. As described in Appendix C, this analysis confirms that respondents' self-reports are largely accurate.

C Compliance with Treatment Instructions

Respondents in the treatment group may refrain from using ChatGPT, while respondents in the control group may defy instructions and use AI even when they should not. We therefore designed our financial incentives to increase compliance with the instructions to use or not use ChatGPT. We highlighted the prospect of additional rewards in case of the treatment group and the threat of being excluded from the survey for non-compliance in case of the control group. In addition to visually inspecting the answers, we used Turnitin detection software to identify AI use in the control group and ask respondents after each task to self-report which, if any, online tools they used. It is important to note that the treatment effects presented in the main text thus represent intention-to-treat (ITT) effects, since we do not have full control over compliance in either the treatment or control group. However, our diagnostic measures indicate high compliance. Moreover, any residual bias is likely to be conservative, as undetected AI use in the control group is likely to narrow performance gaps.

Self-Reported Compliance and Treatment Effects

Respondents' self-reported use of AI arguably allows us to capture an imperfect measure of the ATE. For this, we asked participants after every task whether they used AI. People generally report high compliance with the instructions: well over 90% in the control group state for each task that they did not use AI, and between two thirds and three quarters of respondents in the treatment group report that they did (see Table D2). We call these respondents compliers. In Figure C1 we plot the predicted values of time spent and grade achieved by treatment and compliance status. The overall increase in performance in the treatment group is entirely driven by the compliers. Non-compliers in the treatment group perform very similar to compliers in the control group, implying that self-reports of AI use are generally truthful.[12] [13]

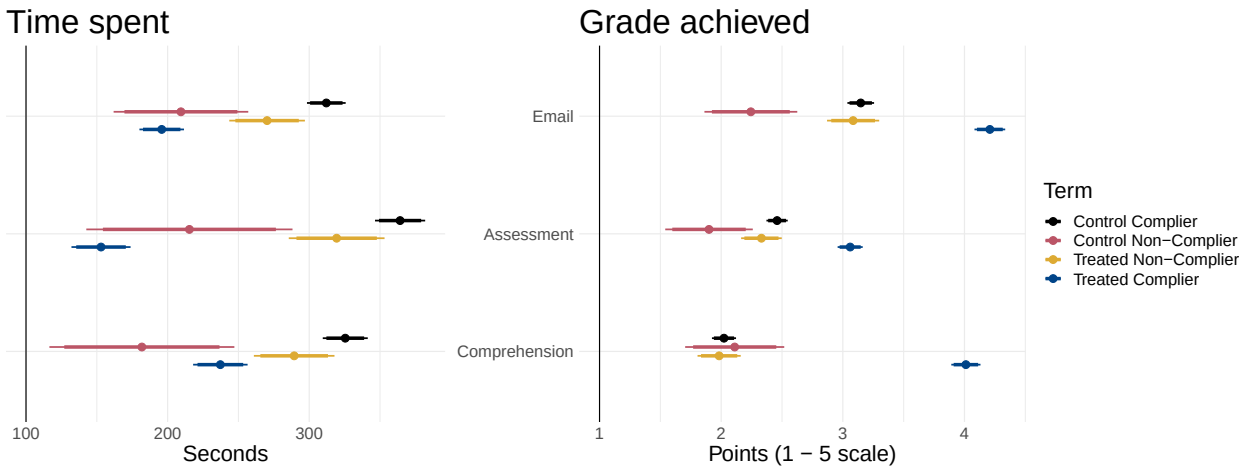
Importantly, this analysis can only approximate the ATE, since selection into compliance is not random. For example, compliance in the control group increases with age, while it decreases with age in the treatment group (see Table D3). For this reason, in the remainder of the paper we analyse ITT effects, which can be interpreted as a lower-bound estimate of the true effect of AI use, while excluding non-compliers would yield an upper-bound estimate.[14] Nevertheless, the above analysis illustrates, firstly, that self-reports of AI use are generally reliable, and secondly, that the estimated treatment effects are not due to some confounding

[12] Based on the reasonable assumption that people in the treatment group have no incentive to underreport their use of AI, we can compare compliers in the control group and non-compliers in the treatment group. If there was widespread underreporting of AI use in the control group, we would expect compliers in the control group to exhibit substantially higher performance than non-compliers in the treatment group.

[13] The low performance of the few non-compliers in the control group (between 30 and 41 individuals) is likely due to a number of inattentive respondents who exerted low effort on the tasks and selected random answers to the self-report question. Similarly, non-compliers in the treatment group spend less time and have slightly lower scores than control-compliers, indicating that they are the less motivated participants.

[14] Panel C of Table E2 shows significantly larger treatment effects when we exclude non-compliers.

Figure C1: Predicted values by treatment and compliance status



Note: The figure shows predicted values calculated from linear probability models interacting the treatment dummy with a complier dummy, with 90% and 95% confidence intervals (thick and thin lines). N = 1,041. Full model output in Table E2.

attributes of the treatment group. This further strengthens the evidence that ChatGPT allows people to perform standard work tasks such as editing sensitive emails, appraising large quantities of written information, and extracting information from complicated texts both better and faster.

Turnitin Estimates of AI Use

We used Turnitin's AI detection tool to verify whether respondents in the control group complied with the survey instructions to avoid using AI chatbots to assist with their responses. To do this, we uploaded a random selection of responses from the treatment and control groups for each of the tasks into Turnitin and compared AI detection scores from the two groups. We repeated this exercise six times with a different set of randomly selected responses. We used a random sample of responses rather than the full sample to comply with Turnitin's word limits for AI detection. In accordance with the maximum word limit, we selected 40 random samples for the email task, 100 random samples for the assessment task, and 35 random samples for the comprehension task. Scores range from 0-100 and estimate the percentage of the document that is AI generated.

As expected, the treatment group consistently received higher AI detection scores than the control group for all three tasks, as outlined in Table C1. AI scores are relatively high for the email task for both groups because the original text was generated in part using ChatGPT — nonetheless, there is still a noticeable difference in the mean scores between treatment and control groups on this task.

While we acknowledge that Turnitin's AI detection tool cannot definitively prove the extent to which

Table C1: Comparison of Turnitin AI detection scores by treatment

Sample	**Email** (n=40)		**Assessment** (n=100)		**Comprehension** (n=35)	
	Treat.	Control	Treat.	Control	Treat.	Control
Sample 1	84	79	16	0	38	0
Sample 2	91	75	14	4	20	0
Sample 3	82	71	14	0	28	0
Sample 4	79	81	10	0	36	9
Sample 5	79	74	18	0	38	2
Sample 6	86	86	24	0	25	0
Mean	**83.5**	**77.67**	**16.0**	**0.67**	**30.83**	**1.83**

AI was used by respondents, these results give us more confidence that both groups largely complied with instructions on AI use.

D Descriptive Statistics

Before the first task, we asked respondents to assess their own skills on a number of dimensions: understanding written information, writing informative texts, communicating with others, performing under pressure, using new technologies, and English proficiency. We find no meaningful differences between the treatment and control group on any of the measures. Respondents overwhelmingly described themselves as "somewhat skilled" in all five dimensions, as we show in Figure D1. Furthermore, the vast majority of respondents are native English speakers, as can be seen in Figure D2. After each task, we asked respondents how frequently they perform tasks similar to the one they just did in their daily life. As Figure D3 shows, familiarity with the tasks is very similar in both groups. The levels of familiarity correspond to task complexity, with people being most likely to be familiar with the email task and least likely with the comprehension task. This, alongside the balance tests on demographic characteristics reported in Table A2, provides reassurance that the treatment effects detailed below do not result from different characteristics of the samples.

Table D1: Performance on the tasks by treatment

		Treatment	
Task	**Variable**	**Control**	**Treatment**
Email	Median score	3	5
	Mean score (SD)	3.1 (1.2)	3.9 (1.4)
	Median time	263.2	182.4
Assessment	Median score	2	3
	Mean score (SD)	2.4 (0.9)	2.9 (1.2)
	Median time	287.3	132.4
Comprehension	Median score	1.7	3.7
	Mean score (SD)	2.0 (1.0)	3.4 (1.6)
	Median time	272.3	212.6

One potential objection to our results is that they might be influenced by the type of device respondents used to complete the survey. In particular, people using a mobile device might find it more onerous to access ChatGPT and read the texts that make up the tasks, leading to lower ChatGPT uptake in the treatment group and lower response quality. To address this concern, we recorded the operating system of the device used to complete the survey, based on which we created a dummy for respondents who used a mobile device (approximately 70% of the sample). In Table D2 we show that mobile users are evenly distributed across treatment groups. Self-reported AI uptake indeed tends to be slightly lower among users of mobile devices, by about 10 percentage points in the control group and by up to 7 percentage points in the treatment group. Nevertheless, performance, both in terms of grades and time spent on the tasks, is very similar for respondents who used a mobile device and those who used a laptop or desktop computer, with the possible exception of the comprehension task. Thus, the patterns documented above are not driven by differences in

the type of devices used.

Table D2: Compliance by treatment group and device type

Task		Control group		Treatment group	
		Mobile device	Other device	Mobile device	Other device
Email	AI: No	367 (95%)	129 (85%)	88 (25%)	42 (27%)
	AI: Yes	19 (5%)	22 (15%)	258 (75%)	116 (73%)
	Mean score (SD)	3.1 (1.2)	3.1 (1.2)	3.9 (1.4)	3.9 (1.4)
Assessment	AI: No	375 (97%)	132 (87%)	103 (30%)	36 (23%)
	AI: Yes	11 (3%)	19 (13%)	243 (70%)	122 (77%)
	Mean score (SD)	2.4 (0.9)	2.4 (0.8)	2.8 (1.2)	2.9 (1.2)
Comprehension	AI: No	377 (98%)	130 (86%)	115 (33%)	43 (27%)
	AI: Yes	9 (2%)	21 (14%)	231 (67%)	115 (73%)
	Mean score (SD)	1.9 (1.0)	2.3 (1.1)	3.3 (1.6)	3.6 (1.5)

Note: Cross-classification of treatment group by type of device used. The share of mobile device users is slightly higher in the control group, as is compliance with the instructions. Within-treatment group performance is not affected by type of device used, except in the comprehension task.

Table D3: Compliance by treatment group and age

Task	Age	Control group		Treatment group	
		Complier	Non-complier	Complier	Non-complier
Email	15 - 35	186 (88%)	26 (12%)	159 (84%)	31 (16%)
	36 - 50	219 (95%)	12 (5%)	163 (75%)	53 (25%)
	51+	91 (97%)	3 (3%)	52 (53%)	46 (47%)
Assessment	15 - 35	194 (92%)	18 (8%)	153 (81%)	37 (19%)
	36 - 50	221 (96%)	10 (4%)	154 (71%)	62 (29%)
	51+	92 (98%)	2 (2%)	58 (59%)	40 (41%)
Comprehension	15 - 35	194 (92%)	18 (8%)	153 (81%)	37 (19%)
	36 - 50	220 (95%)	11 (5%)	144 (67%)	72 (33%)
	51+	93 (99%)	1 (1%)	49 (50%)	49 (50%)

Note: Cross-classification of age by treatment group and compliance status. Compliance increases with age in the control group and decreases with age in the treatment group.

Figure D1: Skill self-assessments

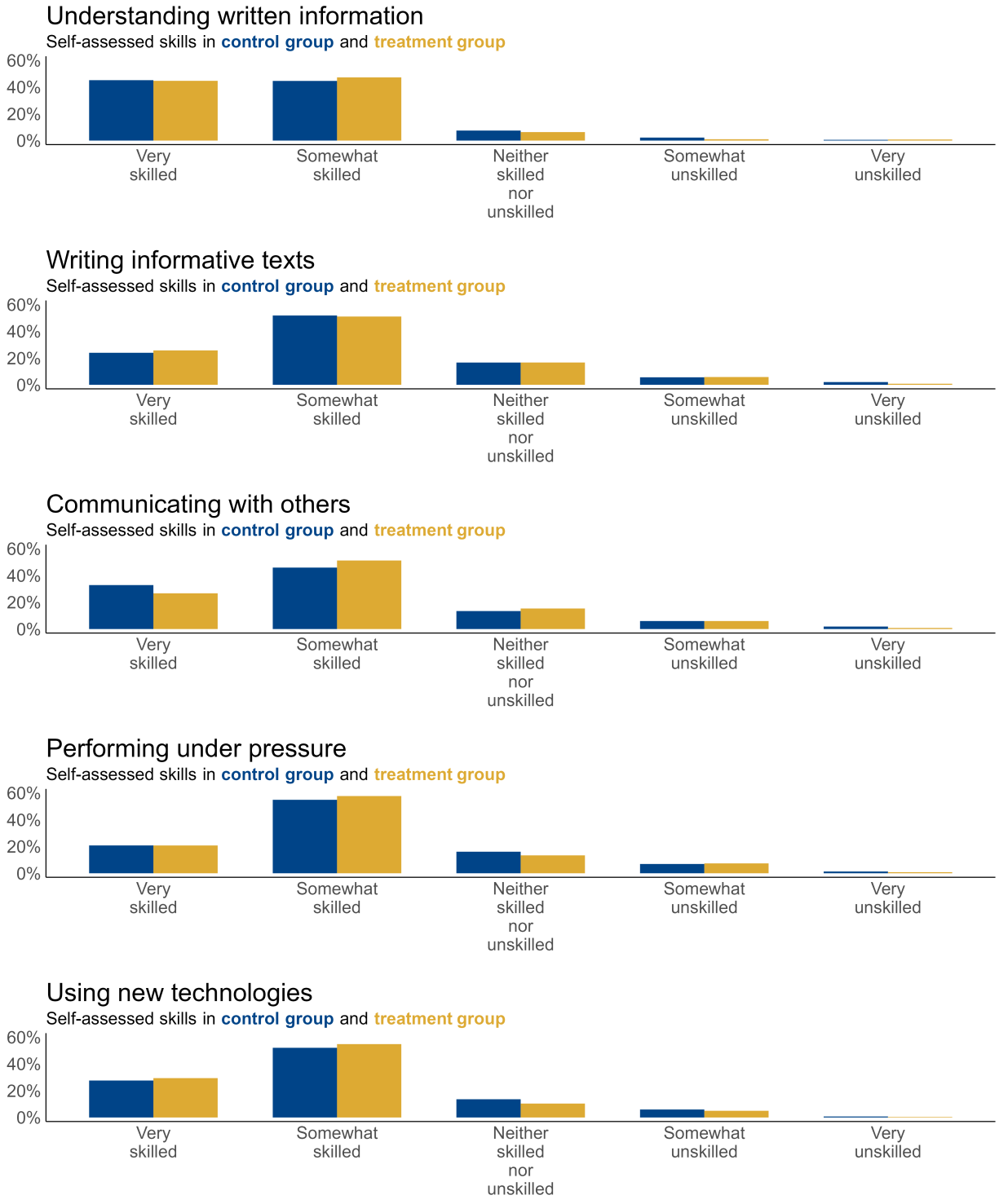
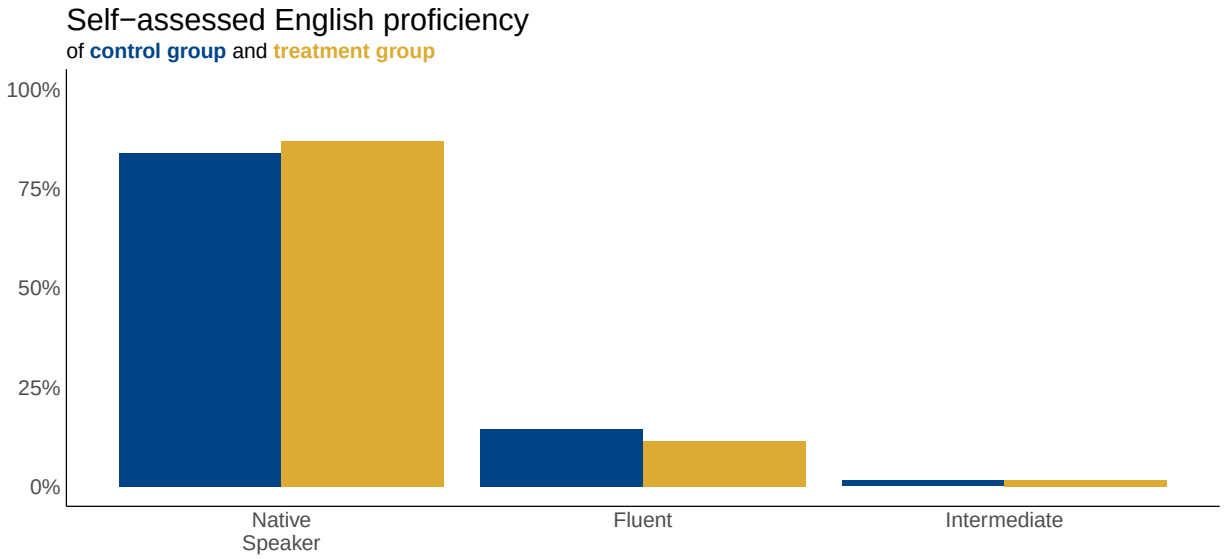


Figure D2: English proficiency



E Supplementary Results

Table E1: Within-occupational group variation in performance

Occ. type	Task	CV Treated	CV Control	F-statistic	p-value	Emp. share
Professionals	Email	34.072	36.407	1.142	0.171	39.7
	Assess.	40.196	33.423	1.446	0.004	
	Comp.	44.391	48.751	1.206	0.090	
Managers	Email	37.391	37.462	1.003	0.492	27.7
	Assess.	40.317	37.156	1.177	0.164	
	Comp.	46.790	51.613	1.217	0.123	
Clerical & Service	Email	31.771	39.273	1.528	0.013	22.6
	Assess.	40.574	36.275	1.251	0.113	
	Comp.	44.325	47.075	1.128	0.263	
Manual	Email	60.384	46.716	1.671	0.100	6.4
	Assess.	56.921	39.038	2.126	0.030	
	Comp.	69.474	44.183	2.473	0.012	
Other	Email	25.590	46.810	3.346	0.009	3.7
	Assess.	31.180	32.856	1.110	0.421	
	Comp.	44.525	52.980	1.416	0.242	

Note: Coefficient of variation of task quality, including all respondents. AI tends to reduce variation in performance in all groups except manual workers in the email and comprehension tasks, and increases variation for all groups except "Other" in the assessment task.

Figure D3: Frequency with which people perform similar tasks

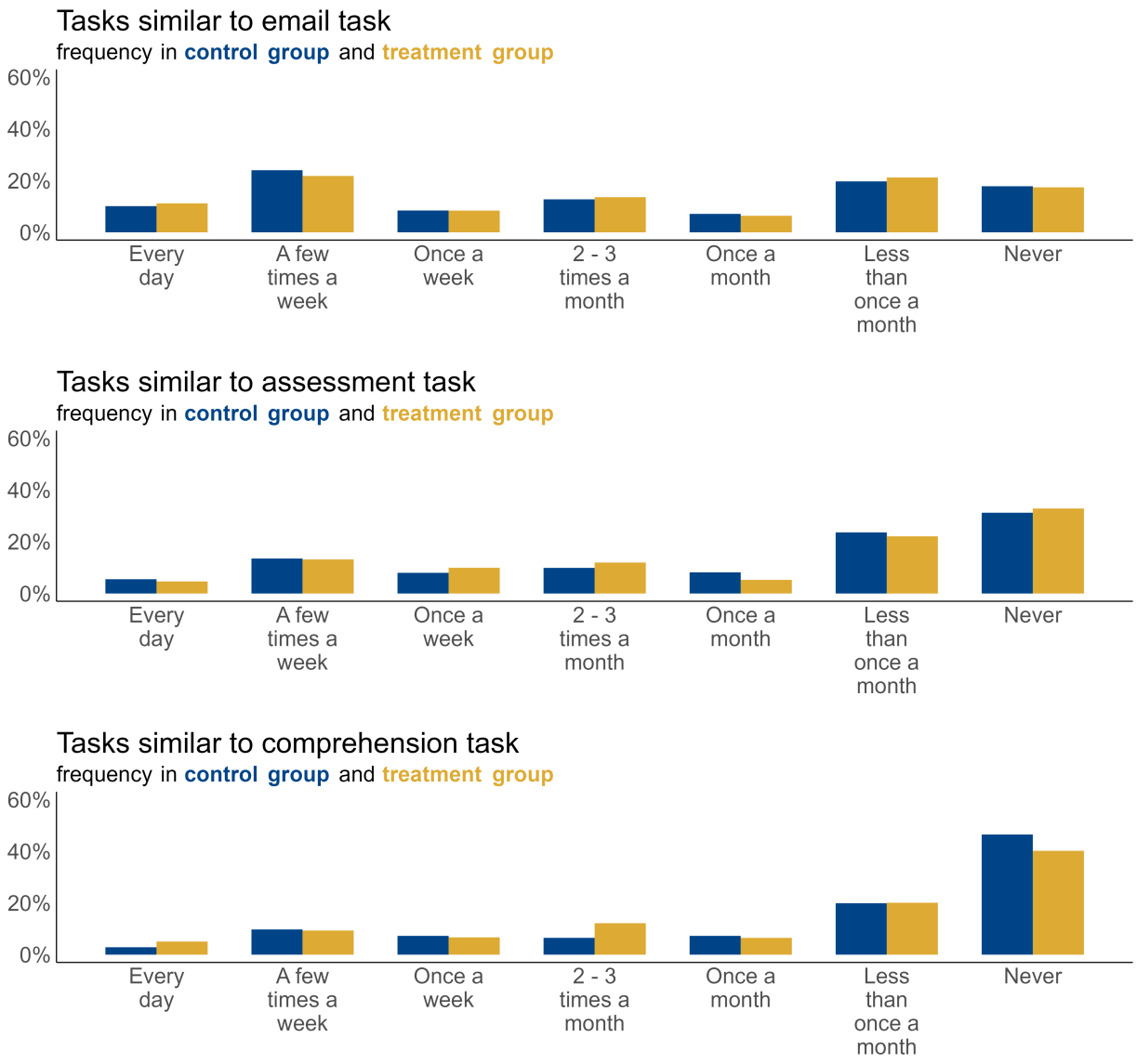


Figure D4: Density plots of performance by treatment group

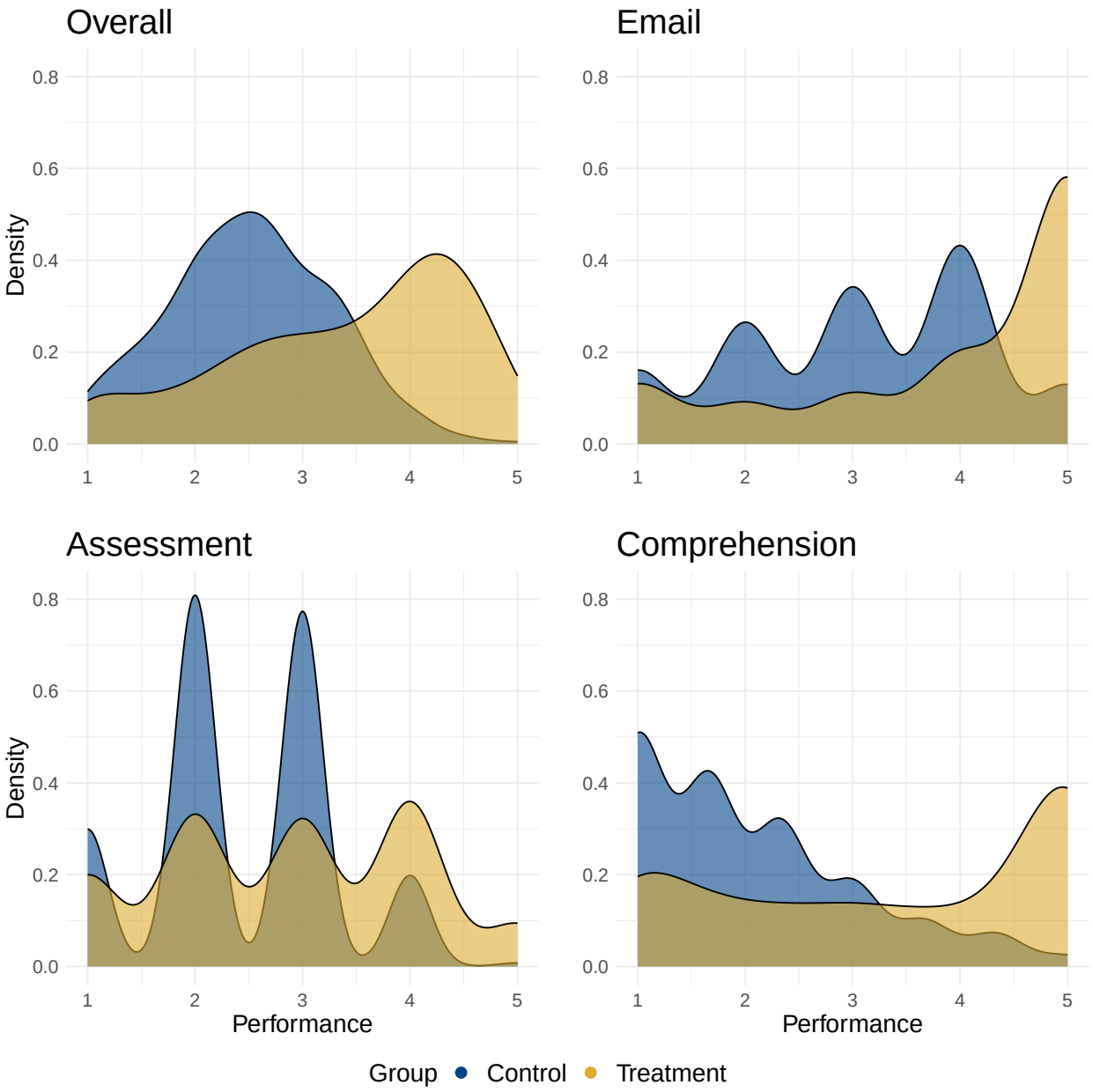


Table E2: Model results for Figures 2 and C1

	Dependent variable:					
	Email Time	Ass. Time	Comp.Time	Email Score	Ass. Score	Comp. Score
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Unconditional treatment effects						
Treatment	−89.212***	−156.995***	−63.840***	0.840***	0.431***	1.348***
	(9.790)	(13.084)	(11.439)	(0.081)	(0.064)	(0.081)
Constant	304.234***	355.858***	317.427***	3.078***	2.428***	2.028***
	(6.812)	(9.104)	(7.959)	(0.056)	(0.045)	(0.057)
Observations	1,041	1,041	1,041	1,041	1,041	1,041
R^2	0.074	0.122	0.029	0.094	0.042	0.209
Panel B: Interaction treatment x compliance						
Treatment	−34.870**	−44.803**	−36.078**	−0.082	−0.129	−0.040
	(15.783)	(19.457)	(16.611)	(0.127)	(0.096)	(0.104)
AI Use	−118.234***	−148.783***	−143.619***	−0.954***	−0.560***	0.088
	(24.461)	(38.186)	(34.256)	(0.197)	(0.189)	(0.214)
Treatment x AI Use	37.150	−17.611	91.493**	2.032***	1.289***	1.940***
	(29.381)	(43.225)	(38.470)	(0.237)	(0.214)	(0.240)
Constant	306.484***	364.170***	325.450***	3.146***	2.460***	2.023***
	(7.021)	(9.026)	(8.097)	(0.057)	(0.045)	(0.050)
Observations	1,041	1,041	1,041	1,041	1,041	1,041
R^2	0.109	0.187	0.053	0.160	0.096	0.407
Panel C: Unconditional treatment effects excluding non-compliers						
Treatment	−116.245***	−211.197***	−88.204***	1.061***	0.601***	1.989***
	(10.466)	(13.946)	(12.315)	(0.084)	(0.069)	(0.080)
Constant	312.072***	364.170***	325.450***	3.147***	2.460***	2.023***
	(6.862)	(9.023)	(7.843)	(0.055)	(0.045)	(0.051)
Observations	870	872	853	870	872	853
R^2	0.124	0.209	0.057	0.155	0.080	0.418

Note: Figures 2 and C1 are created based on the models in Panels A and B. Panel C shows that excluding non-compliers increases effect sizes. Note that the interaction coefficients in Panel B can be interpreted as the difference in differences of the treatment effect on compliers and non-compliers. *p<0.1; **p<0.05; ***p<0.01.

F Time Spent and Performance

In general, our results show that respondents who use ChatGPT perform better even as they take significantly less time to complete the tasks, in line with H1b. As Table D1 shows, on each of the tasks, the median respondent in the treatment group spends at least one minute less than the median respondent in the control group. Summing up the time spent on all three tasks, the median difference amounts to approximately five minutes.[15] None of this is surprising, however, Figure F1 also shows differences in the relationship between time spent a task and answer quality between the two groups: while spending more time on a task is, up to about 10 minutes, associated with a higher-scoring answer in the control group, the relationship between time spent and answer quality starts out or turns negative earlier in the treatment group. This may indicate two things: first, people who take longer to get an answer from ChatGPT may be less familiar with the tool and struggle to use it effectively. The second possibility is that people who try to improve on ChatGPT

[15]However, this does not translate into much longer overall response times for the non-treated respondents, whose median time for completing the survey is only approximately 57 seconds longer than for the treatment group. This indicates that some respondents in the control group compensate for the extra effort required on the tasks by speeding through the remainder of the survey.

Table E3: Model results for Figures 3 - 6 (heterogeneous treatment effects)

	Dependent variable:					
	Email Time	Ass. Time	Comp.Time	Email Score	Ass. Score	Comp. Score
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	−90.147***	−145.297***	−69.409***	1.022***	0.466***	1.468***
	(15.687)	(20.798)	(18.336)	(0.130)	(0.103)	(0.130)
Age 36-50	−9.789	4.196	3.608	0.026	−0.151	−0.171
	(14.935)	(19.801)	(17.456)	(0.124)	(0.098)	(0.124)
Age 51+	58.833***	110.745***	53.484**	0.060	−0.173	−0.157
	(19.458)	(25.798)	(22.743)	(0.161)	(0.128)	(0.161)
Treatment x Age 36-50	14.515	−14.185	6.257	−0.239	−0.019	−0.046
	(21.609)	(28.651)	(25.258)	(0.179)	(0.142)	(0.179)
Treatment x Age 51+	−33.137	−39.910	9.544	−0.415*	−0.123	−0.500**
	(27.568)	(36.550)	(32.223)	(0.228)	(0.181)	(0.229)
Constant	298.146***	334.667***	306.513***	3.057***	2.524***	2.129***
	(10.784)	(14.298)	(12.605)	(0.089)	(0.071)	(0.089)
Observations	1,041	1,041	1,041	1,041	1,041	1,041
R^2	0.087	0.148	0.042	0.099	0.050	0.223
Treatment	−93.433***	−166.938***	−77.178***	0.926***	0.448***	1.325***
	(11.895)	(15.898)	(13.888)	(0.098)	(0.078)	(0.099)
Female	4.576	−13.900	−17.172	0.247**	0.002	0.039
	(14.649)	(19.580)	(17.104)	(0.121)	(0.096)	(0.122)
Treatment x Female	12.741	30.833	41.309*	−0.270	−0.053	0.069
	(20.959)	(28.013)	(24.471)	(0.173)	(0.137)	(0.174)
Constant	302.785***	360.258***	322.863***	3.000***	2.428***	2.015***
	(8.242)	(11.017)	(9.624)	(0.068)	(0.054)	(0.068)
Observations	1,041	1,041	1,041	1,041	1,041	1,041
R^2	0.075	0.123	0.032	0.098	0.042	0.210
Treatment	−82.562***	−131.173***	−53.678**	0.728***	0.428***	1.321***
	(21.767)	(28.944)	(25.337)	(0.178)	(0.141)	(0.179)
Uni	4.230	3.598	6.371	0.181	0.160	0.264*
	(17.066)	(22.692)	(19.864)	(0.140)	(0.111)	(0.140)
Treatment x Uni	−7.644	−33.169	−13.461	0.146	0.012	0.030
	(24.451)	(32.513)	(28.461)	(0.200)	(0.159)	(0.201)
Constant	301.123***	353.965***	313.569***	2.936***	2.303***	1.820***
	(15.210)	(20.225)	(17.704)	(0.125)	(0.099)	(0.125)
Observations	1,027	1,027	1,027	1,027	1,027	1,027
R^2	0.073	0.124	0.030	0.101	0.047	0.214
Treatment	−84.341***	−132.305***	−71.796***	0.851***	0.389***	1.420***
	(17.176)	(22.941)	(20.071)	(0.141)	(0.112)	(0.142)
Professional	11.502	10.584	−10.424	0.152	0.174*	0.275**
	(14.401)	(19.234)	(16.828)	(0.119)	(0.094)	(0.119)
Treatment x Professional	−7.522	−36.356	11.975	−0.021	0.054	−0.115
	(20.914)	(27.934)	(24.440)	(0.172)	(0.136)	(0.173)
Constant	296.630***	348.861***	324.318***	2.978***	2.313***	1.846***
	(11.709)	(15.639)	(13.683)	(0.096)	(0.076)	(0.097)
Observations	1,041	1,041	1,041	1,041	1,041	1,041
R^2	0.075	0.123	0.029	0.097	0.050	0.214

Note: Figures 3, 4, 5, and 6 are created based on these models. Note that the interaction coefficients can be interpreted as the difference in differences of the treatment effect on the respective groups. *p<0.1; **p<0.05; ***p<0.01.

output often end up making things worse. It may also be that those who take longer in the treatment group are disproportionally non-compliers. For people in the treatment group, there is an optimal amount of time (around 600 seconds or 10 minutes for each task), either side of which performance tends to be lower.[16]

[16]The precise form of this finding is somewhat sensitive to where we top-code the answer times.

Table E4: Main analyses with weights

	Dependent variable:					
	Email Time	Ass. Time	Comp.Time	Email Score	Ass. Score	Comp. Score
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	−86.298***	−156.017***	−65.938***	0.810***	0.446***	1.424***
	(9.925)	(13.113)	(11.491)	(0.082)	(0.064)	(0.082)
Constant	296.744***	347.756***	312.276***	3.068***	2.414***	1.988***
	(6.873)	(9.081)	(7.958)	(0.056)	(0.044)	(0.057)
Observations	1,041	1,041	1,041	1,041	1,041	1,041
R^2	0.068	0.120	0.031	0.087	0.045	0.226
Treatment	−85.904***	−149.777***	−70.437***	0.907***	0.463***	1.519***
	(14.608)	(19.212)	(16.939)	(0.121)	(0.095)	(0.120)
Age 36-50	−6.618	−1.284	−2.983	0.009	−0.125	−0.141
	(14.622)	(19.229)	(16.955)	(0.121)	(0.095)	(0.120)
Age 51+	69.398***	108.918***	53.797**	0.004	−0.176	−0.111
	(22.087)	(29.047)	(25.611)	(0.182)	(0.143)	(0.182)
Treatment x Age 36-50	12.276	−5.881	8.499	−0.112	−0.0005	−0.041
	(21.207)	(27.890)	(24.591)	(0.175)	(0.137)	(0.175)
Treatment x Age 51+	−47.964	−40.448	1.078	−0.373	−0.102	−0.561**
	(31.179)	(41.004)	(36.154)	(0.257)	(0.202)	(0.257)
Constant	291.089***	335.107***	307.003***	3.063***	2.488***	2.060***
	(10.044)	(13.208)	(11.646)	(0.083)	(0.065)	(0.083)
Observations	1,041	1,041	1,041	1,041	1,041	1,041
R^2	0.079	0.139	0.040	0.090	0.051	0.237
Treatment	−88.236***	−159.058***	−72.931***	0.863***	0.459***	1.406***
	(11.621)	(15.365)	(13.460)	(0.096)	(0.075)	(0.096)
Female	12.215	2.567	−5.273	0.164	0.057	0.130
	(15.595)	(20.620)	(18.063)	(0.128)	(0.101)	(0.129)
Treatment x Female	6.441	10.875	25.519	−0.198	−0.050	0.059
	(22.347)	(29.546)	(25.883)	(0.184)	(0.144)	(0.184)
Constant	293.520***	347.079***	313.668***	3.024***	2.399***	1.954***
	(8.011)	(10.592)	(9.279)	(0.066)	(0.052)	(0.066)
Observations	1,041	1,041	1,041	1,041	1,041	1,041
R^2	0.070	0.120	0.032	0.088	0.045	0.228
Treatment	−66.839***	−106.135***	−37.459*	0.683***	0.518***	1.503***
	(19.235)	(25.259)	(22.151)	(0.157)	(0.123)	(0.156)
Uni	15.847	25.856	27.676	0.152	0.179*	0.412***
	(15.243)	(20.017)	(17.554)	(0.124)	(0.097)	(0.124)
Treatment x Uni	−25.873	−68.680**	−39.357	0.174	−0.090	−0.117
	(22.569)	(29.637)	(25.990)	(0.184)	(0.144)	(0.183)
Constant	285.749***	330.370***	293.855***	2.961***	2.289***	1.698***
	(12.795)	(16.803)	(14.735)	(0.104)	(0.082)	(0.104)
Observations	1,027	1,027	1,027	1,027	1,027	1,027
R^2	0.067	0.125	0.033	0.095	0.051	0.240
Treatment	−75.754***	−116.239***	−59.637***	0.803***	0.408***	1.571***
	(16.303)	(21.499)	(18.889)	(0.134)	(0.105)	(0.134)
Professional	17.600	31.638*	5.618	0.080	0.121	0.357***
	(14.048)	(18.525)	(16.276)	(0.116)	(0.090)	(0.115)
Treatment x Professional	−17.507	−63.466**	−10.098	0.004	0.050	−0.251
	(20.567)	(27.122)	(23.829)	(0.169)	(0.132)	(0.169)
Constant	286.139***	328.694***	308.891***	3.019***	2.341***	1.773***
	(10.904)	(14.379)	(12.634)	(0.090)	(0.070)	(0.090)
Observations	1,041	1,041	1,041	1,041	1,041	1,041
R^2	0.069	0.125	0.031	0.087	0.049	0.233

Note: *p<0.1; **p<0.05; ***p<0.01.

Figure F1: Relationship between time spent and performance

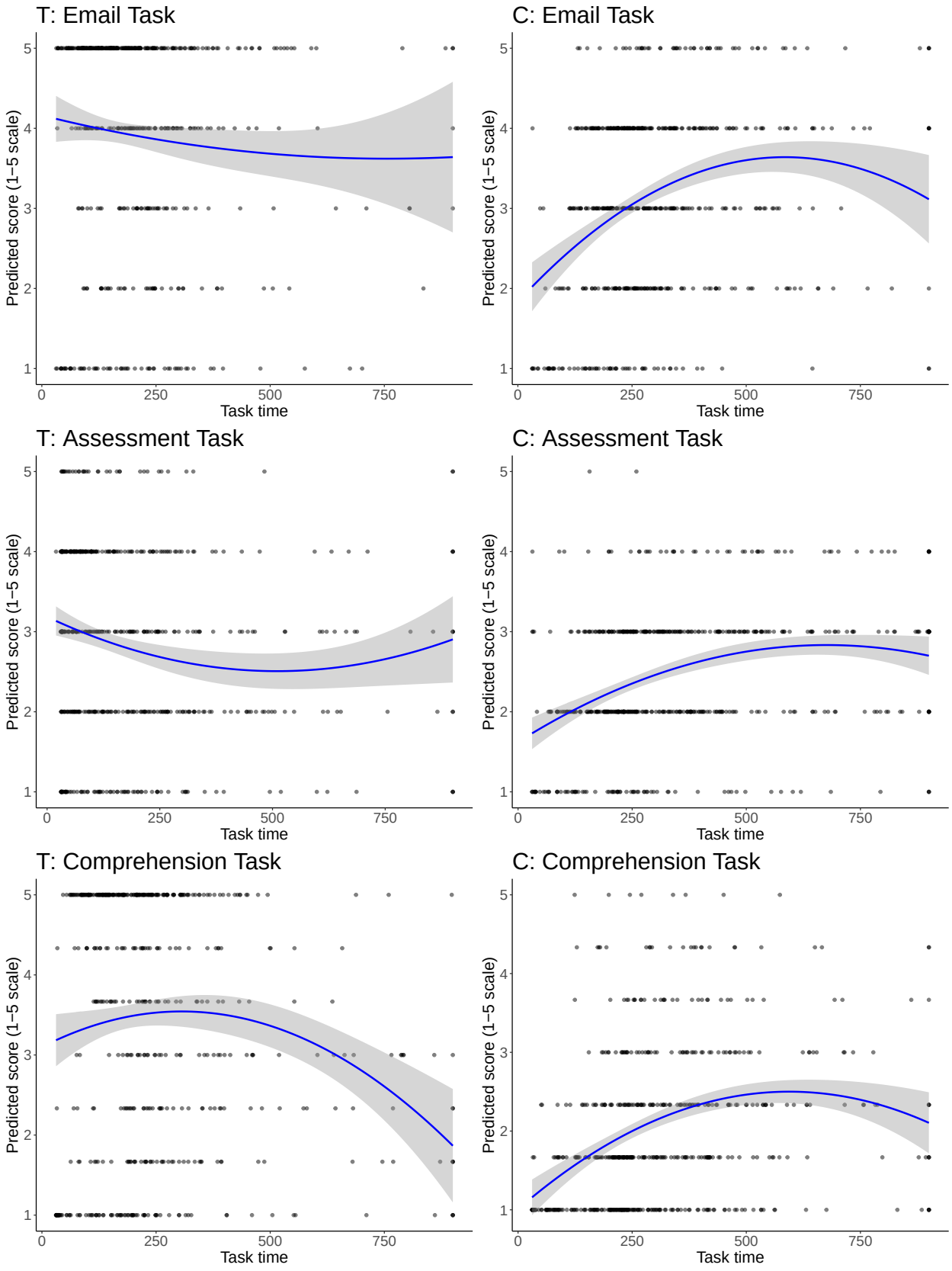


Table E5: Main analysis with controls

	Dependent variable:					
	Email Time	Ass. Time	Comp.Time	Email Score	Ass. Score	Comp. Score
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	−89.619***	−158.758***	−65.124***	0.841***	0.437***	1.346***
	(9.875)	(13.024)	(11.484)	(0.081)	(0.064)	(0.081)
Age 36-50	−2.146	1.103	8.221	−0.108	−0.183**	−0.206**
	(11.078)	(14.610)	(12.883)	(0.091)	(0.072)	(0.091)
Age 51+	42.462***	94.971***	61.324***	−0.149	−0.282***	−0.446***
	(14.201)	(18.730)	(16.516)	(0.117)	(0.092)	(0.116)
Female	13.753	5.645	6.273	0.092	−0.043	0.026
	(10.698)	(14.110)	(12.442)	(0.088)	(0.069)	(0.088)
Uni	0.173	−3.994	5.922	0.203*	0.094	0.188*
	(12.740)	(16.803)	(14.817)	(0.105)	(0.083)	(0.104)
Professional	4.239	−14.909	−12.660	0.122	0.211***	0.232**
	(11.125)	(14.672)	(12.938)	(0.091)	(0.072)	(0.091)
Constant	290.549***	350.725***	305.888***	2.881***	2.358***	1.885***
	(14.436)	(19.039)	(16.789)	(0.119)	(0.093)	(0.118)
Observations	1,027	1,027	1,027	1,027	1,027	1,027
R^2	0.084	0.149	0.044	0.104	0.063	0.229

Note: Main analysis with demographic controls. The results of the interaction models with controls are similar. *p<0.1; **p<0.05; ***p<0.01.

G Tasks and Grading Scheme

After answering the self-assessment questions, the participants were informed that they would now start their first task. Depending on the group they were in, they were told to complete the tasks with/without using ChatGPT. Participants were also reminded that they may earn a substantial bonus if they perform well and that we expect them to take approximately four minutes for each task.

The grading scheme works as follows. ChatGPT (GPT-4) is prompted with the evaluation prompt and the relevant text for the task. The next prompt consists of a number of responses, identified with the respondent ID and clearly separated from one another. ChatGPT then provides a score and a short justification for the score for each answer. The scores are recorded in an Excel file. This process is repeated iteratively. Periodically, ChatGPT has to be re-prompted with the initial prompt and text, owing to its limited context window.

Email Task

Participant instructions: Please carefully read the following...

Lars is a 28 year old manager at a local restaurant. The restaurant owner has noticed that the cleanliness of the restaurant has deteriorated in recent months, and customers are waiting longer for their orders to arrive. Lars has been struggling with scheduling shifts to match customer demand, leaving him short staffed on occasion and over staffed at other times. The restaurant owners are sending Lars the following email. Could you improve it? You may copy the text below and make any changes you deem appropriate.

[Email text here]

Email text: "Dear Lars,

I am writing to express my disappointment in your recent performance. Your work has not been up to the standard that is being expected from our employers, and it has been impacting the overall productiveness of the team.

You lack attention to detail and failure to meet deadlines has resulted in missed opportunities and last revenue for the company. This is unacceptable and cannot continued.

I want to make it clear that if your performance does nor improve significantly in the coming weeks, we will have no choice but to conclude your employment with us. We have highest standards here, and we expect all employees to meet them.

I strongly encourage you to take a clear look at your life and make the necessary changes to improve. This includes better time management, increased attention to detail, and increased communications with your colleagues. We are here to support you, but ultimately it is up to you to make the initiative to improve your work.

I hope to see a significant improvement of your performance in the future weeks. If you have any questions or concerns, please do not hesitate to reach us.

Sincerely,

Restaurant Owner"

ChatGPT evaluation prompt: Below is a draft of an email. I want you to remember it. In the following, I will show you revised versions of the email. Evaluate whether the revised versions constitute an improvement, considering the following criteria: language (Have all or almost all spelling and grammar mistakes been corrected? Is the language clear and concise, without adding redundant information?) and tone (The revised version should be more constructive and not directly threaten termination).

Apply a 5-point scale, where the levels mean the following:

5/5: Noticeable improvement on both dimensions

4/5: Noticeable improvement only on one dimension (e.g., corrected all or almost all spelling and grammar mistakes but no improvement in the tone of the email)

3/5: No improvement on either dimension (this includes cases where only a minority of the mistakes have been corrected) OR improvement in one dimension counterbalanced by worsening in the other (e.g., corrected all or almost all spelling and grammar mistakes but adopted even harsher tone)

2/5: Revised version is worse than the draft on one dimension and unchanged on the other (e.g., spelling mistakes have not been corrected and tone is even harsher)

1/5: Revised version is worse than the draft on both dimensions, or is a non-answer such as N/A.

There are no half-grades (e.g., 4.5/5). Ignore encoding errors. "||||||||" denotes a line break.

Original draft:

[Email text here]

The revised versions to evaluate will be below, in the format "answer ID": "answer text". Provide output in the following format: answer ID $\sim$ x/5 $\sim$ comments.

Assessment Task

Participant instructions: Please carefully read the following...

The following two texts, adapted from a UK newspaper, take opposing views on the question of whether a universal basic income (UBI) is a good idea.

[Text 1 here]

[Text 2 here]

Please provide a short assessment (no more than 200 words) of which of the texts is more convincing. Do not simply state which text you agree with more, but provide a reasoned assessment as to why one of the texts is more convincing than the other.

There is no right or wrong answer, people may have valid reasons for finding either text more convincing. You may answer in bullet points.

Text 1: "The idea of a UBI has made recurrent appearances in history. This time, though, it is likely to have greater staying power, as the prospect of sufficient income from jobs grows bleaker for the poor and less educated.

UBI is a somewhat uneasy mix of two objectives: poverty relief and the rejection of work as the defining purpose of life. The first is political and practical; the second is philosophical or ethical.

The main argument for UBI as poverty relief is the inability of available paid work to guarantee a secure and decent existence for all. In the industrial age, factory work became the only source of income for most people – a source that was interrupted by periodical bouts of unemployment. The labour movement responded by demanding "work or maintenance". This led to the creation of a system of social security,

"welfare capitalism", designed to provide people an income during enforced interruptions of work. Soon, the idea of interruption from work was extended to include the disabled and women bringing up children.

In the 1980s, conservative governments unwittingly extended the scope of welfare further, as they dismantled institutions and legislation designed to protect wages and jobs. "In work" benefits, were introduced to enable employed workers to earn a "living wage". At the same time, governments started to cut back on welfare entitlements. In this newly precarious environment of work and welfare, UBI is seen as guaranteeing the basic income previously promised by work and welfare, but no longer reliably secured by either.

The ethical case for UBI is different. Its source is the idea, found both in the Bible and in classical economics, that work is a curse or "cost", undertaken only for the sake of making a living. As technological innovation causes per capita income to rise, people will need to work less to satisfy their needs. UBI provides a practical path to navigate this transition. Most of the hostility to UBI has come when it is stated in this second form. But to argue that an income independent of the job market is bound to be demoralising is historically inaccurate. If it were true, we would want to abolish all inherited income.

A standard objection to UBI is that it is unaffordable. This partly depends on what parameters are set: the UBI's level; which benefits (if any) it replaces; whether only citizens or all residents are eligible; and so on. But this is not the main point. The overwhelming evidence is that the lion's share of productivity gains in the last 30 years has gone to the very wealthy. Even a partial reversal of this long regressive trend for wealth and income would fund a modest initial basic income.

Unless we change our system of income generation, there will be no way to check the concentration of wealth in the hands of the rich. A UBI that grows in line with capital productivity would ensure that the benefits of growth go to the many, not just to the few."

Text 2: "A recent study sheds doubt on ambitious claims made for a UBI, the scheme that would give everyone regular, unconditional cash payments that are enough to live on. Its advocates claim it would help to reduce poverty, and narrow inequalities.

New research reviewed for the first time 16 projects that have tested different ways of distributing regular cash payments to individuals across a range of countries, as well as copious literature on the topic. It could find no evidence to suggest that such a scheme could be sustained for all individuals in any country – or that this approach could achieve lasting improvements in wellbeing or equality. The research confirms the importance of generous income support, but everything turns on how much money is paid, under what conditions and with what consequences for the welfare system.

Cash payment schemes around the world have been claimed to show that UBI "works". In fact, what's been tested in practice is almost infinitely varied. The Alaska Permanent Fund pays all adults and children

a dividend each year – in 2018, it was $1,600 (£1,230). The scheme is popular and enduring; but it makes no claim to sufficiency and has done nothing to reduce child poverty or to prevent widening income inequalities. Finland undertook a two-year trial of modest monthly payments of €560 (£477) to 2,000 unemployed people – but the government has refused to fund further expansion. It told us little about UBI except that, when push comes to shove, elected politicians may balk at paying for a universal scheme.

The cost of a sufficient UBI scheme would be extremely high according to the International Labour Office, which estimates average costs equivalent to 20-30% of GDP in most countries. Costs can be reduced by paying smaller amounts to fewer individuals. But there is no evidence that a partial or conditional UBI could do anything to mitigate, let alone reverse, current trends towards worsening poverty and inequality.

Costs may be offset by raising taxes or shifting expenditure, but either way there are huge trade-offs. Money spent on cash payments cannot be invested elsewhere. As the report observes, "If cash payments are allowed to take precedence, there's a serious risk of [...] setting a pattern for future development that promotes commodification rather than emancipation." This may help to explain why UBI has attracted support from Silicon Valley tycoons, who are more interested in defending consumer capitalism than in tackling poverty and inequality.

The report concludes that the money needed to pay for an adequate UBI scheme "would be better spent on reforming social protection systems, and building more and better-quality public services". Collective provision offers more cost-effective, socially just, redistributive and sustainable ways of meeting people's needs."

ChatGPT evaluation prompt: Below are two texts on the merits of universal basic income (UBI). Remember them. In the next prompts, they will be followed by assessments which text is more convincing. Evaluate the assessments. There is no right or wrong answer, people may have valid reasons for finding either text more convincing. The evaluation should focus on how well-reasoned the assessment is.
Apply a 5-point scale, where the levels mean the following:

> 5/5: The answer provides a comprehensive and well-reasoned appraisal of strengths and weaknesses of both texts. It includes a plausible judgment which text is more convincing which follows from the appraisal.
>
> 4/5: The answer provides a well-reasoned appraisal of strengths and weaknesses of both texts, but it lacks detail. The judgment which text is more convincing follows from the incomplete appraisal of the texts.
>
> 3/5: The appraisal of strengths and weaknesses of both texts is not well-reasoned or comprehensive

(for example, it only discusses the strengths of one text) and/or the judgment which text is more convincing does not follow from the appraisal.

2/5: The answer is an unsupported statement which does not discuss the strengths and weaknesses of the texts, or simply summarises the texts.

1/5: The answer is unrelated to the texts, or is unintelligible, or comments on the merits of UBI itself rather than on the quality of the texts, or is a non-answer such as N/A.

There are no half-grades (e.g., 4.5/5). Ignore encoding errors. "||||||||" denotes a line break.

[Text 1 here]

[Text 2 here]

The assessments to evaluate will be below, in the format "answer ID": "answer text". Provide output in the following format: answer ID ∼ x/5 ∼ comments.

Text Comprehension Questions

Participant instructions: Please carefully read the following...

[Text here]

Please complete the sentences below so that they accurately reflect the information contained in the text.

Regarding the cultivation of cereal grains, the passage indicates that pollen analyses have provided evidence against...

The author's likely position on the limits of historical records is...

In the structure of the author's argument, the relationship between the second paragraph and the final paragraph of the passage is...

Text: "In tracing the changing face of the Irish landscape, scholars have traditionally relied primarily on evidence from historical documents. However, such documentary sources provide a fragmentary record at best. Reliable accounts are very scarce for many parts of Ireland prior to the seventeenth century, and many of the relevant documents from the sixteenth and seventeenth centuries focus selectively on matters relating to military or commercial interests.

Studies of fossilized pollen grains preserved in peats and lake muds provide an additional means of investigating vegetative landscape change. Details of changes in vegetation resulting from both human

activities and natural events are reflected in the kinds and quantities of minute pollen grains that become trapped in sediments. Analysis of samples can identify which kinds of plants produced the preserved pollen grains and when they were deposited, and in many cases the findings can serve to supplement or correct the documentary record.

For example, analyses of samples from Long Lough in County Down have revealed significant patterns of cereal-grain pollen beginning by about 400 A.D. The substantial clay content of the soil in this part of Down makes cultivation by primitive tools difficult. Historians thought that such soils were not tilled to any significant extent until the introduction of the moldboard plough to Ireland in the seventh century A.D. Because cereal cultivation would have required tilling of the soil, the pollen evidence indicates that these soils must indeed have been successfully tilled before the introduction of the new plough.

Another example concerns flax cultivation in County Down, one of the great linen-producing areas of Ireland during the eighteenth century. Some aspects of linen production in Down are well documented, but the documentary record tells little about the cultivation of flax, the plant from which linen is made, in that area. The record of eighteenth-century linen production in Down, together with the knowledge that flax cultivation had been established in Ireland centuries before that time, led some historians to surmise that this plant was being cultivated in Down before the eighteenth century. But pollen analyses indicate that this is not the case; flax pollen was found only in deposits laid down since the eighteenth century.

It must be stressed, though, that there are limits to the ability of the pollen record to reflect the vegetative history of the landscape. For example, pollen analyses cannot identify the species, but only the genus or family, of some plants. Among these is madder, a cultivated dye plant of historical importance in Ireland. Madder belongs to a plant family that also comprises various native weeds, including goosegrass. If madder pollen were present in a deposit it would be indistinguishable from that of uncultivated native species."

Pollen Question

ChatGPT evaluation prompt: Below is a text passage, remember it. It will be followed by different statements. Each of the statements is meant to complete the sentence "Regarding the cultivation of cereal grains, the passage indicates that pollen analyses have provided evidence against…". The key takeaway from the passage is that pollen analyses have provided evidence against the view that in certain parts of County Down, cereal grains were not cultivated to any significant extent before the seventh century, when the moldboard plough was introduced.

Evaluate the accuracy of the statements on a scale from 1 to 5 according to the following grading scheme:

5/5: The answer correctly identifies the key takeaway and mentions cereal grains.

3/5: The answer correctly identifies one of the tangential points, such as that pollen evidence challenges the view that soil was not tilled in Country Down before the seventh century or that pollen evidence may challenge documentary evidence.

1/5: The answer misrepresents the information in the text, or is unrelated to the question, or is unintelligible, or is a non-answer such as N/A.

Do not use scores other than 1, 3, and 5. Ignore encoding errors.

[Text here]

The statements to evaluate will be below, in the format "answer ID": "answer text". Provide output in the following format: answer ID ∼ x/5 ∼ comment (max. 1 sentence). Separate the evaluations by an empty line.

Historical Records Question

ChatGPT evaluation prompt: Below is a text passage, remember it. It will be followed by different statements. Each of the statements is meant to complete the sentence "The author's likely position on the limits of historical records is…". The key takeaway is that the author believes that while historical documents are valuable for tracing the past, their record is often fragmentary and focuses selectively on certain aspects. Thus, the author believes these documents should be supplemented with other investigative techniques to provide a more comprehensive understanding of history.

Evaluate the accuracy of the statements on a scale from 1 to 5 according to the following grading scheme:

5/5: The answer correctly identifies a) the author's likely view of historical records as fragmentary and/or selective and b) the author's likely view that additional investigative techniques should be used to supplement or correct the documentary record.

3/5: The answer correctly identifies either a) the author's likely view of historical records as fragmentary and/or selective or b) the author's likely view that additional investigative techniques should be used to supplement or correct the documentary record.

1/5: The answer misrepresents the author's likely position, or is unrelated to the question, or is unintelligible, or is a non-answer such as N/A.

Do not use scores other than 1, 3, and 5. Ignore encoding errors.

[Text here]

The statements to evaluate will be below, in the format "answer ID": "answer text". Provide output in the following format: answer ID ∼ x/5 ∼ comment (max. 1 sentence). Separate the evaluations by an empty line.

Paragraph Structure Question

ChatGPT evaluation prompt: Below is a text passage, remember it. It will be followed by different statements about the relationship between the second and the last paragraph of the text. Each of the statements is meant to complete the sentence "In the structure of the author's argument, the relationship between the second paragraph and the final paragraph of the passage is...". The important takeaway is that the final paragraph qualifies the claim made in the second paragraph.

Evaluate the accuracy of the statements on a scale from 1 to 5 according to the following grading scheme:

> 5/5: The answer correctly identifies that the final paragraph qualifies the claim made in the second paragraph.
>
> 3/5: The answer correctly identifies that the author discusses the limitations of pollen analysis, but fails to make explicit the argumentative structure of the text and relationship between the paragraphs.
>
> 1/5: The answer misrepresents the relationship between the paragraphs, or is unrelated to the question, or is unintelligible, or is a non-answer such as N/A.

Do not use scores other than 1, 3, and 5. Ignore encoding errors.

[Text here]

The statements to evaluate will be below, in the format "answer ID": "answer text". Provide output in the following format: answer ID ∼ x/5 ∼ comment (max. 1 sentence). Separate the evaluations by an empty line.

Inter-Coder Reliability

Table G1: Inter-coder reliability between ChatGPT and human coders

Task	GPT-4 & R1	GPT-4 & R2	R1 & R2
Email	0.702	0.662	0.851
	(5.72)	(4.92)	(12.4)
Assessment	0.72	0.77	0.905
	(6.15)	(7.69)	(20.1)
Comprehension	0.94	0.888	0.874
	(32.3)	(16.9)	(14.9)

Note: Intra-class correlations (F-statistics in parentheses) calculated based on a random sample of 100 responses that were also graded by two of the authors. The results show high ICC for all pairings.