

Artificial Intelligence in Labor Market Matching

by

Emma Benz Wiles

S.M., Massachusetts Institute of Technology, 2022

B.A., University of Washington, 2015

Submitted to the Department of Management
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY IN MANAGEMENT

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2024

© 2024 Emma Benz Wiles. This work is licensed under a [CC BY-NC-ND 4.0](#) license.

The author hereby grants to MIT a nonexclusive, worldwide, irrevocable, royalty-free license to exercise any and all rights under copyright, including to reproduce, preserve, distribute and publicly display copies of the thesis, or release the thesis under an open-access license.

Authored by: Emma Benz Wiles
Department of Management
May 3, 2024

Certified by: John Horton
Department of Management, Thesis Supervisor

Accepted by: Eric So
Professor, Global Economics and Finance
Faculty Chair, MIT Sloan PhD Program

Artificial Intelligence in Labor Market Matching

by

Emma Benz Wiles

Submitted to the Department of Management
on May 3, 2024, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Management

Abstract

In my dissertation I study three applications of AI in labor market matching. In my first chapter I show that AI-improved but not entirely written resumes make workers more likely to be hired with no negative downstream implications to employers or to match quality. However, in my second chapter I show that when employers are given entirely AI written drafts of a job post, the jobs posted are more generic and less likely to make a hire. Lastly, I provide evidence that non-technical workers can use AI to upskill into data science, however those skills do not persist in absence of AI assistance.

My first chapter investigates the association between writing quality in resumes for new labor market entrants and whether they are ultimately hired. I show this relationship is, at least partially, causal: in a field experiment in an online labor market with nearly half a million jobseekers, treated jobseekers received algorithmic writing assistance on their resumes. I find that the writing on treated jobseekers' resumes had fewer errors and was easier to read. Treated jobseekers were hired 8% more often, at 10% higher wages. Contrary to concerns that the assistance takes away a valuable signal, I find no evidence that employers were less satisfied with the quality of work done, using star ratings, the sentiment of their reviews, and their probability of rehiring a worker. The analysis suggests digital platforms and their users could benefit from incorporating algorithmic writing assistance into text-based descriptions of labor services or products without downstream negative consequences.

In my second chapter, I study a randomized experiment conducted on an online labor market that encouraged employers to use a Large Language Model (LLM) to generate a first draft of their job post. Treated employers are 20% more likely to post the job and decrease time spent writing their job post by 40%. Among the posted jobs, treated employers receive 5% more applications. Despite this, they are 18% less likely to hire. I find no evidence that this is driven by treated employers receiving lower quality applicants. Moreover, despite the large increase in the number of jobs posted, there is no difference in the overall number of hires between treatment and control employers. These results imply that the treatment lowered the probability of hiring among at least some jobs which would have otherwise made a hire. I rationalize these results with a model in which employers with heterogeneous values of hiring can attract better matches by exerting effort to precisely detail required

skills. I show how a technology that lowers the cost of writing and imperfectly substitutes for effort causes more posts, but lowers the average hiring probability through both marginal posts (as these are less valuable) and inframarginal posts (as the technology crowds out effort and makes the job posts more generic). I provide evidence for these mechanisms using employer screening behavior and the embeddings of the job posts' texts.

In my third chapter, we investigate if LLMs can be used to help non-technical workers adapt to technology induced, rapidly changing skill demands by “upskilling” into a more technical skillset. With coauthors at Boston Consulting Group, we run a randomized control trial on knowledge workers, who have no data science experience, to test whether workers paired with LLMs are able to perform data science tasks to the level of real data scientists. We give consultants at BCG data science problems, representative of what the data scientist role at the company demands, but which GPT-4 cannot solve on its own. We find that treated workers given access to and training in using ChatGPT are more likely to correctly solve all three tasks, and can perform at the level of real data scientists without GPT-4 on the coding task. These results suggest that LLMs can be used to help workers gain new skills to meet the evolving, more technical demands of the labor market, but that for some types of tasks the work of non-technical workers is not interchangeable with data scientists’.

Thesis Supervisor: John Horton

Title: Richard S. Leghorn (1939) Career Development Professor

Acknowledgments

First and foremost, I want to thank the MIT Sloan School of Management & Economics department, for cultivating the richest possible intellectual community without ever confusing meanness for cleverness.

To my committee members, Catherine Tucker and Dean Eckles, who were always available for a cup of coffee or advice. To all the other faculty who contributed so much to my learning, to the Seattle Minimum Wage Study team, and to Apostolos Filippas. I especially want to thank David Autor and Daron Acemoglu for always treating me like a labor economist. And to all of the labor faculty and students in the economics department—I will so miss labor lunch and the stimulating and downright fun conversations we have at labor coffee every week.

I am so grateful for all my fellow students who made the last five years fly by. I don't know what I would have done without my (M)IT elders in the early days of the PhD: David Holtz, Sebastian Steffen, Alex Moehring, and Zanele Munyikwa. And to Hongyi Tu Ye, Mohammed Alsobay, Ben Manning and all the current IT students—you make this program very hard to leave. I also want to thank the economics students in my year who were so welcoming both academically and socially, from Math Camp in 2019 till today. I can't imagine having gone through this process without Ahmet Gulek who has seen the first & worst version of every talk I've ever given, Advik Shreekumar for being the unofficial co-social chair of our friend group, and Tishara Garg for reminding me not to be too serious. I am so grateful for the gift of these (and more) life-long friends.

I will be forever indebted to my advisor, John Horton, without whom I would not be at MIT. For my entire time as a graduate student, John has been teaching me new skills, helping me navigate academia, and reminding me not to get emotionally attached to results. He has been so generous with his advice, and even more importantly, with his attention. The thing I am most excited about in my next phase of academic life is that I will be able to *try* to be to some new student the kind of advisor John has been to me.

I am so grateful for the support of all my friends and family for always giving me a soft place to land. Especially for my siblings, Alex and Grant van Inwegen, who have been there

to take every stressed or ecstatic phone call I've asked of them. The long distance support of my best friends Maddy Grupper, Erin Tudor, Melissa Rienstra, and Emma Allen have kept me sane and grounded. I want to thank my mother-in-law, Alix Davenport, for being the only person outside of my committee to read every one of my papers, and for preferring my research to her son's. And I want to thank my dad, Gregory Benz van Inwegen who for thirty years has never wavered in his belief of my technical and academic abilities, sometimes despite my best efforts to convince him otherwise.

My last glowing review goes to my husband, Edward Wiles, who is the greatest gift I got from MIT. I met Edward on the first day of Math Camp, and he has made every day since lighter and lovelier than I could've ever imagined.

Finally, I dedicate this dissertation to my grandmother Karen Benz Scarvie, who has been my role model ever since I was born on her birthday.

Contents

1	Algorithmic Writing Assistance on Jobseekers' Resumes Increases Hires	11
1.1	Introduction	11
1.2	Empirical context and experimental design	16
1.2.1	Search and matching on the platform	17
1.2.2	Experimental intervention at the resume-writing stage of profile creation	18
1.2.3	The algorithmic writing assistance	19
1.2.4	Platform profile approval	19
1.2.5	Description of data used in the analysis	21
1.2.6	Constructing measures of writing quality	22
1.3	Observational results	23
1.3.1	The association between writing quality and hiring probabilities	23
1.4	Experimental results	26
1.4.1	Algorithmic writing assistance improved writing quality	26
1.4.2	Effects to workers employment outcomes	28
1.4.3	Employers satisfaction was unaffected by the treatment	35
1.4.4	Heterogeneous treatment effects to hiring and ratings	39
1.4.5	Robustness checks	41
1.4.6	Direct tests of the clarity view	41
1.4.7	What happens in general equilibrium?	42
1.5	Conclusion	44
1.6	Appendix	46
1.6.1	Summary statistics and descriptives	46

1.6.2	Effect of the treatment on workers’ earnings	63
1.7	A simple model of the “clarity view” of resume writing	68
1.7.1	A mass of jobseekers with heterogeneous productivity	68
1.7.2	Jobseekers decide whether to put effort into resume-writing	68
1.7.3	The equilibrium fraction of high-type workers putting effort into resume-writing	70
1.7.4	A shift in the resume writing cost distribution leads to more high-type workers choosing to exert effort	71
1.7.5	The effects of lower costs to welfare are theoretically ambiguous	72
2	More, but Worse: The Impact of AI Writing Assistance on the Supply and Quality of Job Posts	73
2.1	Introduction	73
2.2	The setting	78
2.3	Experimental design	80
2.3.1	Experimental intervention at the start of posting a job	80
2.3.2	Description of data used in the analysis	81
2.3.3	Treatment take up	83
2.4	Experimental Results	84
2.4.1	Treated employers were more likely to post a job	84
2.4.2	Treated employers spent less time writing the job post	84
2.4.3	Treated job posts were longer	87
2.4.4	Treated job posts listed different skill requirements	89
2.4.5	Treated job posts received more applications	90
2.4.6	Treated employers job posts were less likely to make an offer, conditional on posting a job	91
2.4.7	Treated non-native English speakers experienced more rejections after making an offer	92
2.4.8	Treated employers were no more likely to make a hire	92
2.5	Mechanisms	93

2.5.1	Employers exhibited lower search effort	94
2.5.2	Treated job posts were more “generic”	94
2.5.3	Treated job posts had a higher fraction of their applications in common with other job posts	97
2.5.4	Applicant pools were not worse overall	99
2.6	Conceptual Framework	100
2.6.1	Period 2: The decision to hire	100
2.6.2	Period 1: The decision to post	101
2.6.3	Treatment	102
2.6.4	Welfare	105
2.7	Difference in differences analysis on near full roll out	105
2.8	Conclusion	109
2.9	Appendix	110
2.9.1	Additional tables and figures	110
2.9.2	Additional tables and figures	112
2.10	Second experiment to understand selection into receiving the AI generated first draft	116
2.10.1	Description of data used in the analysis	117
2.10.2	Experimental intervention at the job description writing stage of job posting	117
2.11	First stage	118
2.12	Results	118
2.12.1	Treated employers were more likely to post a job	118
2.12.2	Employers who opt-in to treatment are slightly positively selected	120
3	Using AI to Upskill Non-Technical Workers into Data Science: A Field Ex- periment	123
3.1	Introduction	123
3.2	Experimental Design	127
3.3	Methods	130

3.3.1	Primary outcomes	130
3.3.2	Task grading	132
3.3.3	Estimating treatment effects	134
3.3.4	Heterogeneous treatment effects	134
3.4	Results	134
3.4.1	Treated workers performed better on data science tasks	134
3.4.2	Treated workers are slightly more likely to complete tasks	135
3.4.3	Treated workers completed tasks faster	137
3.4.4	Treated workers are more confident in their data science skills	138
3.4.5	No impact to workers ability to answer technical problems without help of ChatGPT	139
3.4.6	Treated workers exhibit overconfidence in AI's current capabilities	139
3.4.7	Heterogeneous treatment effects	141
3.5	Conclusion	143
.1	Appendix Tables & Figures	144
.2	Pre-Experiment Survey: Registration for Generative AI Experiment with OpenAI	152
.3	Main experiment	166
.3.1	Survey	166
.4	Main experiment	170
.4.1	Pre-Experiment Survey	170
.4.2	Task details	214
.4.3	Post-Experiment Survey	246
.4.4	Grading Rubrics	261

Chapter 1

Algorithmic Writing Assistance on Jobseekers' Resumes Increases Hires

1.1 Introduction

For most employers, the first exposure to a job candidate is typically a written resume. The resume contains information about the applicant—education, skills, past employment, and so on—that the employer uses to draw inferences about the applicant’s suitability for the job. A well-written resume might influence an employer’s perception of a candidate. One perspective is that a better-written resume—without any change in the underlying facts—might make it easier for the employer to draw the correct inferences about a candidate’s abilities, potentially improving the chance of an interview or job offer. We call this the “clarity view” of the role of resume writing quality. From another perspective, a resume might not merely be a conduit for match-relevant information; the resume’s writing itself could signal ability. In particular, writing quality might provide signals about the jobseeker’s communication skills, attention to detail, or overall quality, potentially leading to a greater chance of a positive outcome. We call this the “signaling view” of the role of resume writing quality.

In this paper, we explore the mechanics of how resume writing quality affects the hiring process. First, using observational data from a large online labor market, we document a strong positive relationship between writing quality in resumes and hiring that persists

even after controlling for obvious confounders. Second, we report the results of a field experiment in which we exogenously vary writing quality in the same market. This experiment directly tests whether there is a causal effect of writing quality on job market outcomes and provides a testing ground to distinguish between the clarity and signaling views.

Our main substantive finding is evidence for the “clarity view.” Evidence for this conclusion is possible because we trace the whole matching process from resume creation all the way to a measure of post-employment satisfaction with a sample of 480,948 jobseekers. This sample size is an order of magnitude larger than the next largest experiments.

Treated jobseekers were more likely to get hired (consistent with both signaling and clarity explanations), but we find no evidence that employers were later disappointed in the quality of work by the treated group, which refutes what the “signaling view” explanation would predict.

To create random variation in writing quality, we intercept new jobseekers at the resume-writing stage of registering for the online labor market. We randomly offer some of them—the treatment group—algorithmic writing assistance, while others—the control group—write their resume under the status quo experience of no assistance. We will discuss this assistance in depth, which we refer to as the Algorithmic Writing Service, but, at a high level, it improves writing by identifying and providing suggestions to resolve common errors. After resume creation, we observe both treated and control jobseekers as they engage in search and, in the case of completed jobs, receive reviews.

In the experimental data, there are quantifiable improvements to resume-writing quality among the treatment group. For example, we find fewer grammar errors, redundancies, and commonly confused words in the resumes of the treated group of jobseekers. These positive effects to writing were greatest at the low-end of the distribution in writing quality, as jobseekers with already excellent resumes benefited little from writing assistance.

One might worry that the treatment could affect behavior. However, we find that, during job search, treated workers did not send out more applications than workers in the control group, nor did they propose higher wage bids. This is a convenient result, as it allows us to focus on the decision-making of the employers. If jobseekers had altered their application behavior—perhaps sending more applications with their stronger resumes—we might

wrongly attribute greater job-market success to the resume rather than this endogenous change in effort.

As for the effect of writing assistance on hiring, we find that treated jobseekers had a 8% increase in their probability of being hired within their first month on the platform relative to the control group. If hired, treated workers' hourly wages were 10% higher than the hourly wages of workers in the control group. This result is downstream of hiring and we provide evidence that it is due to changes in the composition of which workers were hired.

The data make the impact of resume quality on hiring decisions apparent. In order to differentiate between the “signaling view” and “clarity view,” we look at the effect of the treatment on a few different proxy's for employers' satisfaction with the quality of work. We do not find any significant treatment effects to revealed preference measures like hours worked or whether or not workers were ever rehired.

Unique to our setting, we also have explicit measures of employer disappointment, as both sides privately rate each other at the conclusion of the contract. Employer ratings provide a direct way to analyze the informational role of the resume. Specifically, since the treatment removes or at least weakens a credible signal of jobseeker ability, the “signaling view” would suggest that hiring decisions made without this signal should leave employers disappointed. We find no statistically or economically significant treatment effects for any of these ratings. Given the 10% higher average wages in the treatment group, if employers were simply tricked into hiring worse workers generally, these higher wages should have made it even more likely to find a negative effect on ratings (Luca and Reshef, 2021). Moreover, we find that workers are hired for at least as many hours of labor, and are just as likely to be rehired.

One possible explanation for our results is that employers are simply wrong to consider resume writing quality as informative about ability. However, the “clarity view” can also rationalize our results without making this assumption.¹ Our results are consistent with

¹It is helpful to formalize this notion to contrast it with the more typical signaling framing of costly effort and hiring. To that end, we present a model in Appendix Section 1.7 where jobseekers have heterogeneous private information about their productivity but can reveal their type via writing a “good” resume. This model assumes that there are some workers who are unable to write a good resume, for reasons independent of their quality—e.g. due to their English language ability or lack of communication training. We show that relaxing this friction, due to the introduction of a technology which improves those workers writing, and lowers the cost of good writing can generate our findings of more hires, higher wages, and equally satisfied employers.

jobseekers with heterogeneous productivity, where those who receive algorithmic writing assistance face a lower cost to writing clear resumes which reveal their type to employers. We provide evidence for our underlying mechanism, that algorithmic writing assistance improves the clarity of the writing by looking at measures of writing readability (Kincaid et al., 1975). We find consistent evidence that the writing on the resumes of the treated group is easier to read than the resumes in the control group.

We perform this study in the context of a large literature on how experimentally varying applicant attributes affects callback rates (Moss-Racusin et al., 2012; Bertrand and Mullainathan, 2004; Kang et al., 2016; Farber et al., 2016). More specifically, we contribute by showing the importance of text in understanding matching (Marinescu and Wolthoff, 2020). The notion that better writing can help a reader make a better purchase decision is well-supported in the product reviews literature (Ghose and Ipeirotis, 2010) but is a relatively novel finding in labor markets.

Writing has long been used for evaluation across many spheres, for example school essays, personal statements, and cover letters in job applications. While we are not the first to investigate how writing matters to employers² (Sterkens et al., 2021; Martin-Lacroux and Lacroux, 2017), we believe we are the first to do so in a field experiment with natural variation in writing quality. In one related example, Sajjadi et al. (2019) analyze resumes of applicants to public school teaching jobs and find that spelling accuracy is correlated with a higher probability of being hired. Hong, Peng, Burtch and Huang (2021) further show that workers who directly message prospective employers (politely) are more likely to get hired, but the politeness effect is muted when the workers' messages contain typographic errors. Weiss et al. (2022) conducts a lab experiment and finds that the use of AI in jobseekers' writing resulted in employer perceptions of lower competence, warmth and social desirability (however, of particular importance is that in their experiment, the use of AI was disclosed to employers).

These results come at a key time for the evolution of hiring decisions—the practical implications of these two views can inform employers who need to adapt their hiring practices

²While the reason this preference exists is not known, recruiters report, anecdotally, caring about a resume's writing quality (Oreopoulos, 2011).

to a world in which AI can provide substantial quality improvements to application materials. AI capable of generating text is already leaving its mark on labor markets (Eloundou et al., 2023; Felten et al., 2023), and understanding the role of individuals’ writing abilities in predicting their quality is becoming increasingly crucial. Recent research has demonstrated that Large Language Models like ChatGPT can significantly improve worker productivity, particularly by raising bottom of the skill distribution (Noy and Zhang, 2023; Brynjolfs-son et al., 2023a). Similar findings have been reported in other studies on technological advancements, such as the benefit that surgical robots provide to the least proficient surgeons (Tafti, 2022). Our own findings are consistent with these results, as we observed the greatest effects of our treatment among individuals with lower writing quality.

These results can only describe a partial equilibrium. Crowd-out effects are possible if not likely (Crépon et al., 2013; Marinescu, 2017), which are relevant to discuss the welfare implications of any market intervention. Our primary purpose is understanding how employers make decisions with respect to resumes and their role as a tool for lessening information frictions. However there are different implications to platform designers and managers if the introduction of algorithmic writing assistance increases the absolute number of matches or simply changes which jobseeker gets hired. We show that in this setting, the treatment effect is largest for jobseekers who are not competing with as many treated jobseekers, and dissipates based on how much they compete with other treated jobseekers. In the case of the clarity view, even if rolling out the algorithmic writing assistance platform-wide sees a smaller increase in matches than what is found experimentally, there are still benefits to revenue and match quality by introducing the algorithmic writing assistance as a platform-wide policy.

If the “clarity view” is more important to future hiring decisions, then any intervention that encourages better writing will be weakly beneficial for all parties. There will likely be little loss in efficiency if parties are better informed. Even better, as we show, this kind of assistance can be delivered algorithmically. These interventions are of particular interest because they have zero marginal cost (Belot et al., 2018; Briscese et al., 2022; Horton, 2017), making a positive return on investment more likely, a consideration often ignored in the literature (Card et al., 2010). On the other hand, if the “signaling view” is more important,

then providing such writing assistance will mask important information and lead to poor hiring decisions, particularly if writing skills can be conceived of as social skills.³ As for the treatment itself, unlike general advice, algorithmic interventions are adaptive. In our study, the algorithm took what the jobseeker was trying to write as input and gave targeted, specific advice that likely improved it.⁴ This is likely more immediately useful than more vague recommendations, such as telling jobseekers to “omit needless words.”

The rest of the paper proceeds as follows. Section 1.2 describes the online labor market which serves as the focal market for this experiment. Section 1.3 provides evidence on the relationship between writing quality and labor market outcomes from observational data from the market before any intervention. Section 3.4 reports the experimental results of the treatment effects on writing quality and subsequent labor market outcomes. In Section 1.7 we present a simple model that can rationalize our findings. Section 3.5 concludes.

1.2 Empirical context and experimental design

The setting for this experiment is a large online labor market. Although these markets are online, with a global user base, and with lower search costs (Goldfarb and Tucker, 2019), they are broadly similar to more conventional markets (Agrawal et al., 2015). Employers post job descriptions, jobseekers search among job posts and apply. Employers then decide if and who to interview or hire. Jobs can be hourly, or project based. One distinctive feature of online labor markets is that both the employer and the worker provide ratings for each other at the end of a contract.

Because of the many similarities between on and offline labor markets, a substantial body of research uses online labor markets as a setting, often for randomized experiments. Many researchers have used platforms to study the role of information in hiring, as they are difficult to study elsewhere (Stanton and Thomas, 2016; Agrawal et al., 2016; Chan and Wang, 2018; Kokkodis and Ransbotham, 2022). Online labor markets also allow researchers to broaden the range of hypotheses to test (Horton, 2010) because platforms store detailed

³Deming (2017) suggests that there are labor market returns to social skills because they reduce coordination costs and are complementary to cognitive skills.

⁴The Algorithmic Writing Service does not provide whole paragraphs of text, nor is it able to be prompted.

data down to the microsecond on things like applications, text, length of time spent working on an application, speed of hire, and more.

The online labor market which serves as the setting for this experiment is a global marketplace, and not representative of, say, the US workforce. About 20% of the sample comes from anglophone countries US, Canada, UK, and Australia. However, less than 6% of the world’s population comes from these Anglophone countries.⁵ The sample also overweights India, which make up 17% of the global population but 24% of the workers in our sample. As a global marketplace, this market has features that distinguish it from local labor markets. All of the work is Internet-mediated, removing frictions based on geography. Still, there exist frictions based on language and communication skills, which is one of the reasons it makes a good setting to study the role of the resume in hiring. We provide summary statistics about jobs worked on the platform in Appendix Table F1. The average job lasts two months, takes 201 hours of labor with average wages of \$17 per hour. Most of the work measured by the wagebill on the platform consists of hourly jobs, but fixed price jobs make up two-thirds of the total number of contracts formed.

1.2.1 Search and matching on the platform

A would-be employer writes job descriptions, labels the job opening with a category (e.g., “Graphic Design”), lists required skills, and then posts the job opening to the platform website. Jobseekers generally learn about job openings via electronic searches. They submit applications to jobs they are interested in and are required to include a wage bid and a cover letter.

In addition to jobseeker-initiated applications, employers can also use the interface to search worker profiles and invite workers to apply to particular jobs. The platform uses jobseekers’ on-platform history and ratings to both recommend jobseekers directly to would-be employers and to rank them in order of relevance and quality. At no point do these algorithmic recommendations consider the writing quality of the jobseeker’s resume. By using recommendation systems, algorithms can help reduce randomness in the hiring process and provide employers with quality signals about potential hires (Horton, 2017). In terms of se-

⁵https://en.wikipedia.org/wiki/List_of_countries_and_dependencies_by_population

lection, [Pallais \(2014\)](#) shows that employers in an online labor market care about workers' reputation and platform experience when hiring. After jobseekers submit applications, employers screen the applicants. The employers can highlight applications of interest through the platform interface to save in a separate tab, their "shortlist." Then the employer decides whether to conduct interviews, and whether to make an offer(s). If a match is formed, the platform observes the wages, hours worked, earnings, and ratings at the conclusion of the contract. Although these ratings have been shown to become inflated over time and can be distorted when they are public and reciprocal ([Bolton, Greiner and Ockenfels, 2013](#)), they are still a useful signal of worker performance ([Fradkin et al., 2021](#); [Cai et al., 2014](#)). We consider the impact of the treatment to the public and private numerical ratings the employers give to the workers, as well as the "sentiment" of the written text of reviews, which are less prone to inflation ([Filippas et al., 2022](#)).

1.2.2 Experimental intervention at the resume-writing stage of profile creation

When new jobseekers sign up to work on the platform, their first step is to register and create their profile. This profile serves as the resume with which they apply for jobs. This profile includes a list of skills, education, and work experience outside of the platform. It also includes a classification of their primary job category (e.g., "Graphic Design"), mirroring what employers select when posting a job. The interface consists of a text box for a profile title and a longer text box for a profile description. Their finished profile will include their profile description and a "profile hourly wage," which is the wage they offer to employers searching for workers.

During the experimental period, jobseekers registering for the platform were randomly assigned to an experimental cell. The experimental sample comprises jobseekers who joined the platform between June 8th and July 14th, 2021. For treated jobseekers, the text boxes for the profile description are checked by the Algorithmic Writing Service. Control jobseekers received the status quo experience. The experiment included 480,948 jobseekers, with 50% allocated to the treated cell. [Table 1.1](#) shows that it was well-balanced and the balance

of pre-treatment covariates was consistent with a random process.⁶

1.2.3 The algorithmic writing assistance

Words and phrases that a language model determines to be errors are underlined by the Algorithmic Writing Service. See Figure 1-1 for an example of the interface. By hovering a mouse cursor over the underlined word or phrase, the user sees suggestions for fixing spelling and grammar errors. The Algorithmic Writing Service also advises on punctuation, word choice, phrase over-use, and other attributes related to clarity, engagement, tone, and style.

According to the Algorithmic Writing Services website, the software uses a combination of transformer models and rule-based systems to provide its recommendations. Unlike Large Language Models like ChatGPT or BingChat, this system is not generative—it cannot be prompted or asked questions, it simply takes the text the user has provided and suggests improvements to it.

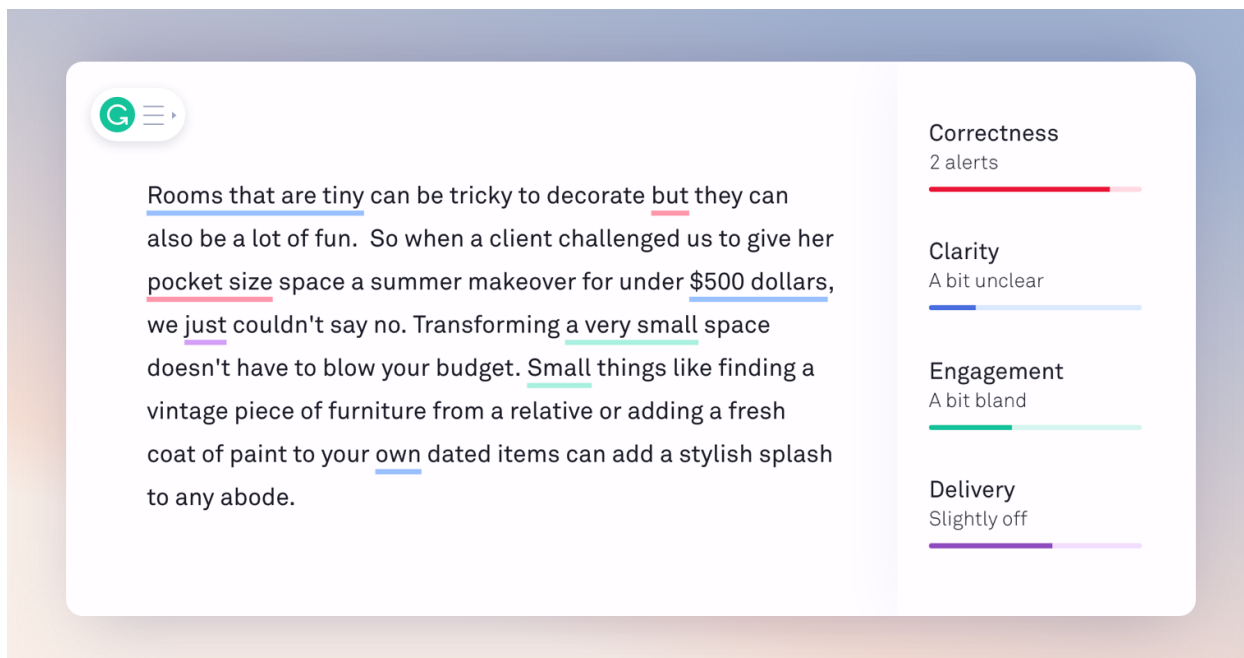
1.2.4 Platform profile approval

Of the experimental sample, 46% of workers allocated into the experiment upon registration complete and submit their profiles. When jobseekers finish setting up their profiles, they have to wait to be approved by the platform. The platform approves jobseekers who have filled out all the necessary information, uploaded ID, and provide bank details so they can be paid. The platform can reject jobseekers at their discretion. However, platform rejection is somewhat rare. About 10% of profiles are rejected, usually as a part of fraud detection or because the jobseekers leave a completely empty profile. About 41% of workers who begin registering get all the way through the approval process.

As approval is downstream of profile creation, this step creates a potential problem for interpreting any intervention that changes profile creation. For example, it could be that the treatment leads to a greater probability of platform approval. Or, the treatment could have made jobseekers more likely to complete the registration process and submit their profile,

⁶In Appendix Figure .11-1 we show the allocations by treatment status over time and find they track closely.

Figure 1-1: Example of the Algorithmic Writing Service's interface showing suggestions on how to improve writing



Notes: Example of the Algorithmic Writing Service applied to a paragraph of text. To receive the suggestions, users hover their mouse over the underlined word or phrase. For example, if you hover over the first clause “Rooms that are tiny” underlined in blue, “Tiny rooms” will pop up as a suggestion.

both of which could effect hiring. While unlikely given the mechanistic rules the platform applies, this is possible, and we investigate this potential issue in multiple ways.

First, we check whether there is any evidence of selection and find no evidence that treated jobseekers are no more likely either to submit their profiles and or to receive approval.⁷

Second, in our main analysis, we condition on profile approval in our regressions. We also perform robustness checks where we report the same analysis not conditioned on profile approval and where we control for profile approval as a covariate. All findings are robust to these strategies, a result described in Section 1.4.5.

⁷See Appendix Table F3 for regression output.

Table 1.1: Comparison of jobseeker covariates, by treatment assignment

	Treatment mean: \bar{X}_{TRT}	Control mean: \bar{X}_{CTL}	Means difference: $\bar{X}_{TRT} - \bar{X}_{CTL}$	p- value
<i>Full sample description: N = 480,948</i>				
Resume submitted	0.45 (0.00)	0.46 (0.00)	0.00 (0.00)	0.45
Platform approved	0.41 (0.00)	0.41 (0.00)	0.00 (0.00)	0.19
Resume length	32.91 (0.12)	32.86 (0.12)	0.05 (0.17)	0.76
Profile hourly rate	18.84 (0.13)	18.92 (0.13)	-0.076 (0.18)	0.68
<i>Flow from initial allocation into analysis sample</i>				
	<i>Treatment (N)</i>	<i>Control (N)</i>	<i>Total (N)</i>	
Total jobseekers allocated	240,231	240,717	480,948	
↪ who submitted their profiles	109,638	109,604	219,242	
↪ and were approved by the platform	97,859	97,610	195,469	
↪ with non-empty resumes	97,479	97,221	194,700	
<i>Pre-allocation attributes of the analysis sample: N = 194,700</i>				
From English-speaking country	0.18 (0.00)	0.18 (0.00)	-0.00 (0.00)	0.36
US-based	0.14 (0.00)	0.14 (0.00)	-0.00 (0.00)	0.22
Specializing in writing	0.17 (0.00)	0.18 (0.00)	-0.00 (0.00)	0.11
Specializing in software	0.16 (0.00)	0.12 (0.00)	0.00 (0.00)	0.77
Resume length	70.39 (0.22)	70.26 (0.22)	0.13 (0.31)	0.67

Notes: This table reports means and standard errors of various pre-treatment covariates for the treatment group and the control group. The first panel shows the post-allocation outcomes of the full experimental sample i) profile submission, ii) platform approval, iii) length of resume in the number of words, iv) profile hourly wage rate in USD. The means of profile hourly rate in treatment and control groups are only for those profiles which report one. The reported p-values are for two-sided t-tests of the null hypothesis of no difference in means across groups. The second panel describes the flow of the sample from the allocation to the sample we use for our experimental analysis. The complete allocated sample is described in the first line, with each following line defined cumulatively. The third panel looks at pre-allocation characteristics of the jobseekers in the sample we use for our analysis, allocated jobseekers with non-empty resumes approved by the platform. We report the fraction of jobseekers i) from the US, UK, Canada, or Australia, ii) from the US only, iii) specializing in writing jobs, iv) specializing in software jobs, and v) the mean length of their resumes in the number of words.

1.2.5 Description of data used in the analysis

The dataset we use in the analysis consists of the text of jobseekers’ resumes as well as all of their behavior on the platform between the time they registered—between June 8th and

July 14th 2021—and August 14th, 2021. We construct jobseeker level data, including the title and text of their profile, the number of applications they send in their first month on the platform, the number of invitations they receive to apply for jobs, the number of interviews they give, and the number of contracts they form with employers. The most common categories listed as worker’s primary job categories are, in order of frequency, Design & Creative, Writing, Administrative Support, and Software Development.

In Table 1.1 we present summary statistics about the jobseekers in the full experimental sample, as well as the sample conditioned on platform approval. Jobseekers with writing as their primary area of work make up 17% of the sample. Only 14% of jobseekers are based in the US, and over 80% are based in a country where English is not the native language.

1.2.6 Constructing measures of writing quality

We do not observe the changes that the Algorithmic Writing Service suggested—we simply observe the resumes that result. As such, we need to construct our own measures of writing quality to determine if the treatment was delivered.

Algorithmic Writing Service provides text improvement suggestions along several dimensions. We measure writing quality of each resume by using a different service, LanguageTool, an open-source software that uses language models to determine various types of writing errors.⁸ LanguageTool is a rule-based dependency parser that identifies errors (rule violations) and categorizes them. Some example categories include “Nonstandard Phrases,” “Commonly Confused Words,” “Capitalization,” and “Typography.” For example, the non-standard phrase “I never have been” would be flagged with a suggestion to replace it with “I have never been.”⁹ Our primary measures of writing quality are the error rates for each of these error types, as well as the overall error rate. The error rate is determined by totaling the number of all error types classified by LanguageTool, normalized by number of words in the resume.

⁸This is a different software than the Algorithmic Writing Service.

⁹For a more detailed explanation of all of the rule categories, see Appendix Table A4.

1.3 Observational results

Before presenting results of the field experiment, we explore the relationship between resume writing quality and hiring using observational data from this market.

1.3.1 The association between writing quality and hiring probabilities

More writing errors are associated with lower hiring probabilities in the observational data. In Figure 1-2, we plot jobseekers' hiring outcomes versus the error rate, controlling for the length of the resume. The sample is the resumes of all jobseekers who registered for the platform over the month of May 7th through June 7th, 2021, prior to the experiment. The distribution of the error rate is very right skewed—over 95% of jobseekers' resumes have error rates of less than 25%. In Figure 1-2, the x-axis is the deciles of error rate, truncated to include only jobseekers whose resumes have error rates of less than 25%. The y-axis is the residuals from regressing the error rate on whether or not the jobseeker is hired, controlling for number of words in the resume. Generally, jobseekers with resumes with a lower error rate (deciles to the left of the plot) are more likely to be hired.

In order to unpack the various types of errors, in Figure 1-3 we show the correlation between hiring outcomes and each individual type of language error in the observational data.¹⁰ In the first specification, we show the correlation between the error rate for the various types of language errors and an indicator for whether or not the jobseeker is ever hired in their first 28 days after registering for the platform. In the second specification, the outcome is simply the number of contracts formed over the jobseeker's first month. In the second specification, we control for the jobseekers' profile hourly rate and primary category of work.

Resumes with more per word grammar errors, typos, typography errors, and miscellaneous errors are all hired less. This linear model places some unreasonable assumptions like constant marginal effects on the relationship between various writing errors and hir-

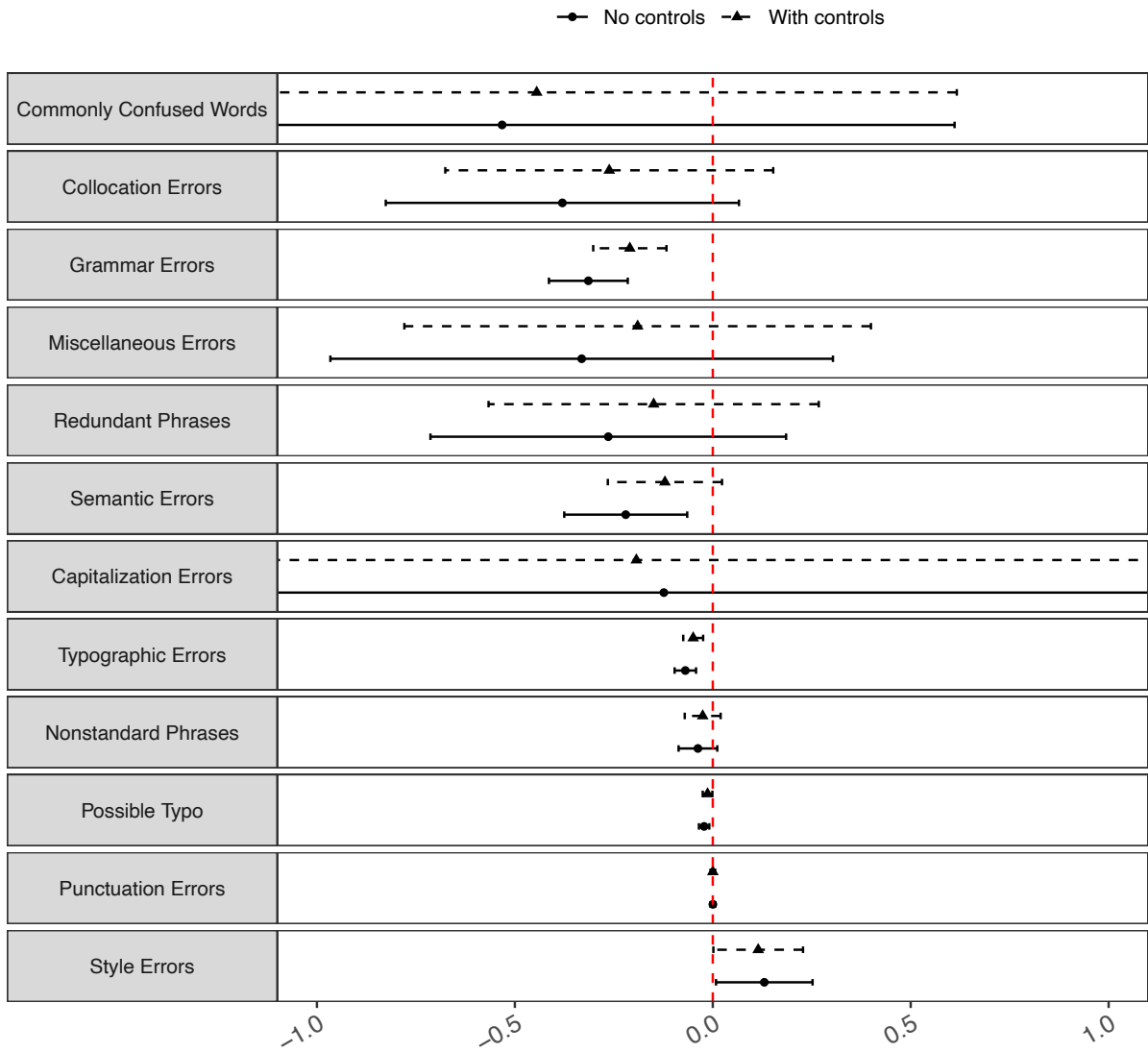
¹⁰In Appendix Table F4 we show the table of these estimates. In Appendix Table F2 we summarize the frequency of these error types in the observational data.

Figure 1-2: Association between resume error rate and if a jobseeker is hired in observational data



Notes: These data show the relationship between the error rate on a jobseekers' resume with the probability they were hired within a month, controlling for resume length. The error rate is determined by totaling the number of all error types classified by LanguageTool, normalized by number of words in the resume. A 95% confidence interval is plotted around each estimate. The sample is all new jobseekers who were approved by the platform between June 1st and June 7th, 2021, with resumes of more than 10 words. The x-axis is error rate deciles on the sample of resumes where the error rate is less than 25%.

Figure 1-3: Relationship between writing error rate and if a jobseeker is hired in observational data



OLS estimate for relationship between each error rate and if a jobseeker gets hired

Notes: These data show the relationship between the error rate on a jobseekers' resume with the probability they were hired within a month. Specification with controls include resume length, jobseeker category, and profile hourly rate. The error rate is determined by the number of each error type classified by LanguageTool, normalized by number of words in the resume. Error type definitions can be found in Appendix Table A4. A 95% confidence interval is plotted around each estimate. The sample is all new jobseekers who were approved by the platform between June 1st and June 7th, 2021. Regression tables the plot is based on can be found in Appendix Table F4.

ing. There may be interactions between these error types. However, it is still useful to summarize the relationships. We can see generally negative relationships between writing error rate and hiring. In the second specification where we add controls, we see coefficients get smaller in magnitude as we would expect, but the significance does not disappear. For robustness we repeat these analysis in levels in Appendix Table F6.

In terms of magnitude, one additional error of any type is associated with that jobseeker being hired 1.4% less. In Appendix Table F5 we show the relationship between total number of errors and hiring outcomes and report these results in both levels and normalized by resume length. The negative relationship between writing errors and hiring persists in all specifications.

1.4 Experimental results

We look at three main kinds of experimental results. First, we examine how the treatment affected the text of resumes. Next, we look at employment outcomes for those treated workers. Third, we will look at how the treatment impacted the quality of work, as assessed by employer reviews and whether or not a worker is rehired.

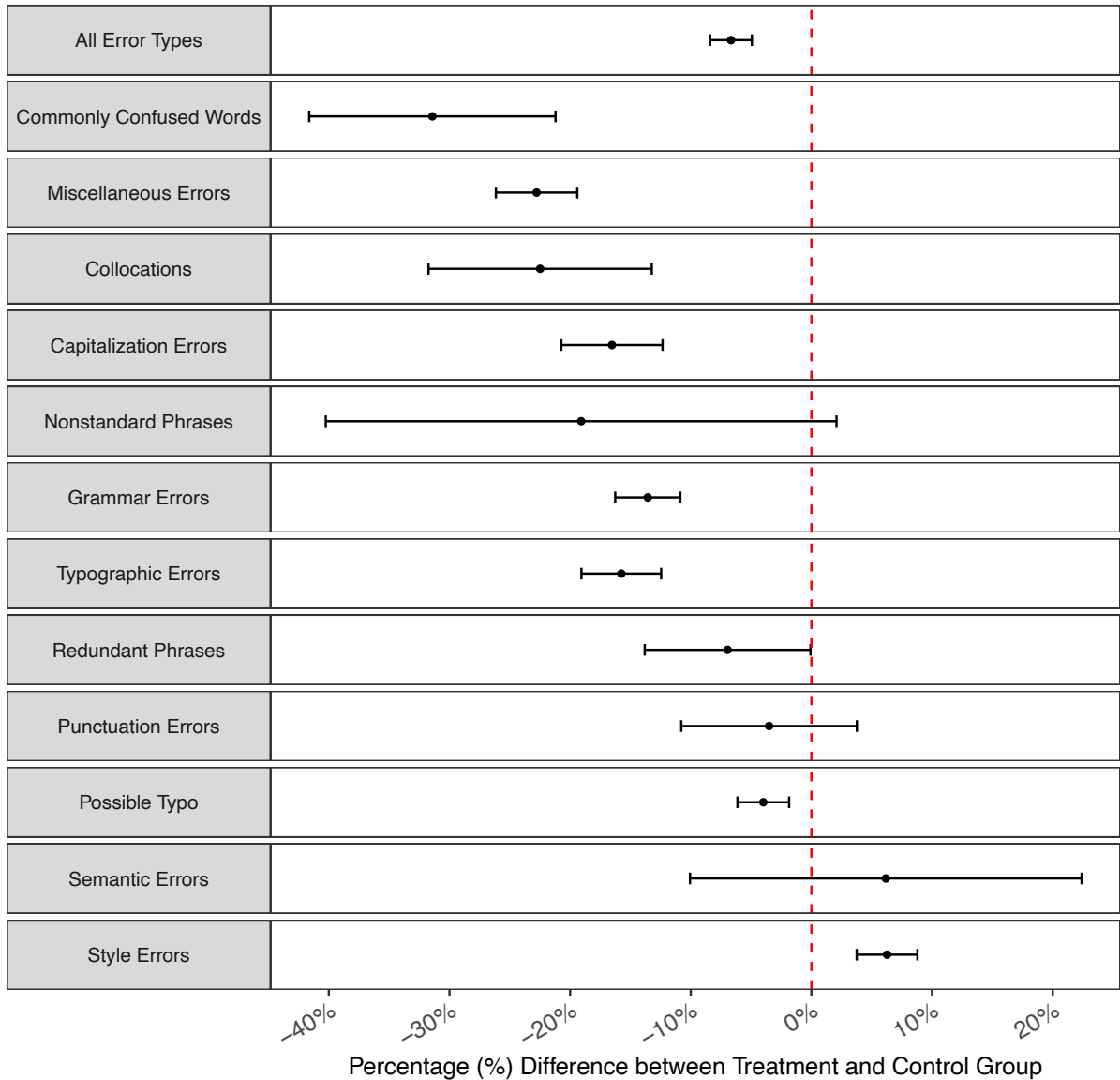
1.4.1 Algorithmic writing assistance improved writing quality

The first step of our analysis is to measure the effect that the Algorithmic Writing Service has on the writing of the resumes in the treatment group. We begin by analyzing the effect of the treatment on all types of writing error rates, as defined by LanguageTool. Figure 1-4 displays the effect of treatment on the number of each type of writing error, normalized by resume length.¹¹ For treatment effects measured in percentage terms, we calculate the standard errors using the delta method.

In the first facet of Figure 1-4, we find that jobseekers in the treatment group made 5% fewer errors in their resumes. Breaking these down by error type, we find that jobseekers in the treatment group had a significantly lower rate of errors of the following types: capitalization, collocations, commonly confused words, grammar, possible typos, miscellaneous,

¹¹The treatment had no effect on the length of resumes—see Appendix Table F7.

Figure 1-4: Effect of the algorithmic writing assistance on resume error rates



Notes: This plot shows the effect of the treatment on various writing errors in jobseekers' resumes, normalized by number of words in the resume. Point estimates are the percentage change in the dependent variable versus the control group for the treatment groups. A 95% confidence interval based on standard errors calculated using the delta method is plotted around each estimate. The experimental sample is of all new jobseekers who registered and were approved for the platform between June 8th and July 14th, 2021, and had non-empty resumes, with $N = 194,700$. Regression details can be found in Appendix Tables F8 and F9. Bonferroni Corrected standard errors can be found in Appendix Table F10.

and typography. We find larger treatment effects for errors associated with writing clarity than for many others. For example, two of the largest magnitudes of differences in error rate were commonly confused words and collocations, where two English words are put together that are not normally found together. Interestingly, the treatment group had more “style” errors, paralleling our results from the observational data (see Table F4).

Heterogeneous treatment effects to writing quality

A natural question is which jobseekers benefited most from the treatment. Appendix Table F13 interacts pre-randomization jobseeker attributes with the treatment. We can see that jobseekers from the US or from English-speaking countries,¹² all have fewer errors in “levels.”

The treatment negatively impacted the writing error rate of all subgroups by country of origin. We find that jobseekers from non-native English-speaking countries experience significantly larger treatment effects to their error rate. Still, effects to their Anglophone counterparts are negative and significant.

In Appendix Table F14 we focus on jobseekers who list their primary category of work as “Writing” and in Column (1) show that the treatment even has a significant impact on the writing on writers’ resumes.

1.4.2 Effects to workers employment outcomes

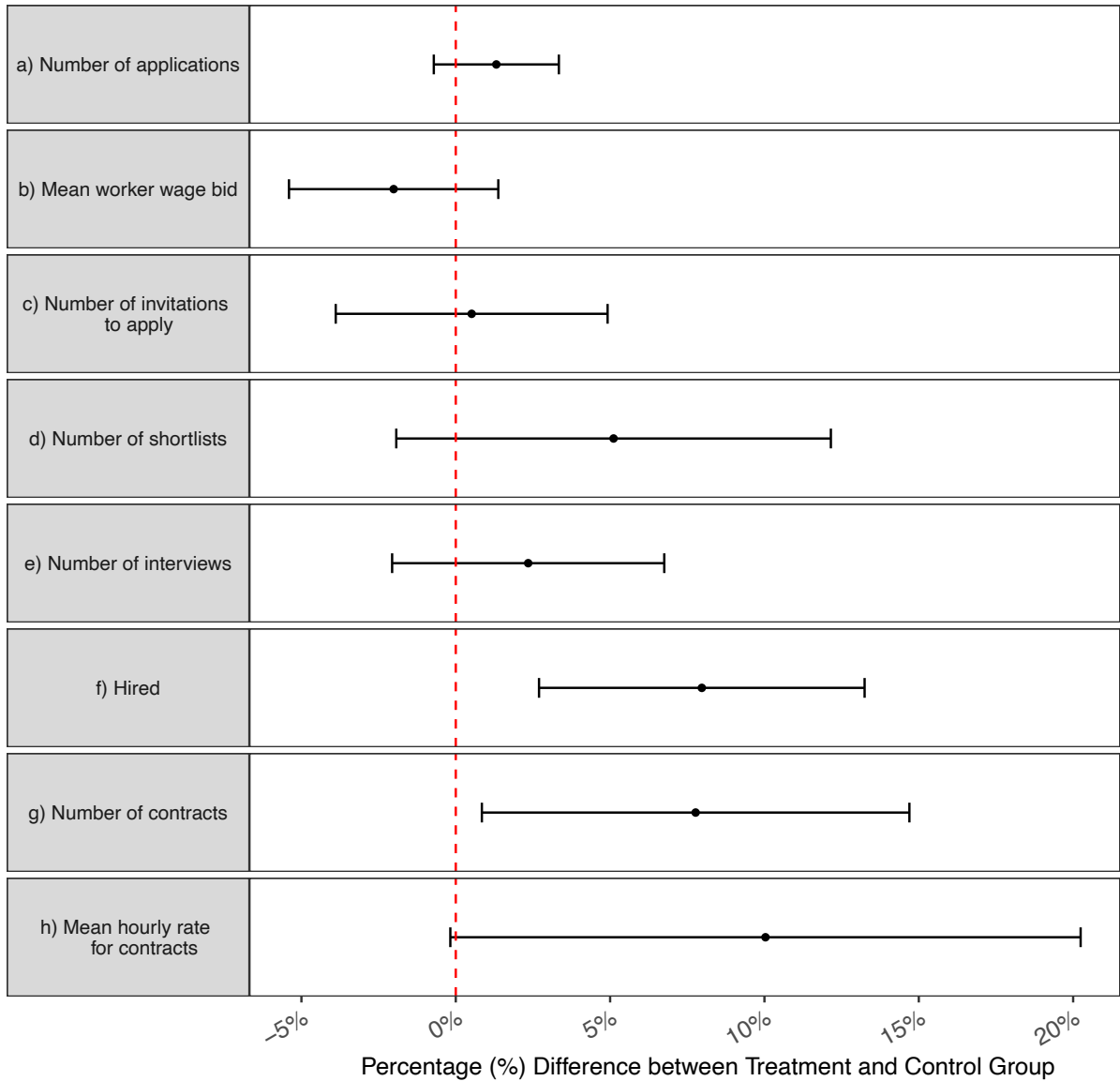
Access to the treatment impacted whether or not jobseekers were hired. Figure 1-5 summarizes the treatment effects on the primary hiring outcomes.

Treated workers did not change their job search strategy or behavior

The potential for the treatment to impact jobseeker search behavior or intensity could complicate our desire to focus on employer decision-making. Job applications have been shown to be costly (Abebe, Caria and Ortiz-Ospina, 2021) and job search intensity could depend on

¹²We define whether a jobseeker is from an Anglophone country, by whether they login to the platform from USA, UK, Canada, Australia, or New Zealand.

Figure 1-5: Effect of algorithmic writing assistance on hiring outcomes



Notes: This analysis looks at the effect of treatment on hiring outcomes on jobseekers in the experimental sample. The x-axis is the difference in the mean outcome between jobseekers in the treated group and the control group. A 95% confidence interval based on standard errors calculated using the delta method is plotted around each estimate. The experimental sample is of all new jobseekers who registered and were approved for the platform between June 8th and July 14th, 2021, and had non-empty resumes, with $N = 194,700$. Regression details on the number of applications and wage bid can be found in Table 1.2. Regression details on invitations, interviews, hires, and the number of contracts can be found in Table 1.3. Regression details on hourly wages can be found in Table 1.4.

jobseekers expectation of their own hireability. It is possible that treated jobseekers realized they were in an experiment and increased their search efforts, knowing they had higher quality resumes. In that case, we could not interpret our treatment effect as being driven by employers' improved perceptions of treated jobseekers. We therefore plot the percentage change in job search metrics for jobseekers in the treatment versus those in the control group in Figure 1-5a) and Figure 1-5b) and find no evidence that jobseekers changed their search behavior.

Table 1.2 provides regression results for these effects of the treatment on jobseekers' search behavior. In Column (1) the outcome is the number of applications a jobseeker sends out over their first 28 days after registering. In the control group, jobseekers send on average 2 applications in their first month on the platform. We find no effect of the treatment on the total number of applications sent.

In Table 1.2 Column (2), the outcome is the mean wage bid proposed by the jobseekers on those applications. We find that treated jobseekers do not apply to more hourly jobs than those in the control group. They also could have bid for higher wages knowing they had better-looking resumes. In Table 1.2 Column (3), the outcome is the mean wage bid proposed by the jobseekers on those applications. Average wage bids in both the treatment and control groups were \$24 per hour. This lack of impact to jobseeker's behavior makes sense as jobseekers were not made aware of the fact that they were in an experiment.

The treatment did not affect employer recruiting

Employers are able to seek out workers using the platform's search feature to invite jobseekers to apply to their job openings. In Figure 1-5c), the outcome is the number of invitations to apply for a job that the jobseeker receives in their first month. We find no effect of the treatment on employer invitations.

This result makes sense given that our experimental sample consists of only new jobseekers to the platform. New entrants almost never appear in the search results when employers search for jobseekers, given that their rank is determined by their platform history. Given that the search feature is how employers find jobseekers to invite to jobs, we would not expect the treatment to affect invitations to apply. Table 1.3 Column (1) provides

Table 1.2: Effects of writing assistance on jobseekers' application behavior

	<i>Dependent variable:</i>		
	Num Apps	Num Hourly Apps	Mean Hourly Wage Bid
	(1)	(2)	(3)
Algo Writing Treatment	0.023 (0.018)	0.012 (0.011)	-0.492 (0.427)
Constant	1.768*** (0.013)	0.919*** (0.008)	24.425*** (0.302)
Observations	194,700	194,700	59,854
R ²	0.00001	0.00001	0.00002

Notes: This table analyzes the effect of the treatment on jobseekers' application behavior. The experimental sample is made up of all new jobseekers who registered and were approved by the platform between June 8th and July 14th, 2021 and had non-empty resumes. The outcome in Column (1) is the number of total applications a jobseeker sent out between the time the experiment began and one month after it ended. The outcome in Column (2) is the number of specifically hourly applications sent out in that same time period. The outcome in Column (3) is the mean hourly wage bid they proposed for those hourly jobs, and the sample narrows to only jobseeker who submitted at least one application to an hourly job.

Significance indicators: $p \leq 0.10$: *, $p \leq 0.05$: ** and $p \leq .01$: ***.

the details of this regression.

After jobseekers apply, employers can sort through the applications to their job and highlight applications they are especially interested in through a feature called shortlisting. In Figure 1-5d) we observe that jobseekers in the treatment group had applications shortlisted 5% more than jobseekers in the control group, although this effect is not significant. Table 1.3 Column (2) provides the details of this regression.

The treatment had no significant impact to number of interviews. In Figure 1-5e), we show the effect of the treatment on number of interviews. Interviews, while technically feasible, are rare on this platform, and do not correspond the types of interviews given in offline labor markets. Here, an interview is defined as any correspondence via message between the employer and applicant, prior to an offer being made. In the control group the average jobseeker gives 0.18 interviews over the course of their first month after registering, with the treatment group receiving 2.5% more interviews. Table 1.3 Column (3) provides the details of this regression.

Treated jobseekers were more likely to be hired

The treatment raised jobseekers’ hiring probability and the number of contracts they formed on the platform. In Figure 1-5f), the outcome is a binary indicator for whether or not a jobseeker is ever hired in their first 28 days on the platform. During the experiment, 3% of jobseekers in the control group worked at least one job on the platform. Treated jobseekers see an 8% increase in their likelihood of being hired in their first month on the platform.

Jobseekers in the treated group formed 7.8% more contracts overall. In Figure 1-5g), the outcome is the number of contracts a jobseeker worked on over their first month. In Table 1.3 Column’s (4) and (5) we report these results in levels.

Table 1.3: Effect of algorithmic writing assistance on hiring outcomes

	<i>Dependent variable:</i>				
	Num Invitations (1)	Num Shortlists (2)	Num Interviews (3)	Hired x 100 (4)	Num Contracts (5)
Algo Writing Treatment	0.001 (0.003)	0.002 (0.001)	0.004 (0.004)	0.247*** (0.080)	0.004** (0.002)
Constant	0.142*** (0.002)	0.039*** (0.001)	0.178*** (0.003)	3.093*** (0.057)	0.047*** (0.001)
Observations	194,700	194,700	194,700	194,700	194,700
R ²	0.00000	0.00001	0.00001	0.00005	0.00003

Notes: This analysis looks at the effect of treatment on hiring outcomes on jobseekers in the experimental sample. The Column (1) outcome Invitations is the number of times they were recruited to a job over their first month. Column (2) is the number of times their application was shortlisted over that month. Column (3) is the number of interviews they gave over that month. Column (4) defines Hired x 100 as one hundred times the probability the jobseeker was hired over that month. Column (5) defines Number of Contracts as the number of unique jobs they work over the month after they register for the platform. The experimental sample is of all new jobseekers who registered and were approved for the platform between June 8th and July 14th, 2021 and had non-empty resumes. Significance indicators: $p \leq 0.10$: *, $p \leq 0.05$: ** and $p \leq .01$: ***.

Hourly wages in formed matches were higher

Treated workers had 10% higher hourly wages than workers in the control group. In Figure 1-5h), the outcome is the mean hourly rate workers earned in jobs they worked over their first month on the platform.¹³

¹³Hourly wage rates for new entrants are not representative of rates on the platform. If a new entrant gets hired for their first job, they tend to experience rapid wage growth.

In Table 1.4 Column (1) we show that in the control group, workers on average made \$17.25 per hour. In the treatment group, workers made \$19.01 per hour, with a p-value of 0.05. Since workers did not bid any higher, it is possible that employers are hiring more productive workers, or that they thought the treated workers were more productive. If that is the case, the “signaling view” would predict that employers would then be disappointed with the workers they hired, which we should be able to observe in worker ratings.

Because these effects are downstream of hiring, these higher wages could be a result of bargaining or due to a composition effect. We find that the initial wage bids are almost always the same as the hourly wage and there is very little evidence of bargaining. In this sample of hires, in only 0.2% of contracts the freelancer proposes more than one bid before being hired. Initial wages and bids are 92% correlated for hourly jobs and 95% correlated for fixed price jobs. In Table 1.4 Column (2) we regress the treatment on an indicator variable defined as 1 if the jobseekers’ initial wage bid is equal to the hourly wage they are hired for, and 0 if not. Using this definition as well, we see no evidence that the treatment increased bargaining.

Taken together with the fact that there is no effect of the treatment to asking wage bids, as we show in Table 1.2, this evidence points to the increase in hourly wages being driven by a composition effect.

Hours worked were unaffected by the treatment

After examining the effects of the treatment on hiring outcomes, we now turn our attention to employer satisfaction with the workers’ labor. One proxy for employer satisfaction is each worker’s total number of hours worked, as this can be an indication of how much demand there is for their services. In Table 1.5 Column (1) we show that treated workers worked no fewer hours than workers in the control group. This sample for this analysis is the entire experimental sample who finished registration and were approved by the platform. The average worker in the control group only works for 2.6 hours during their first month on the platform. However, among those who are ever hired, the average worker in the control group works 238 hours.

Lastly in Column (2) we show the impact of the treatment to the fraction of workers that

Table 1.4: Effect of algorithmic writing assistance on average contract wages

	<i>Dependent variable:</i>	
	Hourly wage rate	I(Bargaining)
	(1)	(2)
Algo Writing Treatment	1.763** (0.834)	-0.027 (0.020)
Constant	17.247*** (0.611)	0.277*** (0.015)
Clustered SEs	X	X
Observations	3,305	1,949
R ²	0.001	0.001

Notes: This analysis looks at the effect of treatment on hourly wages of contracts for jobseekers in the experimental sample, conditional on a hire. The sample is at the job level, and we cluster standard errors at the worker level. The outcome in Column (1), hourly wage rate, is defined as the max hourly wage rate a worker receives for that job. In Column (2) the outcome is an indicator which is 1 if the jobseeker’s wage bid is not equal to the wage they are hired at, and 0 if else. The experimental sample is of all new jobseekers who registered and were approved for the platform between June 8th and July 14th, 2021 and had non-empty resumes, for all jobs they worked within 28 days of registering for the platform. Significance indicators: $p \leq 0.10$: *, $p \leq 0.05$: ** and $p \leq .01$: ***.

Table 1.5: Effects of algorithmic writing assistance on hours worked and rehires

	<i>Dependent variable:</i>	
	Hours worked	Ever rehired
	(1)	(2)
Algo Writing Treatment	0.412 (0.303)	-0.003 (0.007)
Constant	2.649*** (0.214)	0.079*** (0.005)
Observations	194,700	6,263
R ²	0.00001	0.00003

Notes: This table analyzes the effect of the treatment on measures of hours worked and rehires. In Column (1) the outcome is the number of total hours worked by a worker in their first 28 days on the platform. In Column (2) the outcome is the fraction of workers who are ever rehired for different jobs by the same employer, conditional on jobseekers working at least one job. The experimental sample is of all new jobseekers who registered and were approved by the platform between June 8th and July 14th, 2021 and had non-empty resumes. Significance indicators: $p \leq 0.10$: *, $p \leq 0.05$: ** and $p \leq .01$: ***.

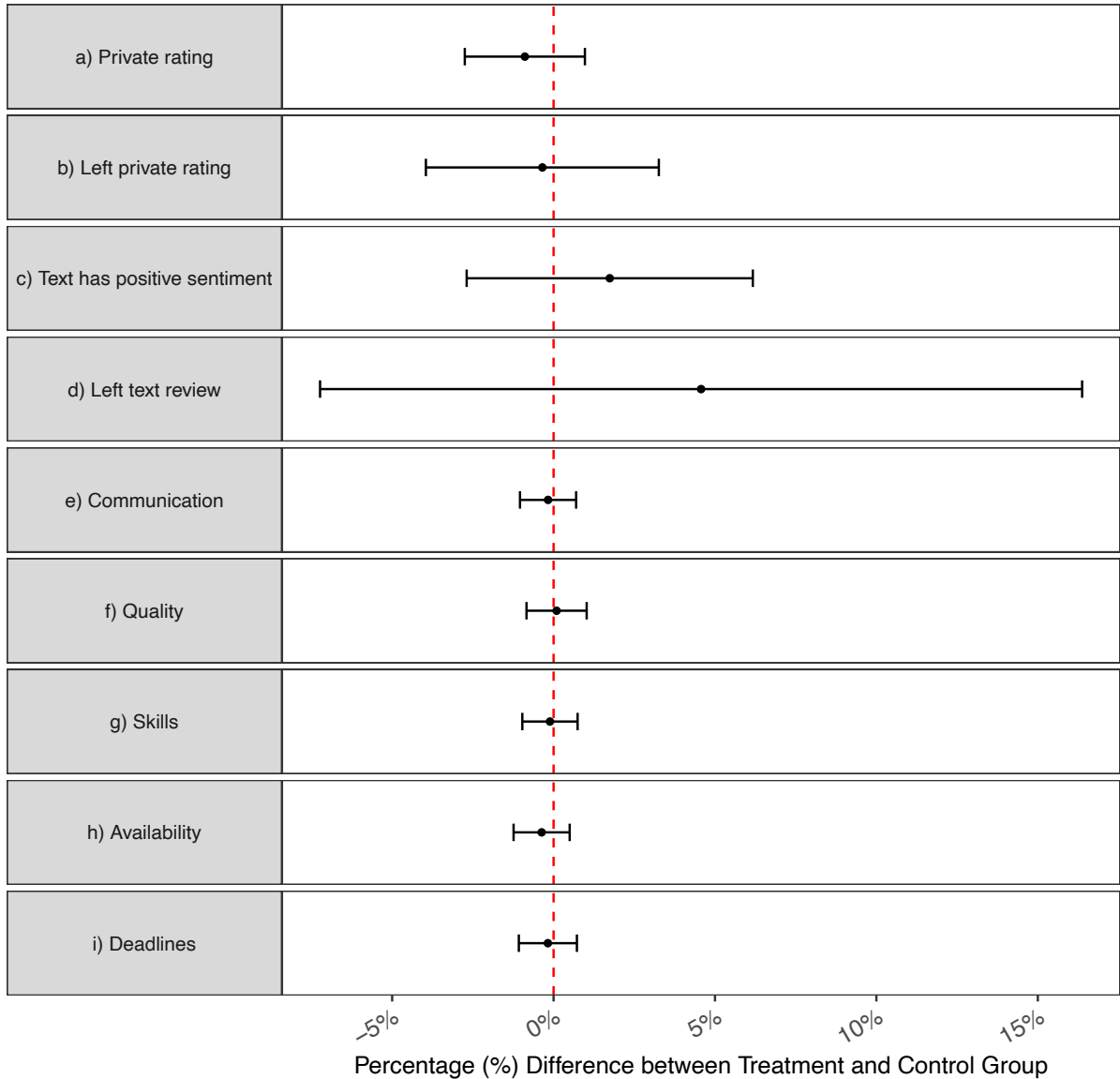
are ever rehired. Unlike the other outcomes, rehires are conditional on a worker being hired at least once over their first month on the platform. All jobseekers in this sample have been hired at least once, and the outcome “ever rehired” is 1 if the jobseeker is ever hired a second time by their first employer and 0 if they are only hired once. About 8% of all workers are rehired by the same employer at least once over the course of the experiment. This fraction does not differ in the treatment and control group.

1.4.3 Employers satisfaction was unaffected by the treatment

At the end of every contract, employers rate and review the workers by reporting both public and private rating to the platform. Private ratings are not shared with the worker. In the control group, workers had an average private rating of 8.63. In Figure 1-6a) we show that treated workers who formed any contracts over the experimental period did not have statistically different private ratings than workers in the control group. In Column (1) of Table 1.6 report the results from this regression. We show that workers in the treated group have an average private rating of 8.56 with a standard error of 0.08. We may also worry that if employers are less happy with the workers quality or productivity, that they may be more or less likely to leave a review at all. Figure 1-6b) we show that workers in the treatment group are not more or less likely to receive any rating than workers in the control group.

When the employers give these ratings they are also able to leave text reviews. While numerical ratings have become inflated in recent years, [Filippas et al. \(2022\)](#) show that the sentiments associated with the text of reviews has increased significantly less over time. This means that text reviews are likely more informative about the workers’ quality than the numerical ratings. We use a BERT text classification model ([HF canonical model maintainers, 2022](#)) to label each review as having positive or negative sentiment. These classifications are significantly correlated with the private ratings, with a Pearson correlation coefficient of 0.54. In Figure 1-6c) we show that the treated workers’ average text reviews are not statistically different from the average sentiment of the reviews for control workers. In Figure 1-6d) we show that workers in the treatment group are not more or less likely to receive any text review than workers in the control group. Results from these regressions can be found in Table 1.6.

Figure 1-6: Effect of algorithmic writing assistance on ratings



Notes: This analysis looks at the effect of treatment on ratings outcomes on jobseekers in the experimental sample. Private ratings are on a scale from one to ten. Communication, Quality, Skills, Availability, and Deadlines ratings are public and left as star ratings, on a scale from one to five. The x-axis is the difference in the mean outcome between jobseekers in the treated group and the control group. A 95% confidence interval based on standard errors calculated using the delta method is plotted around each estimate. The experimental sample is of all new jobseekers who registered and were approved for the platform between June 8th and July 14th, 2021, and had non-empty resumes that were hired in their first month on the platform, with $N = 4,250$. Regression details on private ratings and text reviews can be found in Table 1.6. Regression details on public ratings can be found in Appendix Table F17.

Table 1.6: Effect of algorithmic writing assistance on contract ratings

	<i>Dependent variable:</i>			
	Private rating	Positive text review	Left any rating	Left any text review
	(1)	(2)	(3)	(4)
Algo Writing Treatment	-0.077 (0.082)	0.015 (0.019)	-0.002 (0.012)	0.006 (0.008)
Constant	8.633*** (0.059)	0.859*** (0.014)	0.624*** (0.008)	0.138*** (0.006)
Observations	4,250	1,185	6,263	6,263
R ²	0.0002	0.001	0.00001	0.0001

Notes: This analysis looks at the effect of treatment on contract outcomes for jobseekers in the experimental sample. Column (1) defines private rating as the mean private rating on all jobs given by employers to the workers after the job ended, at the worker level. In Column (2) we take the text of the reviews left by employers on each job and use sentiment analysis (model: distilbert-base-uncased-finetuned-sst-2-english) to impute whether the review is positive, neutral, or negative, labeled one if it is positive or neutral. The outcome is the mean of these ratings over all contracts in the sample. Column (3) is the percentage of contracts worked where the freelancer received any private rating. And Column (4) is the percentage of contracts worked where the freelancer received any text based review. The experimental sample is of all new jobseekers who registered and were approved for the platform between June 8th and July 14th, 2021 and had non-empty resumes, for all jobs they worked within 28 days of registering for the platform. Significance indicators: $p \leq 0.10$: *, $p \leq 0.05$: ** and $p \leq .01$: ***.

In Figure 1-6e) through i) we report the results of the effect of the treatment on the employers' public ratings of the workers. Each outcome is a public rating the employers give to the workers at the end of a contract. Employers rate the workers communication, skills, quality of work, availability, cooperation, and ability to make deadlines. Each rating is given on a five point scale. There is less variation in the public ratings than in the private ones, and the average rating for each attribute is over 4.75 stars. Like the private ratings, there are no significant effects of the treatment to any of the ratings, including to workers' communication skills. And the point estimate of the treatment effect to the quality of the work done is even positive. Results from these regressions can be found in Appendix Table F17.

How much power do we have to detect worse contractual outcomes?

Given the null effect of the treatment to ratings, a natural question is how much power is available to detect effects. While we do find a substantial increase in hiring—8%—these marginal hires are mixed in with a much larger pool of “inframarginal” hires that would likely be hired anyway. How much worse could those marginal applicants have been and

still get our results to private ratings in the treatment?

Let I indicate “inframarginal” jobseekers who would have been hired in the treatment or control. Let M indicate “marginal” jobseekers who are only hired in the treatment. For workers in the control group, the average private rating will be $\bar{r}_C = \bar{r}_I$. But for the treatment, the mean rating is a mixture of the ratings for the inframarginal and the ratings for the induced, marginal applicants, and so

$$\bar{r}_T = \frac{\bar{r}_I + \tau \bar{r}_M}{1 + \tau} \tag{1.1}$$

where τ is the treatment effect. We assume no substitution, making our estimates conservative. The sampling distribution of the mean rating for the marginal group is

$$\bar{r}_M = \frac{\bar{r}_T(1 + \tau) - \bar{r}_C}{\tau} \tag{1.2}$$

Our course, \bar{r}_T , τ and \bar{r}_C are all themselves random variables. Furthermore, they are not necessarily independent. To compute the sampling distribution of \bar{r}_M , we bootstrap sample both the hiring regressions and the private feedback regressions on the experimental sample.¹⁴ Because we do not have feedback on workers who are never hired, we use the estimates values to calculate \bar{r}_M . Figure 1-7 shows the sampling distribution of \bar{r}_M .

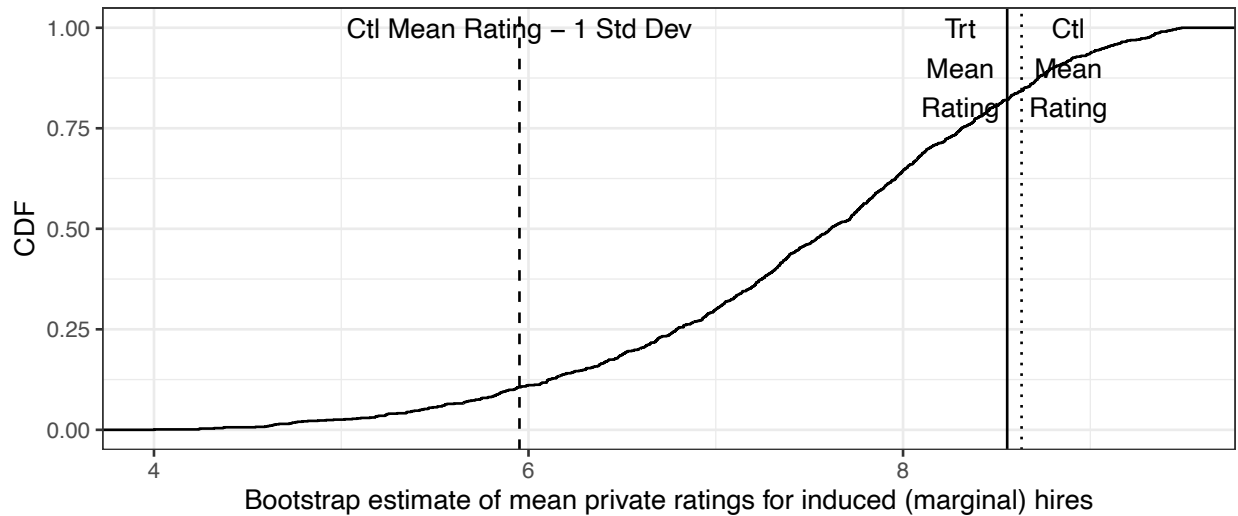
The treatment actual rating is plotted as a dotted line and control actual rating is plotted as a solid vertical line. The distribution is centered at these mean values.

The dashed line indicates the control mean rating minus one standard deviation in the private ratings (where the standard deviation is 2.4). Comparing this value to the distribution of \bar{r}_M , this value (at the dashed line) lies at less than 0.1 of the density. In short, it would be quite surprising for us to get the results we have—an 8% increase in hires and no different ratings if the actual marginal hires were a standard deviation worse.

Due to concerns about the loss of information in ratings caused by ratings inflation, it is reasonable to question the level of variation that could realistically be observed, even

¹⁴We define this sample as the workers allocated into the experiment who were approved by the platform and had non-empty resumes. From this we bootstrap sample with replacement. We run the hiring regressions on this sample and the ratings regressions on the same samples, narrowed to only those workers who were ever hired.

Figure 1-7: Sampling distribution of the private ratings of marginal hired jobseekers



in the presence of real effects. We do find variation in the ratings given to workers on the platform. In particular, workers with profiles written in a language other than English have an average private rating of 7.9 out of 10, which is lower than the average rating of 8.6 out of 10 for workers with profiles in English. Among workers with profiles in English, those based in the US have an average rating of 9.08 (with a standard deviation of 2.8), while workers from outside the US have an average rating of 8.46 (with a standard deviation of 2.14).

We can conduct a power analysis to determine the smallest effect size we could rule out with confidence. With 80% power and a 0.05 significance level, we can rule out any effects larger than 0.2 of a standard deviation. The overall standard deviation of ratings is 2.4, so an effect size of 0.2 standard deviations corresponds to a difference of 0.48 in ratings. This effect size is within the range of variation in ratings that we see within the data. Therefore, we can be reasonably confident that our study design would have been able to detect effects of practical significance.

1.4.4 Heterogeneous treatment effects to hiring and ratings

We have already shown above in Appendix Table F13 that the treatment disproportionately impacted the error rate in non-native English speakers' resumes. If we look downstream to

hiring outcomes, in Appendix Table F15, we interact the same groups with the treatment and look at their effect on an indicator for whether or not they were hired. While non-native English speakers' writing might benefit more from the treatment¹⁵, it does not translate into more hires relative to native English speakers. In fact, we actually see positive point estimates for effects to hiring for US and Anglophone workers, although these interaction effects are not significant. This may appear surprising, but it is important to remember that those workers are much more likely to be hired to begin with. Absent the treatment, the average worker from an Anglophone country is about twice as likely to ever be hired within their first 28 days on the platform. Because of this, in percentage terms, the treatment effect is actually larger for non-anglophone workers, 8.4%, than it is for anglophone workers whose treatment effect is 7.35%. These are not statistically different from each other, and both fit comfortably inside the 95% confidence interval on the hiring effect which is (3%,13%).

Lastly, in Appendix Table F18 we report the same specifications but look for heterogeneity in the effects to private ratings or whether the text of the review had a positive sentiment. These results are conditional on a hire, and therefore the point estimates are generally quite imprecise and we lack the power to conclude much. We can see from Column (2) that Anglophone workers are generally higher rated and Column (3) that US workers are as well. However neither have any additional effect when interacted with the treatment.

While our results suggest robust evidence for the “clarity view”, it is certainly possible that there are some types of work where ones writing ability is an important indicator of their on the job performance. We look specifically at jobseekers whose primary work is in writing in Appendix Table F14. Since workers who specialize in writing make up only 17% of the sample, the standard errors are too large to be able to very confidently say anything about the effect of the treatment to ratings. Therefore, we do not reject the possibility that the signaling view could be important in jobs where writing is an important part of the output.

¹⁵Workers from Anglophone countries have smaller treatment effects to their writing error rate in Table F13 than their Anglophone counterparts, but they still have significant positive treatment effects.

1.4.5 Robustness checks

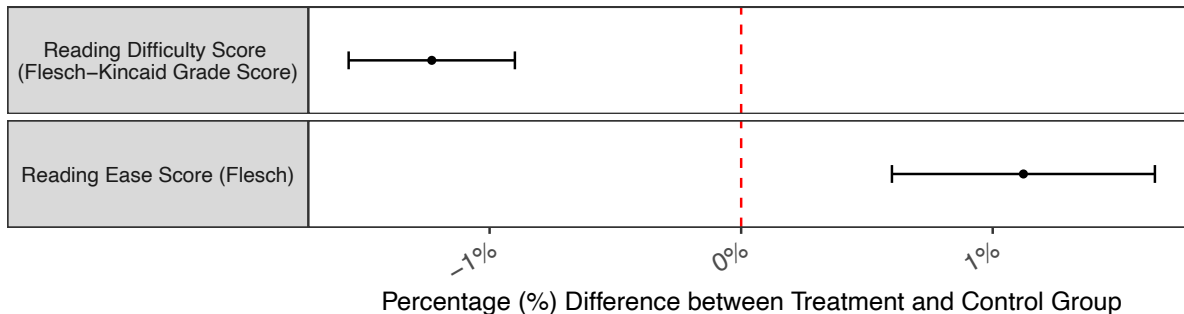
In our main analysis, we narrow the sample to only those jobseekers whose profiles were approved by the platform. In Appendix Table F11 we run a similar regression on the full experimental sample, but we include profile approval as a control to see if it affects the estimates. In this analysis, we find that the treatment effect on the number of hires is slightly smaller than in the analysis conditional on platform approval—conditioning the sample on only jobseekers whose profiles were approved has an estimate of 7.8% while it is 10% in the full sample. The effect on the probability of any hire is 8% in the sample of only approved jobseekers and 8% in the unconditional sample. This approach and narrowing the sample to only approved jobseekers would “block” the approval channel. In Appendix Table F12 we report the same analysis not conditioned on profile approval. None of these robustness checks change the direction or significance of any of the hiring estimates, and the slightly larger estimates in the unconditional sample are unsurprising because platform approval is a necessary condition for a jobseeker to be hired.

1.4.6 Direct tests of the clarity view

In order to provide evidence for our hypothesized mechanism, the clarity view, we use measures of readability from the statistics, psychology, and education literature to proxy for the clarity of the text of the resumes. In Figure 1-8 we report the effect of the treatment on two measures of readability. These measures are based on word length, number of syllables, and sentence length. In the first facet we use a measure of reading difficulty, the Flesch-Kincaid Grade Score (Kincaid et al., 1975). The Flesch-Kincaid Grade does not have bounds, but they roughly approximate grade levels, as in a score of 12 is text approximately at a 12th grade reading level. Higher scores imply text is more difficult to read. We see that the reading difficulty score is 1% lower in the treatment group, a small but significant effect.

Using another measure of readability, Flesch’s Reading Ease Score (Flesch, 1948), we find consistent effects— that the text of the resume’s in the treatment group is easier to read. This outcome is bounded by 1 and 100, with higher scores implying easier text to read. In the control group, the reading ease score is 39.7, and 40.2 in the treatment group.

Figure 1-8: Effect of treatment on measures of readability



Notes: This plot shows the effect of the treatment on the readability score and grade of the profiles. The first facet plots the Reading Difficulty Score, or the Flesch-Kincaid Grade Level Score. The Flesch-Kincaid Grade does not have bounds, but they roughly approximate grade levels. Higher scores imply text is more difficult to read. In the second facet is plotted the Flesch Reading Ease Score, which is a score between 0 and 100 where the higher the score the easier it is to read. The experimental sample is of all new jobseekers who registered and were approved for the platform between June 8th and July 14th, 2021, and had non-empty resumes, with $N = 194,700$. See Appendix Table F19 for the regression table.

While these effects are small, they consistently show that the resumes in the treatment group are easier to understand, and these measures have been used across scientific fields to understand the readability of writing (Singh et al., 2017; Alvero et al., 2021).

1.4.7 What happens in general equilibrium?

A first order question for platform owners and hiring managers is whether these effects would hold up as a market wide policy. As with any experimentally allocated labor market intervention, it is possible that increase in the number of workers hired does not reflect an increase in the supply or demand for labor, but instead reflect employers substituting a worker in the control group or outside of the experiment for one in the treatment group. Crowd-out concerns have been shown to be important with labor market assistance (Crépon et al., 2013).

In order to test how much of the benefits to treated workers came at the expense of other workers, in Figure 1-9 we break down the treatment effect by how much a jobseeker on average competed with jobseekers in the treatment group. Here create a measure of average competition with treated jobseekers. We take each job and count the number of treated jobseekers that apply. For jobseekers in the control group, we calculate the average number of treated jobseekers that apply to the jobs they apply to. To calculate the number of

Figure 1-9: Treatment Effects by Exposure to Treated Jobseekers, with ATE in blue



Notes: This analysis looks at the effect of treatment on whether or not jobseekers are hired within one month. On the x-axis we break down the jobseekers into quantiles based on the average number of treated jobseekers they compete with when they apply to jobs. The 1st quantile is jobseekers who on average apply to jobs which no more than one other treated jobseeker applied to. The 5th quantile is jobseekers who on average apply to jobs which receive 6 or more other applications from treated jobseekers. A 95% confidence interval is plotted around each estimate. The experimental sample is of all new jobseekers who registered and were approved for the platform between June 8th and July 14th, 2021, and had non-empty resumes, with $N = 194,700$.

treated competitors for jobseekers in the treatment group, we count the number of treated jobseekers that apply to the jobs they apply to, minus one. On the x-axis of Figure 1-9 we break down the jobseekers into quantiles based on the average number of treated competitors they have. Jobseekers in the first quantile have one or fewer treated competitors on average, while jobseekers in the fifth quantile have more than six treated competitors on average. We find that the treatment effect diminishes based on how exposed a jobseeker is to treated jobseekers. In the first quantile, the treatment effect is a full percentage point, or more than a 30% increase in the likelihood of being hired within a month on the platform. By the third quantile, the treatment effect is almost exactly the average treatment effect of 8%, although the effect is insignificant. And for those jobseekers who compete the most with other treated jobseekers, the effect is small and insignificant.

1.5 Conclusion

Employers are more likely to hire new labor market entrants with better-written and clearer resumes. We argue that better writing makes it easier for employers to judge the match quality of a particular worker. We show results from a field experiment in an online labor market where treated workers were given algorithmic writing assistance. These jobseekers were 8% more likely to get hired and formed 7.8% more contracts over the month-long experiment. These jobseekers were hired at 10% higher wages than those in the control group, due to a change in the composition of which workers were hired. While one might have expected writing quality to be a reliable indicator of worker quality, the treatment did not affect employers' ratings of hired workers. We provide evidence for “the clarity view” of resume writing—that better writing, without any difference in the underlying facts—makes it easier for employers to correctly judge an applicants abilities.

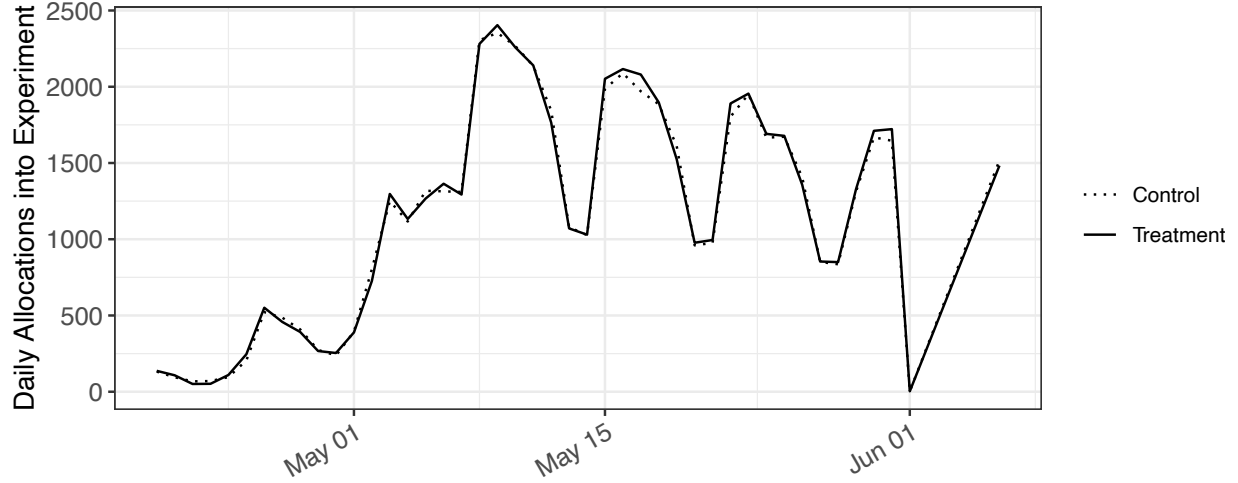
These results have important implications for hiring managers and for platform designers. The change in the composition of hired workers to more expensive workers implies that if this technology was rolled out platform wide, it would increase platform revenue. As for the increase in number of hires, it is possible that the benefits to treated workers came at the expense of other workers, as both treated- and control-assigned workers compete in the same market. We find evidence that the treatment effect dissipates the more treated jobseekers one is in competition with. This suggests that the additional hires driven by the treatment might be crowding out other hires. However, even if additional hires came from experienced workers, this is likely still a positive result. New labor market entrants are uniquely disadvantaged ([Pallais, 2014](#)) in online labor markets. To the extent that the gains to new workers come partially at the expense of experienced workers, this is likely a good trade-off. And lastly, given the wages of the hired workers are higher with no lower ratings, when rolled out platform-wide, algorithmic writing assistance is likely to increase the quality of matches formed.

Conceptualizing AI/ML innovation and proliferation as a fall in the cost of prediction technology fits our setting ([Agrawal et al., 2018a,b](#)). Writing a resume is, in part, an applied prediction task—what combination of words and phrases, arranged in what order, are likely

to maximize my pay-off from a job search? The Algorithmic Writing Service reduces the effort or cost required for making these decisions. When revising their resumes, rather than identifying errors in their own predictions themselves, jobseekers with access to the Algorithmic Writing Service are given suggestions for error correction and cleaned up writing. Furthermore, the treatment, by lowering the costs of error-free writing for at least some jobseekers, causes them to do better at writing their resumes.

These kinds of algorithmic writing assistance will likely “ruin” writing as a signal of ability. With advances in writing technologies with capabilities far beyond what is explored here ([Brown et al., 2020](#)), even if the “signaling view” was at one time dominant, the proliferation of Large Language Models are likely to make it not true in the future.

Figure 1-1: Daily allocations of jobseekers into experimental cells



Notes: This plot shows the daily allocations into the treatment and control cells for the experimental sample of 480,948 new jobseekers to the platform.

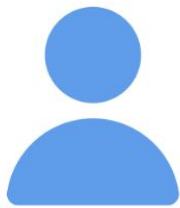
1.6 Appendix

1.6.1 Summary statistics and descriptives

Table F1: Summary statistics of jobs worked in the control group

	Mean	Std Deviation
Length of job in days	59	(114)
Fraction of jobs that are hourly	0.34	(0.47)
Total hours worked	201	(482)
Earnings from hourly jobs (\$)	2,524	(2,524)
Earnings from fixed-price jobs (\$)	386	(386)
Min hourly rate (\$)	17.04	(18.42)
Max hourly rate (\$)	17.22	(19)

Notes: This table reports summary statistics for jobs worked by workers in the control group of the experiment. The experimental sample is of all new jobseekers who registered and were approved for the platform between June 8th and July 14th, 2021 and had non-empty resumes. The sample of jobs include any job where the worker was hired between June 08,2021 and August 14, 2021, with N = 4,521.



Aaron P.

Indianapolis, USA

100%

Average rating

Django/Flask/Rails Web Developer

I have a B.S. in Industrial and Systems Engineering and having been doing entrepreneurial web development for 15 years. I am a full-stack web developer specializing in back-end development with Python frameworks (Django and Flask). I am also proficient with Ruby on Rails.

I have worked with many small startups as well as Fortune 50 companies in capacities ranging from developer to product manager.

I have extensive experience building and interacting with RESTful API's, data modeling, and project management.

Some services I have... [See more](#)

\$90.00

Hourly rate

\$300k+

Earnings

23

Jobs

5,059

Hours worked

Figure 1-2: Stylized version of a workers' resume on the online labor market

Table F2: Summary statistics on error counts and rates in the control group

	Total Errors	Error Rate
All Error Types	4.390 (10.257)	0.080 (0.252)
Capitalization Errors	0.140 (0.543)	0.004 (0.029)
Possible Typo	2.312 (8.768)	0.041 (0.106)
Grammar Errors	0.219 (0.589)	0.004 (0.014)
Punctuation Errors	0.425 (1.629)	0.008 (0.213)
Typographic Errors	0.821 (2.985)	0.016 (0.051)
Style Errors	0.317 (0.890)	0.004 (0.011)
Miscellaneous Errors	0.103 (0.391)	0.002 (0.009)
Redundant Phrases	0.025 (0.162)	0.000 (0.003)
Nonstandard Phrases	0.002 (0.050)	0.000 (0.001)
Commonly Confused Words	0.010 (0.117)	0.000 (0.002)
Collocations	0.012 (0.121)	0.000 (0.003)
Semantic Errors	0.003 (0.061)	0.000 (0.001)

Notes: This table reports means and standard errors of the writing errors in the resumes of the control group. The first column displays the average total error count and the second column displays the average error rate (total errors normalized by word count). Writing errors are defined by LanguageToolR. The sample is made up of all jobseekers in the control group of the experimental sample who submitted non-empty resumes and were approved by the platform.

Table F3: Effects of writing assistance on profile submission and platform approval

	<i>Dependent variable:</i>		
	Profile submitted x 100	Approved x 100	
	(1)	(2)	(3)
Algo Writing Treatment	0.106 (0.144)	0.199 (0.133)	0.186 (0.142)
Constant	45.532*** (0.102)	89.057*** (0.094)	40.550*** (0.100)
Observations	480,948	219,242	480,948
R ²	0.00000	0.00001	0.00000

Notes: This table analyzes the effect of the treatment on whether or not a jobseeker's profile was submitted and approved. In Column (1) the outcome is 100 times a binary indicator for whether or not the jobseeker completed platform registration and submitted their resume, on the full experimental sample. In Column (2) the outcome is 100 times a binary indicator for whether or not the platform approved the resume, on the sample of only those jobseekers who submitted their resumes. In Column (3) the outcome is 100 times a binary indicator for whether or not the platform approved the resume, on the full experimental sample. Significance indicators: $p \leq 0.10$: *, $p \leq 0.05$: ** and $p \leq .01$: ***.

Table A4: Description of Error Rule Categories with Examples

Category	Description	Examples
American English Phrases	Sentence favors the American English spelling of words.	<i>apologize, catalog, civilization, defense</i>
British English, Oxford Spelling	Sentence favors the British English spelling of words.	<i>apologise, catalogue, civilisation, defence</i>
Capitalization	Rules about detecting uppercase words where lowercase is required and vice versa.	<i>This house is old. it was built in 1950. I really like Harry potter.</i>
Collocations	A collocation is made up of two or more words that are commonly used together in English. This refers to an error in this type of phrase.	<i>Undoubtedly, this is the result of an extremely dynamic development of Lublin in the recent years. I will take it in to account. It's batter to be save then sorry.</i>
Commonly Confused Words	Words that are easily confused, like 'there' and 'their' in English.	<i>I have my won bed. Their elicit behavior got the students kicked out of school.It's the worse possible outcome.</i>
Grammar	Violations related to system of rules that allow us to structure sentences.	<i>Tom make his life worse. A study like this one rely on historical and present data.This is best way of dealing with errors.</i>
Miscellaneous	Miscellaneous rules that do not fit elsewhere.	<i>This is best way of dealing with errors. The train arrived a hour ago. It's nice, but it doesn't work. (inconsistent apostrophes)</i>
Nonstandard Phrases		<i>I never have been to London. List the names in an alphabetical order. Why would a man all of the sudden send flowers?</i>
Possible Typo	Spelling issues.	<i>It'a extremely helpful when it comes to homework. We haven't earned anything.This is not a HIPPA violation.</i>
Punctuation	Error in the marks, such as period, comma, and parentheses, used in writing to separate sentences and their elements and to clarify meaning.	<i>"I'm over here, she said. Huh I thought it was done already. The U.S.A is one of the largest countries.</i>
Redundant Phrases	Redundant phrases contain words that say the same thing twice. When one of the words is removed, the sentence still makes sense. Sometimes the sentence has to be slightly restructured, but the message remains the same.	<i>We have more than 100+ customers. He did it in a terrible way. The money is sufficient enough to buy the sweater.</i>
Semantics	Logic, content, and consistency problems.	<i>It allows us to both grow, focus, and flourish. On October 7, 2025 , we visited the client.This was my 11nd try.</i>
Style	General style issues not covered by other categories, like overly verbose wording.	<i>Moreover, the street is almost entirely residential. Moreover, it was named after a poet. Doing it this way is more easy than the previous method. I'm not very experienced too. Anyways, I don't like it.</i>
Typography	Problems like incorrectly used dash or quote characters.	<i>This is a sentence with two consecutive spaces. I have 3dogs.The price rose by \$12,50. I'll buy a new T—shirt.</i>

Table F4: Hiring outcomes predicted based on language errors (normalized by word count) in observational data

	<i>Dependent variable:</i>			
	Hired (1)	Number of Contracts (2)	Hired (3)	Number of Contracts (4)
Capitalization Error	-0.038 (0.025)	-0.075 (0.048)	-0.026 (0.023)	-0.055 (0.045)
Possible Typo	-0.022*** (0.007)	-0.030** (0.013)	-0.013** (0.006)	-0.016 (0.012)
Grammar Error	-0.314*** (0.051)	-0.534*** (0.097)	-0.210*** (0.047)	-0.360*** (0.092)
Punctuation Error	0.0002 (0.003)	-0.0001 (0.006)	0.0001 (0.003)	-0.0002 (0.006)
Typography Error	-0.069*** (0.014)	-0.098*** (0.026)	-0.050*** (0.013)	-0.066*** (0.025)
Style Error	0.130** (0.062)	0.261** (0.119)	0.115** (0.058)	0.234** (0.112)
Miscellaneous Error	-0.220*** (0.079)	-0.414*** (0.151)	-0.121 (0.074)	-0.252* (0.143)
Redundant Phrases	-0.264 (0.229)	-0.433 (0.437)	-0.149 (0.213)	-0.240 (0.414)
Nonstandard Phrases	-0.124 (0.882)	0.804 (1.681)	-0.193 (0.819)	0.699 (1.591)
Commonly Confused Words	-0.331 (0.324)	-0.761 (0.618)	-0.190 (0.301)	-0.531 (0.584)
Collocations	-0.380* (0.228)	-0.637 (0.434)	-0.262 (0.211)	-0.438 (0.411)
Semantic Error	-0.532 (0.583)	-0.340 (1.112)	-0.445 (0.541)	-0.191 (1.052)
Constant	0.036*** (0.001)	0.053*** (0.002)	0.026*** (0.001)	0.036*** (0.002)
Controls			X	X
Observations	65,114	65,114	65,114	65,114
R ²	0.002	0.001	0.140	0.106

Notes: This table analyzes correlation between various writing errors on jobseekers' resumes and their hiring outcomes. The independent variables, writing errors, are divided by the number of words in the jobseekers' resume. Hired is defined as 1 if the jobseeker was ever hired in their first month after registering for the platform, and 0 if else. Number of Contracts is defined as the number of unique jobs they begin working in that time. Columns (3) and (4) include controls for profile hourly rate and job category. Writing errors are defined by LanguageToolR. The sample is made up of all jobseekers who registered for the platform in the week before the experiment who submitted non-empty resumes.

Significance indicators: $p \leq 0.10$: *, $p \leq 0.05$: ** and $p \leq .01$: ***.

Table F5: Hiring outcomes predicted based on language errors in observational data

	<i>Dependent variable:</i>							
	Hired		Number of Contracts		Hired		Number of Contracts	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Total errors	-0.0002** (0.0001)		-0.0003** (0.0001)		-0.0001 (0.0001)		-0.0002 (0.0001)	
Num words	0.0002*** (0.00001)		0.0004*** (0.00002)		0.0001*** (0.00001)		0.0003*** (0.00002)	
Error rate		-0.011*** (0.003)		-0.017*** (0.005)		-0.007*** (0.003)		-0.010** (0.005)
Constant	0.017*** (0.001)	0.033*** (0.001)	0.020*** (0.002)	0.049*** (0.001)	0.014*** (0.001)	0.024*** (0.001)	0.014*** (0.002)	0.034*** (0.001)
Normalized		X		X		X		X
Controls					X	X	X	X
Observations	65,114	65,114	65,114	65,114	65,114	65,114	65,114	65,114
R ²	0.007	0.0002	0.007	0.0002	0.142	0.139	0.108	0.105

Notes: This table analyzes correlation between all writing errors on jobseekers' resumes and their hiring outcomes. The first independent variable is the total number of writing errors on a jobseekers' resume. The second independent variable is the total number of errors divided by the length of their resume, in number of words. Hired is defined as 1 if the jobseeker was ever hired in their first month after registering for the platform, and 0 if else. Number of Contracts is defined as the number of unique jobs they begin working in that time. Columns (5) through (8) include controls for profile hourly rate and job category. Writing errors are defined by LanguageToolR. The sample is made up of all jobseekers who registered for the platform in the week before the experiment who submitted non-empty resumes.

Significance indicators: $p \leq 0.10$: *, $p \leq 0.05$: ** and $p \leq .01$: ***.

Table F6: Hiring outcomes predicted based on language errors in the observational data

	<i>Dependent variable:</i>			
	Hired	Number of Contracts	Hired	Number of Contracts
	(1)	(2)	(3)	(4)
Number of words	0.0002*** (0.00001)	0.0004*** (0.00002)	0.0001*** (0.00001)	0.0003*** (0.00002)
Capitalization Error	-0.002* (0.001)	-0.006*** (0.002)	-0.001 (0.001)	-0.004* (0.002)
Possible Typo	-0.00001 (0.0001)	-0.00000 (0.0002)	0.00003 (0.0001)	0.0001 (0.0002)
Grammar Error	-0.007*** (0.001)	-0.012*** (0.002)	-0.004*** (0.001)	-0.008*** (0.002)
Punctuation Error	0.001*** (0.0004)	0.002* (0.001)	0.001* (0.0004)	0.0004 (0.001)
Typography Error	-0.001*** (0.0002)	-0.001** (0.0005)	-0.001*** (0.0002)	-0.001* (0.0004)
Style Error	0.003*** (0.001)	0.006*** (0.002)	0.003*** (0.001)	0.006*** (0.001)
Miscellaneous Error	-0.007*** (0.002)	-0.011*** (0.003)	-0.003** (0.002)	-0.005 (0.003)
Redundant Phrases	-0.003 (0.004)	-0.009 (0.008)	-0.001 (0.004)	-0.006 (0.008)
Nonstandard Phrases	-0.001 (0.014)	0.009 (0.026)	-0.003 (0.013)	0.005 (0.025)
Commonly Confused Words	-0.008 (0.006)	-0.021* (0.011)	-0.003 (0.006)	-0.015 (0.011)
Collocations	-0.003 (0.006)	-0.013 (0.011)	-0.003 (0.005)	-0.013 (0.010)
Semantic Error	0.004 (0.011)	0.035 (0.021)	0.002 (0.010)	0.031 (0.020)
Constant	0.018*** (0.001)	0.022*** (0.002)	0.015*** (0.001)	0.015*** (0.002)
Controls			X	X
Observations	65,114	65,114	65,114	65,114
R ²	0.009	0.008	0.142	0.109

Notes: This table analyzes correlation between various writing errors on jobseekers' resumes and their hiring outcomes. Hired is defined as 1 if the jobseeker was ever hired in their first month after registering for the platform, and 0 if else. Number of Contracts is defined as the number of unique jobs they begin working in that time. Columns (3) and (4) include controls for profile hourly rate and job category. Writing errors are defined by LanguageToolR. The sample is made up of all jobseekers who registered for the platform in the week before the experiment who submitted non-empty resumes.

Significance indicators: $p \leq 0.10$: *, $p \leq 0.05$: ** and $p \leq .01$: ***.

Table F7: Effects of writing assistance on length of resume

	<i>Dependent variable:</i>
	Number of words in resume
Algo Writing Treatment	0.127 (0.314)
Constant	70.541*** (0.223)
Observations	194,700
R ²	0.00000

Notes: This table analyzes the effect of the treatment on the number of words in a jobseeker's resume. The sample is made up of all jobseekers in the experimental sample who submitted non-empty profiles and were approved by the platform. Significance indicators: $p \leq 0.10$: *, $p \leq 0.05$: ** and $p \leq .01$: ***.

Table F8: Effect of Treatment on Writing Errors, Page 1

	<i>Dependent variable:</i>					
	Capitalization (1)	Possible Typo (2)	Grammar (3)	Punctuation (4)	Typography (5)	Style (6)
Algo Writing Treatment	-0.0005*** (0.0001)	-0.002*** (0.0005)	-0.0005*** (0.0001)	-0.0004 (0.0004)	-0.002*** (0.0003)	0.0003*** (0.0001)
Constant	0.003*** (0.00005)	0.041*** (0.0003)	0.004*** (0.00004)	0.010*** (0.0003)	0.015*** (0.0002)	0.004*** (0.00004)
Observations	194,700	194,700	194,700	194,700	194,700	194,700
R ²	0.0003	0.0001	0.0004	0.0000	0.0004	0.0001

Notes: This table analyzes the effect of the treatment on all types of writing errors, normalized by resume length. Writing errors are defined by LanguageToolR, and divided by the number of words in a jobseekers' resume to calculate their error rate. The sample is made up of all jobseekers in the experimental sample who completed the platform registration page and submitted non-empty resume. Significance indicators: $p \leq 0.10$: *, $p \leq 0.05$: **, and $p \leq .01$: ***.

Table F9: Effect of Treatment on Writing Errors, Page 2

<i>Dependent variable:</i>						
	Redundant Phrases (1)	Nonstandard Phrases (2)	Commonly Confused Words (3)	Collocations (4)	Semantics (5)	get(vars[12]) (6)
Algo Writing Treatment	-0.0004*** (0.00003)	-0.00003* (0.00001)	-0.00001 (0.00000)	-0.00005*** (0.00001)	-0.0001*** (0.00001)	0.00001 (0.00001)
Constant	0.002*** (0.00002)	0.0004*** (0.00001)	0.00003*** (0.00000)	0.0001*** (0.00001)	0.0003*** (0.00001)	0.0001*** (0.00000)
Observations	194,700	194,700	194,700	194,700	194,700	194,700
R ²	0.001	0.00002	0.00001	0.0001	0.0001	0.00000

Notes: This table analyzes the effect of the treatment on all types of writing errors, normalized by resume length. Writing errors are defined by LanguageToolR, and divided by the number of words in a jobseekers' resume to calculate their error rate. The sample is made up of all jobseekers in the experimental sample who completed the platform registration page and submitted non-empty resume. Significance: $p \leq 0.10$: †, $p \leq 0.05$: *, $p \leq 0.01$: ** and $p \leq .001$: ***.

Table F10: Bonferroni Corrected Standard Errors for Figure 1-4

	Unadjusted SE	Bonferroni SE
Capitalization Errors	0.000	0.000
Possible Typo	0.0003	0.0034
Grammar Errors	0.000	0.000
Typographic Errors	0.000	0.000
Style Errors	0.000	0.000
Miscellaneous Errors	0.000	0.000
Redundant Phrases	0.056	0.556
Collocations	0.00002	0.00020
Commonly Confused Words	0.000	0.000
Semantic Errors	0.442	1.000

Table F11: Effect of algorithmic writing assistance on hiring outcomes, controlling for platform approval

	<i>Dependent variable:</i>				
	Num Invitations	Num Shortlists	Num Interviews	Hired x 100	Num Contracts
	(1)	(2)	(3)	(4)	(5)
Algo Writing Treatment	0.0004 (0.001)	0.001 (0.001)	0.002 (0.002)	0.100*** (0.032)	0.001** (0.001)
Approved by Platform	0.142*** (0.001)	0.040*** (0.001)	0.179*** (0.002)	3.204*** (0.033)	0.049*** (0.001)
Constant	0.00001 (0.001)	-0.0003 (0.0005)	-0.001 (0.001)	-0.050* (0.027)	-0.001 (0.001)
Observations	480,948	480,948	480,948	480,948	480,948
R ²	0.023	0.010	0.024	0.019	0.011

Notes: This analysis looks at the effect of treatment on hiring outcomes on jobseekers in the experimental sample. The Column (1) outcome Invitations is the number of times they were recruited to a job over their first month. Column (2) is the number of times their application was shortlisted over that month. Column (3) is the number of interviews they gave over that month. Column (4) defines Hired x 100 as one hundred times the probability the jobseeker was hired over that month. Column (5) defines Number of Contracts as the number of unique jobs they work over the month after they register for the platform.

The sample used in this analysis is the entire experimental sample. Significance indicators: $p \leq 0.10$: *, $p \leq 0.05$: ** and $p \leq .01$: ***.

Table F12: Effect of algorithmic writing assistance on hiring outcomes, unconditional on platform approval

<i>Dependent variable:</i>					
	Num Invitations (1)	Num Shortlists (2)	Num Interviews (3)	Hired x 100 (4)	Num Contracts (5)
Algo Writing Treatment	0.001 (0.001)	0.001 (0.001)	0.002 (0.002)	0.106*** (0.033)	0.002** (0.001)
Constant	0.058*** (0.001)	0.016*** (0.0004)	0.072*** (0.001)	1.249*** (0.023)	0.019*** (0.0005)
Observations	480,948	480,948	480,948	480,948	480,948
R ²	0.00000	0.00001	0.00000	0.00002	0.00001

Notes: This analysis looks at the effect of treatment on hiring outcomes on jobseekers in the experimental sample. The Column (1) outcome Invitations is the number of times they were recruited to a job over their first month. Column (2) is the number of times their application was shortlisted over that month. Column (3) is the number of interviews they gave over that month. Column (4) defines Hired x 100 as one hundred times the probability the jobseeker was hired over that month. Column (5) defines Number of Contracts as the number of unique jobs they work over the month after they register for the platform. The sample used in this analysis is the entire experimental sample. Significance indicators: $p \leq 0.10$: *, $p \leq 0.05$: ** and $p \leq .01$: ***.

Table F13: Effects of writing assistance on error rate, by sub-groups

	<i>Dependent variable:</i>		
	Total Error rate x 100		
	(1)	(2)	(3)
Algo Writing Treatment	-0.512*** (0.070)	-0.628*** (0.077)	-0.594*** (0.075)
Anglophone Country		-4.555*** (0.127)	
Trt × Anglo		0.580*** (0.179)	
US			-4.469*** (0.141)
Trt × US			0.516*** (0.200)
Constant	7.680*** (0.050)	8.531*** (0.055)	8.321*** (0.053)
Observations	194,700	194,700	194,700
R ²	0.0003	0.012	0.009

Notes: In Column (1) we show the overall effect of the treatment to one minus the number of errors on a jobseekers' resume divided by the number of words. In Column (2) we interact the treatment with a dummy variable for if the jobseeker is from the US, UK, Canada, or Australia. In Column (3) we interact the treatment with a dummy for if the jobseeker is in the US. The experimental sample is of all new jobseekers who registered and were approved by the platform between June 8th and July 14th, 2021 and had non-empty resumes. Significance indicators: $p \leq 0.10$: *, $p \leq 0.05$: ** and $p \leq .01$: ***.

Table F14: Effect of algorithmic writing assistance on writers

	<i>Dependent variable:</i>			
	Error Rate X 100	Hires X 100	Private rating	Positive text review
	(1)	(2)	(3)	(4)
Algo Writing Treatment	-0.580*** (0.139)	0.190 (0.184)	-0.053 (0.214)	-0.064 (0.060)
Constant	7.086*** (0.098)	2.855*** (0.130)	8.456*** (0.153)	0.874*** (0.046)
Observations	33,907	33,907	672	149
R ²	0.001	0.00003	0.0001	0.008

Notes: This analysis looks at the effect of treatment on contract outcomes for jobseekers in the experimental sample whose primary job category is listed as Writing. In Column (1) the outcome is 100 times the error rate based on any error type in their resume. In Column (2) the outcome is 100 times whether or not the jobseeker ever is hired over their first month on the platform. In Column (3) the outcome is the mean private rating of jobseeker for any jobs they work in their first month on the platform. In Column (4) we take the text of the reviews left by employers on each job and use sentiment analysis (model: distilbert-base-uncased-finetuned-sst-2-english) to impute whether the review is positive, neutral, or negative, labeled one if it is positive or neutral. The outcome is the mean of these ratings over all contracts in the sample. The experimental sample is of all new jobseekers who registered and were approved for the platform between June 8th and July 14th, 2021 and had non-empty resumes, for all jobs they worked within 28 days of registering for the platform. Significance indicators: $p \leq 0.10$: *, $p \leq 0.05$: ** and $p \leq .01$: ***.

Table F15: Effects of writing assistance on hiring, by sub-groups

	<i>Dependent variable:</i>		
	Hired x 100		
	(1)	(2)	(3)
Algo Writing Treatment	0.247*** (0.080)	0.218** (0.088)	0.242*** (0.086)
Anglophone Country		2.486*** (0.145)	
Trt × Anglo		0.187 (0.205)	
US			2.602*** (0.161)
Trt × US			0.072 (0.228)
Constant	3.093*** (0.057)	2.629*** (0.063)	2.719*** (0.061)
Observations	194,700	194,700	194,700
R ²	0.00005	0.003	0.003

Notes: This table analyzes the effect of the treatment on whether or not a jobseeker was ever hired on the platform in the month after they joined, times 100. In Column (1) we show the overall effect of the treatment to hiring. In Column (2) we interact the treatment with a dummy variable for if the jobseeker is from the US, UK, Canada, or Australia. In Column (3) we interact the treatment with a dummy for if the jobseeker is in the US. The experimental sample is of all new jobseekers who registered and were approved by the platform between June 8th and July 14th, 2021 and had non-empty resumes. Significance indicators: $p \leq 0.10$: *, $p \leq 0.05$: ** and $p \leq .01$: ***.

1.6.2 Effect of the treatment on workers' earnings

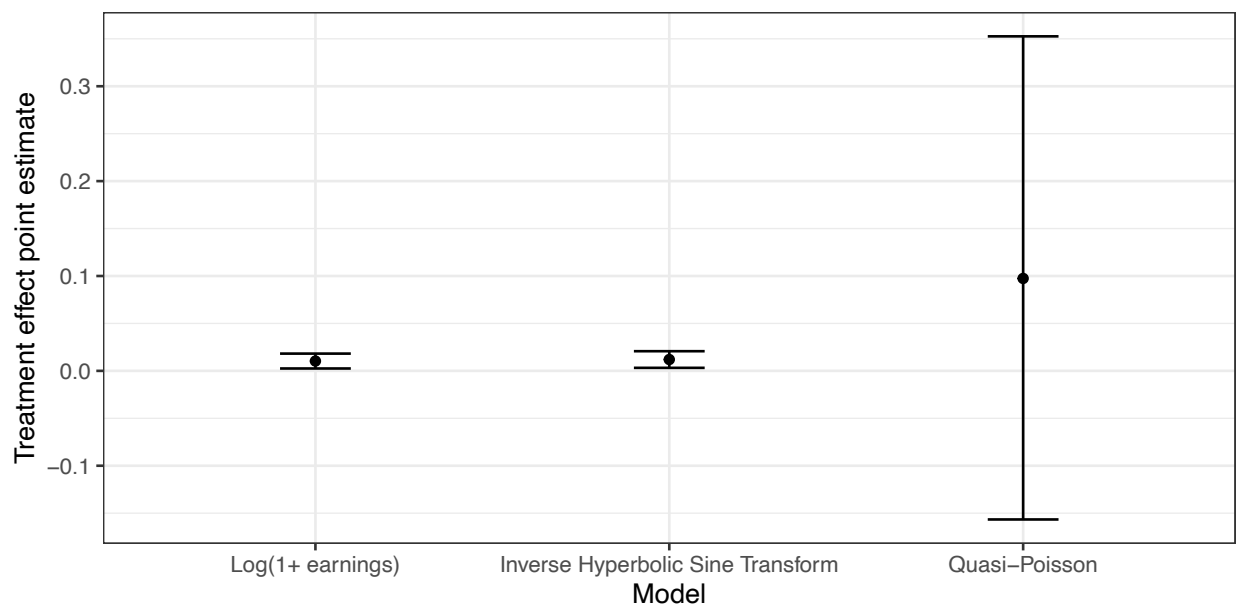
As further robustness to test whether workers in the treatment group underperformed to expectation, we regress treatment on various measures of worker's earnings in their first 28 days on the platform. In Appendix Table F16 Column (2) the outcome is workers' log of 1 plus hourly earnings. As with total hours, there are a lot of zeros, but among workers who have any hourly earnings, the average in the control group is \$1,529 in their first month. In Column (3) the outcome is the log of 1 plus the sum of workers earnings from both hourly and fixed price jobs, unconditional on ever being hired. Among workers who have any earnings, the average in the control group is \$2,957. Small positive effects on hourly earnings are mechanical due to the increase in hourly wages, and there does not seem to be any additional earnings effect to fixed price jobs. Because of the large number of jobseekers who have zero earnings, we report results from specifications which deal better with overdispersion in Appendix Figure 1-3. The treatment effect estimates are all small and positive, but point estimates are sensitive to specification.

Table F16: Effects of algorithmic writing assistance on workers' earnings

	<i>Dependent variable:</i>	
	Log hourly earnings	Log total earnings
	(1)	(2)
Algo Writing Treatment	0.010*** (0.003)	0.010*** (0.004)
Constant	0.061*** (0.002)	0.134*** (0.003)
Observations	194,700	194,700
R ²	0.0001	0.00003

Notes: This table analyzes the effect of the treatment on measures of workers' earnings. In Column (1) the outcome is the log hourly earnings a worker is paid over this time period. In Column (2) the outcome is the log total earnings a worker is paid over this time period. The experimental sample is of all new jobseekers who registered and were approved by the platform between June 8th and July 14th, 2021 and had non-empty resumes. Significance indicators: $p \leq 0.10$: *, $p \leq 0.05$: ** and $p \leq .01$: ***.

Figure 1-3: Robustness tests for total earnings treatment effect



Notes: This plot shows the effect of the treatment on the total earnings on all contracts from the workers first month on the platform. We report results from specifications meant to deal with the fact that most of the observations are zeros. The first specification is an OLS regression where the outcome is $\log(1 + \text{earnings})$, the second specification is OLS where the earnings variable has had an Inverse Hyperbolic Sine transformation. The third is a quasi-poisson generalized linear model. The experimental sample is of all new jobseekers who registered and were approved for the platform between June 8th and July 14th, 2021, and had non-empty resumes, with $N = 194,700$.

Table F17: Effect of algorithmic writing assistance on workers' public ratings

	<i>Dependent variable:</i>					
	Communication (1)	Skills (2)	Quality (3)	Cooperation (4)	Deadlines (5)	Received public rating (6)
Algo Writing Treatment	-0.008 (0.021)	-0.005 (0.021)	0.004 (0.023)	-0.008 (0.020)	-0.009 (0.022)	0.002 (0.012)
Constant	4.811*** (0.015)	4.801*** (0.015)	4.768*** (0.016)	4.840*** (0.015)	4.803*** (0.016)	0.534*** (0.009)
Observations	3,745	3,745	3,745	3,745	3,745	6,263
R ²	0.00004	0.00002	0.00001	0.00004	0.00004	0.00001

Notes: This analysis looks at the effect of treatment on the average public ratings of contracts for jobseekers in the experimental sample. All ratings are on a scale of 1 to 5 and averaged at the worker level. In Column (1) the outcome is the rating the employer gives to the workers' skills. In Column (2) the outcome is the rating the employer gives to the workers' communication ability. In Column (3) the outcome is the rating the employer gives to the overall quality of work completed. In Column (4) the outcome is the rating the employer gives to the workers' cooperation. In Column (5) the outcome is the rating the employer gives to the workers' ability to make deadlines. In Column (6) the outcome is the percentage of jobs worked where the worker was left any public rating by their employer. The experimental sample is of all new jobseekers who registered and were approved for the platform between June 8th and July 14th, 2021 and had non-empty resumes. Significance indicators: $p \leq 0.10$: *, $p \leq 0.05$: **, and $p \leq 0.01$: ***.

Table F18: Effects of writing assistance on private ratings and reviews

	<i>Dependent variable:</i>					
	Private rating			Positive text review		
	(1)	(2)	(3)	(4)	(5)	(6)
Algo Writing Treatment	-0.076 (0.082)	-0.132 (0.098)	-0.126 (0.094)	0.016 (0.019)	0.015 (0.022)	0.014 (0.021)
Anglophone Country		0.481*** (0.126)			0.014 (0.033)	
Trt × Anglo		0.213 (0.177)			0.003 (0.044)	
US			0.516*** (0.135)			-0.002 (0.036)
Trt × US			0.244 (0.190)			0.009 (0.049)
Constant	8.631*** (0.059)	8.480*** (0.071)	8.502*** (0.068)	0.858*** (0.014)	0.854*** (0.016)	0.858*** (0.016)
Observations	4,318	4,318	4,318	1,189	1,189	1,189
R ²	0.0002	0.011	0.011	0.001	0.001	0.001

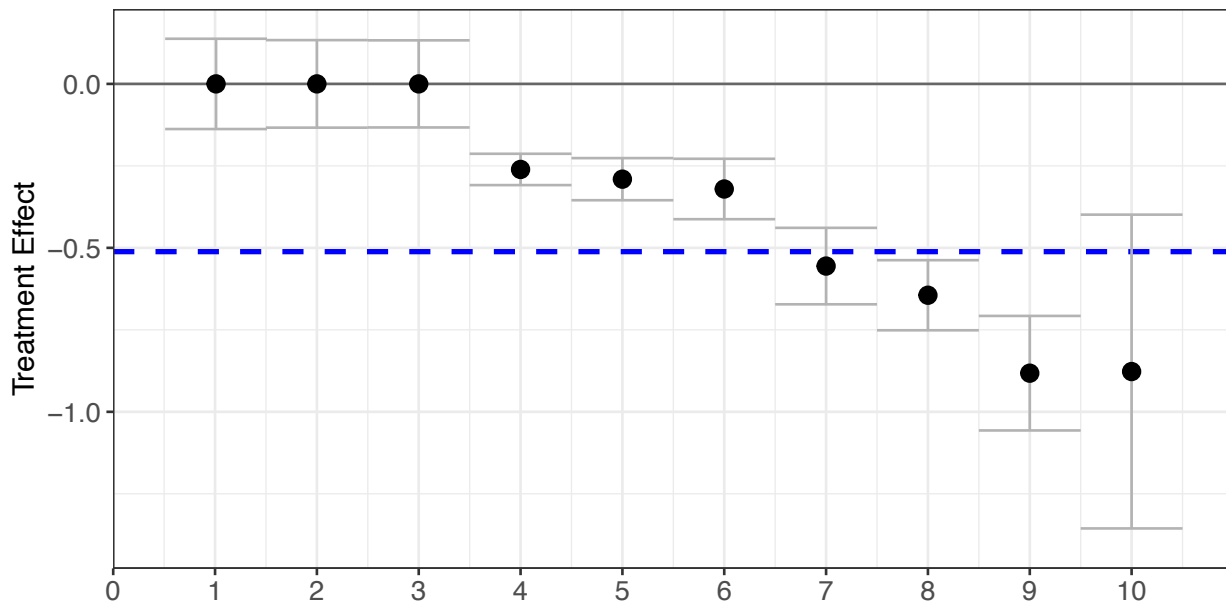
Notes: In this table we report the effect of the treatment to two measures of worker ratings. In Columns (1) through (3) the outcome is jobseekers average private ratings. In Columns (4) through (6) the outcome is the average percent of jobseekers reviews flagged as having a positive or neutral sentiment. In Column (2) and (5) we interact the treatment with a dummy variable for if the jobseeker is from the US, UK, Canada, or Australia. In Column (3) and we interact the treatment with a dummy for if the jobseeker is in the US. The experimental sample is of all new jobseekers who registered and were approved by the platform between June 8th and July 14th, 2021 and had non-empty resumes. Significance indicators: $p \leq 0.10$: *, $p \leq 0.05$: ** and $p \leq .01$: ***.

Table F19: Effects of writing assistance on resume readability

	<i>Dependent variable:</i>	
	Reading Ease Score (1)	Reading Difficulty Score (2)
Algo Writing Treatment	0.446*** (0.105)	-0.154*** (0.021)
Constant	39.779*** (0.075)	12.494*** (0.015)
Observations	195,247	185,487
R ²	0.0001	0.0003

Notes: This table shows the effect of the treatment on various writing readability scores. In Column (1) we show the effect of the treatment to the Flesch Reading Ease Score. This score is bounded by 1 and 100, with higher scores being more readable. In Column (2) we show the effect of the treatment to the Reading Difficulty Score, or the Flesch-Kincaid Grade Score. The score is unbounded, so we remove the most extreme 5 percent of outliers. Lower scores are more readable. The experimental sample is of all new jobseekers who registered and were approved by the platform between June 8th and July 14th, 2021 and had non-empty resumes, with outliers removed for the Flesch-Kincaid Grade Score. Significance indicators: $p \leq 0.10$: *, $p \leq 0.05$: ** and $p \leq .01$: ***.

Figure 1-4: Effect of treatment on the error rate, by deciles



Notes: This plot shows the effect of the treatment on the number of errors divided by the number of words in jobseekers' resumes, by deciles. Jobseekers in the lowest deciles have the least writing errors relative to their length, and jobseekers at the highest deciles have the most errors. The experimental sample is of all new jobseekers who registered and were approved for the platform between June 8th and July 14th, 2021, and had non-empty resumes, with $N = 194,700$.

1.7 A simple model of the “clarity view” of resume writing

In this section, we formalize a rational model of how the writing intervention could (a) increase hiring but (b) not lead to worse matches. We formalize the argument that better writing allowed employers to better ascertain who was a potential match with a simple model, and show how this kind of interplay between resume quality and hiring could exist in equilibrium.

1.7.1 A mass of jobseekers with heterogeneous productivity

There is a unit mass of jobseekers. If hired, their productivity is θ_i . Workers are either high-type ($\theta = \theta_H$) or low-type ($\theta = \theta_L$), with $\theta_H > \theta_L$. Workers know their own type. It is common knowledge that the fraction of high types in the market is γ . All workers, if hired, are paid their expected productivity, from the employer’s point of view. Hires only last one unit of time.

1.7.2 Jobseekers decide whether to put effort into resume-writing

Before being hired, jobseekers write resumes. Jobseekers must decide whether to put effort $e \in \{0, 1\}$ into writing that resume. Effort itself is not observable. The cost of this effort is jobseekers-specific and there is a distribution of individual resume effort costs. The support of the cost distribution is $[0, \bar{c}]$. The distribution has mass everywhere and the CDF is F and PDF is f . Jobseekers who put in no effort have resume costs of 0, while those that put in effort have a cost of c_i .

Before making an offer, firms observe a signal of jobseekers’ type on their resume, $R \in \{0, 1\}$. With effort, a high-type jobseeker generates an $R = 1$ signal; without effort, $R = 0$. A low-type jobseeker generates $R = 0$ no matter what. There is some share of workers λ for whom it is impossible to generate $R = 1$, regardless of their type. This share of workers are hit with a random shock of “bad writing” which make them unable to write clearly in English. There are $\gamma\lambda$ high type workers who are unable to generate $R = 1$, even if they put

in effort.

Clearly, low-types will never put in effort. The question is whether a high type will put in effort. The decision hinges on whether the cost of resume effort is worth the wage premium it creates. Let $w_{R=0}$ be the wage paid in equilibrium to jobseekers with $R = 0$. Note that $w_{R=1} = \theta_H$, as there is no uncertainty about a jobseeker's type if $R = 1$.

A jobseeker i who is a high-type will choose $e = 1$ if $\theta_H - w_{R=0}(c_i) > c_i$. The marginal high-type jobseeker is indifferent between putting in effort or not, and has a resume-writing cost of \hat{c} , where

$$\hat{c} = \theta_H - w_{R=0}(\hat{c}). \quad (1.3)$$

This implies that there are $F(\hat{c})\gamma(1-\lambda)$ jobseekers that choose $e = 1$. These are the high-type jobseekers with relatively low resume writing costs who aren't hit with the "bad writing" shock. The remaining $[1 - F(\hat{c})]\gamma(1-\lambda) + \gamma\lambda$ high-type jobseekers choose $e = 0$. They are pooled together with the $1 - \gamma$ jobseekers that choose $e = 0$ because they are low-types.

From the employer's perspective, if they believe that the resume effort cost of the marginal high-type jobseekers is \hat{c} , the probability an $R = 0$ jobseekers is high-type is

$$p_H^{R=0}(\hat{c}) = \frac{1 - F(\hat{c})\gamma(1-\lambda) + \gamma\lambda}{\gamma\lambda + (1 - F(\hat{c}))\gamma(1-\lambda) + (1-\gamma)}. \quad (1.4)$$

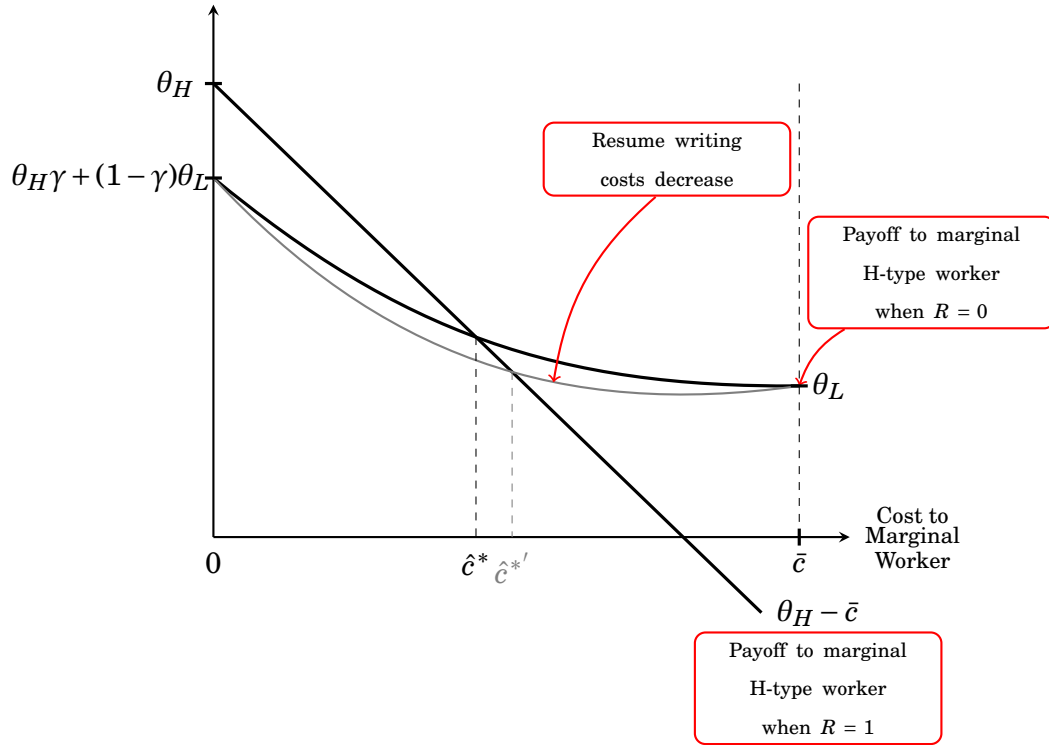
The wage received by an $R = 0$ worker is

$$w_{R=0}(\hat{c}) = \theta_L + (\theta_H - \theta_L)p_H^{R=0}(\hat{c}) \quad (1.5)$$

When the cost of the marginal jobseeker is higher, more jobseekers find it worth choosing $e = 1$, as $F'(\hat{c}) > 0$. This leaves fewer high-types in the $R = 0$ pool, and so

$$\frac{dp_H^{R=0}}{d\hat{c}} < 0. \quad (1.6)$$

Figure 1-5: Equilibrium determination of the marginal high-type jobseeker indifferent between putting effort into a resume



1.7.3 The equilibrium fraction of high-type workers putting effort into resume-writing

In equilibrium, there is some marginal high-type jobseeker indifferent between $e = 0$ and $e = 1$, and so

$$(\theta_H - \theta_L)(1 - p_H^{R=0}(\hat{c}^*)) = \hat{c}^*.$$

Figure 1-5 illustrates the equilibrium i.e., the cost where the marginal jobseeker is indifferent between $e = 0$ and $e = 1$. The two downward-sloping lines are the pay-offs to the marginal jobseeker for each \hat{c} . The pay-off to $R = 1$ is declining, as the wage is constant (at θ_H) but the cost is growing linearly. The pay-off to $R = 0$ is also declining, from Equation 1.6. Both curves are continuous.

Note that when the marginal jobseeker has $\hat{c} = 0$, there is just a point-mass of high-types that have a cost that low, i.e., $f(\hat{c})$. Because the marginal jobseeker is indifferent between

putting in effort and not putting in effort, jobseekers with costs of even ε will not put in effort. Since no one finds it worthwhile to put in effort the $R = 0$ pool is just the expected value of all jobseekers. And the wage is $w_{R=0}(\hat{c}) = \gamma\theta_H + (1 - \gamma)\theta_L$. The marginal jobseeker pays nothing, so the pay-off is θ_H .

At the other extreme, $\hat{c} = \bar{c}$, all but a point mass of jobseekers have a cost less than this. Since the marginal jobseeker is indifferent between putting in effort at a cost of \bar{c} , any jobseeker with cost $\bar{c} - \varepsilon$ or below will put in effort. Then the $R = 0$ pool is purely low-types and the wage is θ_L . For the $R = 1$ market, the marginal jobseeker has a cost of \bar{c} so the pay-off is $\theta_H - \bar{c}$. We know $\theta_H > \gamma\theta_H + (1 - \gamma)\theta_L$. And by assumption, $\theta_L > \theta_H - \bar{c}$, and so by the intermediate value theorem, an equilibrium \hat{c}^* exists on $(0, \bar{c})$.

1.7.4 A shift in the resume writing cost distribution leads to more high-type workers choosing to exert effort

Now suppose a technology comes along that lowers resume writing costs for some, and doesn't increase it for any, jobseekers. This technology also allows jobseekers hit with the "bad writing" shock to be able to generate $R = 1$ if they put in effort. The lower resume writing costs for those who use it shifts F higher for all points except the endpoints of the support, creating a new distribution of costs that first-order stochastically dominates the other.

Before determining the new equilibrium, note that no matter the marginal \hat{c} , when F increases, the probability that an $R = 0$ worker is a high-type declines, as

$$\frac{dp_H}{dF} = -\frac{1}{(F-2)^2} < 0. \quad (1.7)$$

This shifts the $w_{R=0}$ curve down everywhere, without changing the endpoints.

Because $w_{R=1} - \hat{c}$ is downward sloping, it intersects $w_{R=0}(\hat{c})$ at a higher value of \hat{c} . At the new equilibrium, the marginal jobseeker has resumes costs of $\hat{c}^{*'}$, where $\hat{c}^{*'} > \hat{c}^*$. At this new equilibrium, more jobseekers choose $e = 1$, causing more $R = 1$ signals. This lowers wages for the $R = 0$ group.

1.7.5 The effects of lower costs to welfare are theoretically ambiguous

Note that neither the shift in costs nor new found ability for λ high type workers to generate the positive signal are Pareto improving. While these high-types benefit from being able to collect $w_{R=1}$, low-types are made worse off as they find themselves in a pool with fewer high-types. Furthermore, because workers are all paid their expected product, the surplus maximizing outcome would be for everyone to choose $R = 0$. Resume effort purely changes around the allocation of the wage bill, not the total amount. Total surplus is

$$\theta_H \gamma + (1 - \gamma) \theta_L - \int_0^{\hat{c}} c f(c) dc, \quad (1.8)$$

which is maximized at $\hat{c} = 0$, i.e., when no one finds it worthwhile to choose effort. However, with a *shift* in cost distribution (raising F), what matters is whether the marginal decrease in costs for all inframarginal workers i.e, those with $c < \hat{c}$ outweighs the costs borne by the (newly) marginal jobseekers who choose to put in effort.

In our model, all job offers are accepted. However, if we think of jobseekers as having idiosyncratic reservation values that determine whether they accept an offer, the shift in costs makes it more likely that high-types will accept an offer, while making it less likely that low-types will accept an offer. This is consistent with results where there is a greater chance an employer hires at all in the treatment. It is also consistent with our result of higher wages. Finally, if we think of employer ratings being a function of surplus, our finding of no change in employer satisfaction is also consistent, as employers are, in all cases, just paying for expected productivity.

Chapter 2

More, but Worse: The Impact of AI Writing Assistance on the Supply and Quality of Job Posts

WITH JOHN HORTON

2.1 Introduction

The rapid advancement of artificial intelligence technologies, particularly in the field of Large Language Models (LLMs), has sparked considerable interest and speculation on their impact on the labor market. Anecdotaly, generative AI is already being used to generate application materials like cover letters and resumes as well as job posts (Smith, 2023; Mok, 2023). Hiring is costly—beginning with the writing of the job post and followed by applicant search and screening (Barron and Bishop, 1985; Blatter et al., 2012). If generative AI proves to be effective in assisting employers with job post writing, it could lower the cost of job posting. It may also lead to more standardized, coherent, and targeted job descriptions potentially reducing information asymmetry between employers and job seekers. On the other hand, concerns arise regarding the potential homogenization of job postings, the impact on job search strategies, and the downstream matches that result. From the perspective of online labor markets and other platforms, providing or encouraging the use of

generative AI could either improve the efficiency and accuracy of job postings or flood the market with informationless or homogeneous posts.

We analyze an experiment run on a large online labor market. We study the question of how providing would-be employers with LLM generated job posts impacts posting, user behavior, and hiring. A randomly selected treatment group of first time would-be employers were offered first drafts of their job posts written by generative AI. First, this experiment directly tests whether a technology which lowers the cost of posting increases the number and share of jobs posted. Second, we can see what types of jobs the treatment induces, and what types of applications they receive. Third, it allows us to test the efficiency of providing this kind of AI assistance to platforms who might consider this as a policy.

If firms find hiring to be costly in terms of time and domain knowledge, using LLMs to generate hiring materials has the potential to increase the supply of jobs. The existence of a multi billion dollar HR & recruiting industry suggests this is the case. [Blatter et al. \(2012\)](#) show that hiring one skilled worker costs 10 to 17 weeks of wages, and that these costs increase with the skill requirements of the position. Hiring online is also costly (despite fewer frictions), and digital platforms use recommendation systems to lower the costs of search and screening ([Oestreicher-Singer and Sundararajan, 2012](#); [Horton, 2017](#)).

Creating the job post itself can also be costly. In search and matching markets, employers create job openings and adjust their search depending on how costly a vacancy is ([Rogerson, Shimer and Wright, 2005](#)). There are mechanisms that search and matching models abstract away from that could be important for practice, like the decision to finish posting a job once started. In writing job posts, employers have traditionally had to rely on their own expertise or outsourced this work to recruiting agencies.

In the platform on which the experiment is run, 92% of employers who have posted jobs before publish a job post that they have started. A technology which makes it easier to post a job could benefit such employers. This is especially true for first time job posters who are less familiar with norms of the platform. On this platform, only 25% of those who begin the job posting process for the first time eventually publish a post. If this intervention can lessen frictions and make it easier for employers to post, in addition to whatever resources firms might allocate themselves, platforms and other social planners might consider further

expenditure to subsidize job posting, given the financial and social returns to job formation and employment.

Our first finding is that there is significant interest from employers—the employer can choose to opt in to receive an AI written first draft of the job post or opt out and write the job post themselves. 75% of employers opted in to receive the AI written first draft. We are able to track both which employers received the AI generated draft and the edits that they made to it before publishing the post.

Because we generated AI-written draft for both job posts in the treatment and control group (despite only revealing the AI-written draft to the treatment group), we can estimate a treatment effect for the similarity of the AI writing to the post that resulted. We use a measure of similarity where 1 means the two documents are identical and 0 means they share no elements. We find that job posts in the treatment group had mean similarity of 0.65 as opposed to jobs in the control group which had a similarity coefficient of 0.3.

We find that treated employers are 20% (or 6 percentage points) more likely to post a job than employers in the control group. Among those who do post a job, treated employers spend about 40% less time writing the job post than employers in the control group, on a base of 8 minutes. The distribution of length is compressed—the job posts which would have been short get longer and those which would have been very long get more compact. While the difference in mean number of words is small, this two-sided distributional shift causes the median word count to increase by 60%.

We also look downstream to how these treated job posts fared among jobseekers. Jobseekers rely on noisy signals of fit and job quality from job posts to decide whether or not to apply, which can vary by employer and job type. For example, non-native English¹ speaking employers in the control group receive fewer significantly fewer applications than native speaking employers. We find that the treatment was particularly useful to non-native English speakers—for them, treated job posts got significantly more applications. Since native English speakers saw no effect to the number of applications they received, the treatment significantly tightened the gap between the number of applications received by employers

¹We proxy for native English or not using the country that the employer reported registering from. We classify those from Anglophone countries US, UK, Ireland, Canada, New Zealand, and Australia as native English speakers.

along this dimension. We also test whether applicant pools for treated job posts are lower quality on average, using a measure of quality defined by the platform based on jobseekers' prior experience. We find no evidence to support this.

Despite this increase in applicants, treated employers are 18% less likely to make a hire on their first job post. The overall share of treated employers that hire a worker is no different to the share of control employers who hire. It may have saved the employers time, but access to AI written drafts resulted in no more matches.

In order to reconcile the large treatment effect to the number of job posts with no effect to the number of hires, we present a model where would-be employers decide whether to post a job with effort, post a job without effort, or not post a job at all. When writing a job description, employers can decide how much time and effort to put into carefully detailing the specifics of the tasks the job required, and the skills necessary to complete it. Therefore we model effort as something which causes the range of applications the job post induces to shrink, making it more likely for at least one application to come from a worker similar enough to what the job post requires to be worth hiring.

We introduce the AI as a technology which lowers the cost of posting a job but crowds out effort that some employers would have otherwise put into making the job post precise. If the cost of posting goes down for a subset of the job posts (the treated group with access to AI), a higher share of employers will post a job. However, the marginal jobs induced by the lower costs in the first period are ones with lower value to the employer, and causes ambiguous effects to hiring unconditional on posting. This rationalizes our otherwise surprising result that the treatment group had no more hires, despite 20% more job posts. Not only does the treatment induce lower value jobs that are less likely to hire, but it even makes the inframarginal jobs less likely to make a hire by decreasing the specificity of the posts.

To empirically test these hypotheses, we first show that employers' in the treatment group exhibited lower search effort than those in the control group. This is consistent with the hypothesis that the jobs posted in the treatment group were of lower value to the employers. Next, we embed the text of the job posts using OpenAI's "text-embedding-ada-002" model to create numerical representations of the texts. We first plot the embeddings of the job posts in the treatment and control group and show that the job posts in the treatment

group are clustered closer together than the treatment group. We then calculate the cosine similarity between each job post's embeddings to show that job posts in the treatment group are on average more similar to each other than those in the control group. This is consistent with the hypothesis that the text of the job posts in the treatment group are more generic.

The main contribution of this paper is to provide early evidence on the impact of generative AI in hiring. [Tambe et al. \(2019\)](#) suggests that using ML algorithms for recruiting can provide new knowledge that the recruiters missed. And [Van den Broek et al. \(2021\)](#) shows that humans can use ML in a hybrid practice in which candidates are judged and selected by relying on a combination of ML and recruiters domain expertise. While much of the literature on AI/ML in hiring is focused on algorithmic approaches to search and screening, there are a few papers on the use of AI/ML for application materials. Early evidence suggests that using algorithmic writing assistance on resumes makes workers more likely to be hired ([Wiles, Munyikwa and Horton, 2023](#)). But there is evidence that when the use of AI in application materials is disclosed, that people perceive the applicants as less competent and warm ([Weiss, Liu, Mieczkowski and Hancock, 2022](#)).

We also contribute to a very young literature on generative AI and productivity. Across multiple domains and versions of LLMs, there is evidence of large productivity effects. [Noy and Zhang \(2023\)](#) find that the use of ChatGPT on writing tasks caused treated workers to take 0.8 SD less time to create work that was even higher quality than the work from the control group. When paired with GitHub Copilot, treated workers completed coding tasks 55% faster than a control group without [Peng et al. \(2023\)](#). We contribute to this literature by providing a case study in a real labor market where access to a LLM saves users time and increases their engagement with the platform, but that the positive results do not exist downstream of posting.

Lastly, we contribute to a literature on the role of cost of entrance to market quality. [Mankiw and Whinston \(1986\)](#) theorize that free entry is not always socially optimal. In one paper analyzing the effect of Chinese export subsidy program, the author finds that the subsidy made exporters worse off by polluting the market with low-quality firms ([Zhao, 2018](#)). [Filippas et al. \(2023\)](#) find that when subsidizing entrance to an online labor market, the financial benefits to workers outweighed the cost of the resulting increase in job search.

Our results suggest that if employers have perfect information about the technology, generative AI is weakly welfare increasing for employers. They saved time, and the technology made it possible for some jobs that would otherwise been abandoned to be posted. However, no more hires resulted from these matches, and we suggest that for most employers, the use of the AI crowded out effort that they would have otherwise exerted to make a more specific job post.

Our results suggest that workers are made worse off. The flood of job posts with no increase in hires increased search costs for workers, and makes it harder for them to tell apart good and bad jobs. They also “wasted” their time on applications—despite resulting in no more hires, job posts in the treatment group got 106,565 more applications. From a platform’s perspective, the usefulness of such a tool depends on if the increase in likelihood of posting a first job posts induced by the treatment caused employers to keep coming back to the platform for future jobs. If not, the overwhelming result of the treatment was to flood the market with low-quality job posts.

The rest of the paper proceeds as follows. Section 2.2 describes the online labor market which serves as the focal market for this experiment. Section 2.3 describes the experimental design and results from the first-stage. Section 2.4 reports the experimental results of the treatment on job posting and hiring. Section 2.5 provides evidence for our proposed mechanism. In Section 2.6 we present a simple model that can rationalize our findings. Section 2.8 concludes.

2.2 The setting

This experiment was conducted on a large online labor market. In online labor markets, employers² search for and hire workers to complete jobs that can be done with only a computer and an internet connection. These markets can differ in their scope and focus, and platforms have different responsibilities they provide to employers and workers. Some common services platforms provide include soliciting and promoting job openings, hosting profile pages,

²We use the terms “employer,” “job opening,” and “application” for consistency with the economics literature and not as a commentary on the legal nature of the relationships created on the platform.

processing payments, certifying worker skills, and maintaining a reputation system (?).

In the platform which we use as our empirical setting, employers post job openings on the platform website with job descriptions, required skills, and scope of project. First, a would-be employer gives the title of the job, for example “E-commerce website copywriter”, “Web developer”, or “Executive assistant.” They then report the skills necessary to complete the job. Next, the employer picks the broad category the job falls into, for example, as “Administrative Support”, “Data Entry”, “Software Development”, among others. The jobs can either be one-off projects called “fixed price jobs” or hourly jobs, in which case the employer estimates how many hours they expect the job to take.

Workers find out about job openings in three ways. They can use electronic search to seek job posts in specific categories or for job openings that require specific skills. They can receive email notifications from the platform when a job is posted in a particular category. And finally, they can receive invitations from employers to apply to specific jobs.

Employers find workers in two ways. They receive organic applications from workers who find the job opening independently, or they search for workers themselves and invite specific workers to apply. Employers can search through worker “profiles.” These profiles contain workers’ history of work on the platform (jobs, hours, hourly rates, ratings) as well as their education history and skills. For both workers and employers, the platform verifies some of the information available to the other side of the market.

When a worker chooses to apply to a job opening, they apply with a cover letter and an hourly wage bid or a total project bid for fixed-price jobs. The employer determines whether to hire and, if so, which worker(s) to select. To complete the work on hourly jobs, workers install custom tracking software that serves as a digital punch clock. The software records not only the time spent working but also keystroke count and mouse movements. The software also captures images of the worker’s computer screen randomly. This information is all sent to the platform’s servers, and made available to the employer for monitoring in real-time. At the end of the contract, both parties give a reason for ending the contract (usually that the project was completed successfully) and provide both written and numerical feedback about each other.

2.3 Experimental design

This experiment intercepts would-be employers at the moment they begin to post their first job. From June 7, 2023, through July 20, 2023, newly registered employers on the platform were randomly allocated into a treatment and control group.

The experimental sample includes 181,962 employers who post 50,125 openings between them. Appendix Figure 2-10 shows the daily allocations of employers into the treatment and control groups.

2.3.1 Experimental intervention at the start of posting a job

When an employer on the platform wants to post a job, they go through a series of steps. First they provide a job title, the length of time they expect the job to last, and a list of skills required or demanded of the job. After they provide this information, they report some information on their expected budget and then move on to a page where they can input a job description. For employers in the control group, here they type in their job description and then submit the job to be posted.

For employers in the treatment group, as soon as they start to post a job, they are offered two options. They can either “get started using AI” or “I’ll do it without AI.” If they click on the latter button, they receive the status quo job posting experience. If they elect to “get started using AI” they are asked to describe the job they want to post in a sentence or two. See Figure 2-1 for a stylized version of the interface that the employers use.

As an example, after being asked to describe the job in a sentence or two, one employer wrote:

I need someone to generate a an Excel database showing the frequency of a search term in a list of targeted business media

This is incorporated into a prompt, calling a popular generative AI service.³ On the next page, the employer is shown the job post as written by the AI as well as a list of required skills. In the case of the above input, the employer would be shown:

We are looking for an expert who can generate an Excel database that displays the fre-

³The exact prompt is listed in Appendix 2.9.1.

quency of a specific search term in a list of targeted business media. The ideal candidate should have the following skills:

- *Strong knowledge of Excel*
- *Ability to work with large sets of data*
- *Research skills*
- *Attention to detail*
- *Time management skills*

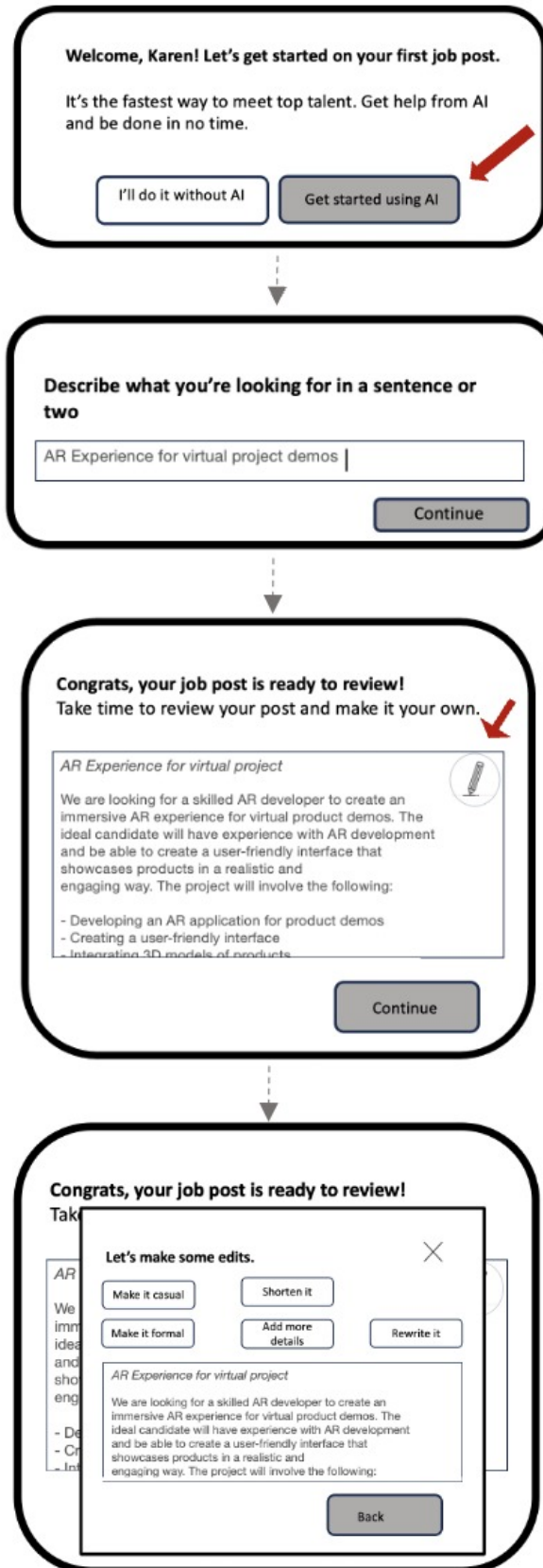
The page containing this draft contains the message to “take time to review your job post and make it your own.” Employers are able to edit the job post however they want and they are also shown a series of options for how the AI can edit the job post for them. The options are “Make it casual”, “Make it formal”, “Shorten it”, “Add more details”, “Rewrite it”. They will observe an have the option to edit a job post category which is determined by the API call as well.

2.3.2 Description of data used in the analysis

The dataset we use in this analysis consists of all job posts posted by employers in the experimental sample between the moment they were allocated into the experiment and August 4, 2023, 14 days after allocation ended. We construct job post level data with all posts, applications, and hires they have within 14 days of posting. Our economic outcomes of interest are 1) whether the employer eventually completed the job posting, 2) the number of applications to the job posts, and 3) whether or not anyone is ever hired for the job. We also collect the text of the job posts themselves, the amount of time the employer spent writing the post, and the count of skills required for the job. Lastly, we collect the country that the employer reported being in when they registered for the platform. We construct an imperfect definition of “native English speaker” which includes all employers who registered from the United States, Canada, United Kingdom, Ireland, Australia, or New Zealand.

We also use the output generated by the AI for both posts from treated and control employers. For employers in the treatment group, we observe the text generated each time they call the API, either through the initial job post generation or the later buttons used to

Figure 2-1: Stylized job post process for employers in the treatment group



have the AI edit the job post after. For employers in the control group, an API call is made using the job title, budget amount, expected length of job, and skills required. We observe the output, although the control employers do not.

2.3.3 Treatment take up

Of all the job posts by treated clients, 75% opted in to receive an AI generated first draft. The platform records every action and even click taken by each user to the microsecond. This helps us to see the ‘first-stage’ of the treatment. While opting into receiving the first draft was widely used, the personalization buttons were not. Treated employers made on average 1.2 API calls through the buttons, the first of which generated the draft job post. Employers used the “Make it casual” feature the most, 3.3% of the time. The next most used feature was “Add more details” which was used in 2% of the job posts. “Shorten it”, “Make it formal”, and “Rewrite it” were used on between 1 and 2% of the job posts.

Next, we calculate the Sørensen–Dice coefficient of each job post measuring the similarity between the job post an employer submitted and the one that the AI generated, regardless if the employer was in the treatment or control group.⁴ It measures the proportion of common elements or features shared by two sets, relative to the total number of elements present in both sets. A dice coefficient of 1 means the job posts were identical, whereas a dice coefficient of 0 means they have nothing in common. While only treated employers who opt-in to the treatment ever receive these generated job posts, they are calculated for all job postings regardless of treatment group. We find that job posts where employers opt-ed out of the AI treatment had low levels of similarity, which matched the similarity of job posts in the control group, of around 0.3. In the treatment group the average dice coefficient is 0.65. This gives us a first stage of the treatment–job posts in the treatment group share more than double the elements with those generated by the API call than job posts in the control group.

We see an even more pronounced difference between those that opted in to the treatment

⁴We cannot use the same prompt for job posts in the treatment and control group because the inputs from the employer are different. For job posts in the control group we generate a job post based on the title, skills, and proposed job duration that the employer inputs before writing their job description.

compared to those that opted out, we can see that on job posts where employers opted in the dice coefficient is 0.89. This last measure gives us a magnitude for how much employers who use the AI generated first drafts are editing them before they post them.

2.4 Experimental Results

Since employers were offered the choice to opt out of getting help from AI, our estimates are intent-to-treat effects for the entire experimental sample. In addition to the overall treatment effects, we present results interacting the treatment with a dummy variable for whether or not the employer registered for the platform from an anglophone country.

2.4.1 Treated employers were more likely to post a job

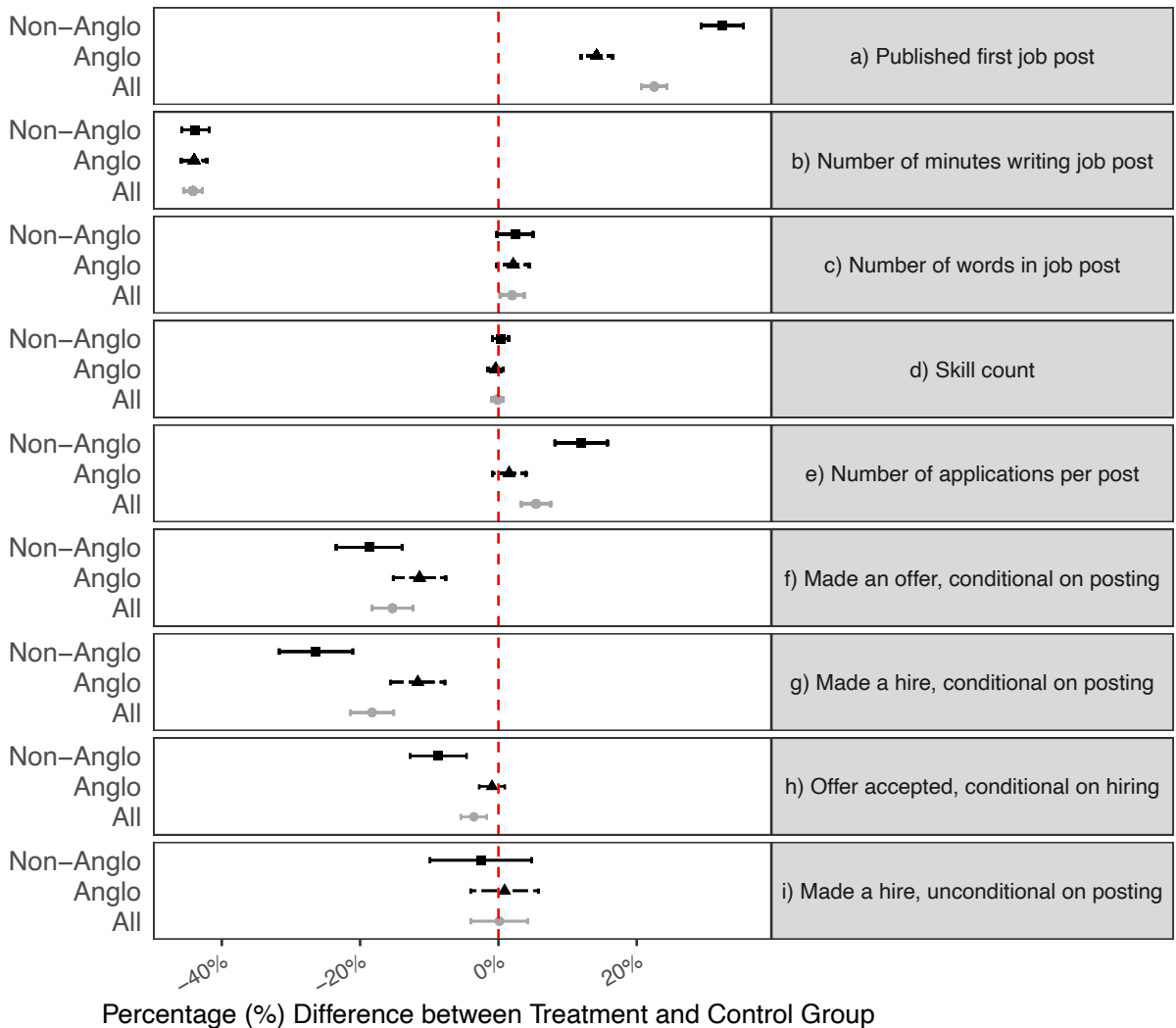
We will begin by observing that treated employers are about 20% more likely to post a job than employers in the control group. In Table D1 we show that only 25% of first time would-be employers who get to the landing page ever publish a job. This is very low compared to the employers who have posted a job before, for whom 92% who start the process for a given job publish it. There is clear room for improvement for keeping these would be employers in the hiring funnel. Treated job posts are 5 percentage points more likely to publish.

The effect of the treatment to whether an employer posts is significantly larger for employers who are native English speakers. Non native English speakers in the treatment group are 6.5 percentage points, or 24% more likely to publish than non native English speakers in the control group, while native English speakers only experience a 10% increase in likelihood of posting.

2.4.2 Treated employers spent less time writing the job post

The treatment caused employers to spend less time writing the job post. The outcome in Table D2, minutes, is defined as the difference in the timestamps from when the employer first clicks on the page to post a job and when the employer finally presses submit on the job post. Column (1) shows that employers in the control group spend on average 8 minutes

Figure 2-2: Experimental Estimates



Notes: This analysis looks at the effect of being assigned treatment on outcomes for employers in the experimental sample. The x-axis is the percentage difference in the mean outcome between employers in the treated group and the control group. The outcome a) published first job post is a 0 if the employer never posts a job after allocations and 1 if they do. The outcomes b), c), d), e), f), and g) are all conditional on the employer posting a job. The outcome h) offer accepted is conditional on the employer posting a job and making an offer. The outcome i) made a hire is unconditional on posting a job, it is 0 if the employer doesn't hire anyone after allocations and 1 if they do. A 95% confidence interval based on standard errors calculated using the delta method is plotted around each estimate. The experimental sample is of employers who posted a job between June 7th and July 20th, 2023, with $N = 181,962$. Regression details on the number of jobs posted can be found in Table D1, on minutes in Table D2, on number of words in Table D3, and on skill count in Table D4. Regression details on the number of applications can be found in Table D5, on offers in Appendix Table I14, on hires in Table D6, on offers accepted in Table D7, and on hires unconditional on posting in Table D8.

Table D1: Effects of generative AI on employer proclivity to post jobs

	<i>Dependent variable:</i>	
	Indicator for if first job is posted	
	(1)	(2)
GenAI Treatment Assigned (Trt)	0.056*** (0.002)	0.065*** (0.003)
Anglophone		0.109*** (0.003)
Anglophone X Trt		-0.021*** (0.004)
Constant	0.248*** (0.001)	0.200*** (0.002)
Observations	181,962	181,962
R ²	0.004	0.016

Notes: This table analyzes the effect of the treatment on the number of jobs the employer posts over the experimental period. Likelihood of completing first job post is a binary variable for the job post that the employer was working on when they were allocated into the experiment. “Anglophone” is 1 if the employer registers from an anglophone country, defined as the United States, Canada, United Kingdom, Ireland, or New Zealand. The sample is made up of all employers in the experimental sample. Significance indicators: $p \leq 0.10$: *, $p \leq 0.05$: ** and $p \leq .01$: ***.

Table D2: Effects of generative AI on length of time employer worked on job post

	<i>Dependent variable:</i>	
	Minutes writing job post	
	(1)	(2)
GenAI Treatment Assigned (Trt)	-3.581*** (0.071)	-3.302*** (0.105)
Anglophone		1.033*** (0.108)
Anglophone X Trt		-0.467*** (0.143)
Constant	8.107*** (0.054)	7.537*** (0.080)
Observations	38,841	38,841
R ²	0.061	0.064

Notes: This table analyzes the effect of the treatment on the number of minutes the employer spent working on the job post. Minutes is the difference in the timestamps from when the employer starts the job post till the timestamp when they publicly post it. “Anglophone” is 1 if the employer registers from an anglophone country, defined as the United States, Canada, United Kingdom, Ireland, or New Zealand. The sample is conditioned on employers who posted a job. Significance indicators: $p \leq 0.10$: *, $p \leq 0.05$: ** and $p \leq .01$: ***.

writing the job post while employers in the treatment group spend only 4.5 minutes.

2.4.3 Treated job posts were longer

Table D3: Effects of generative AI on length of job post

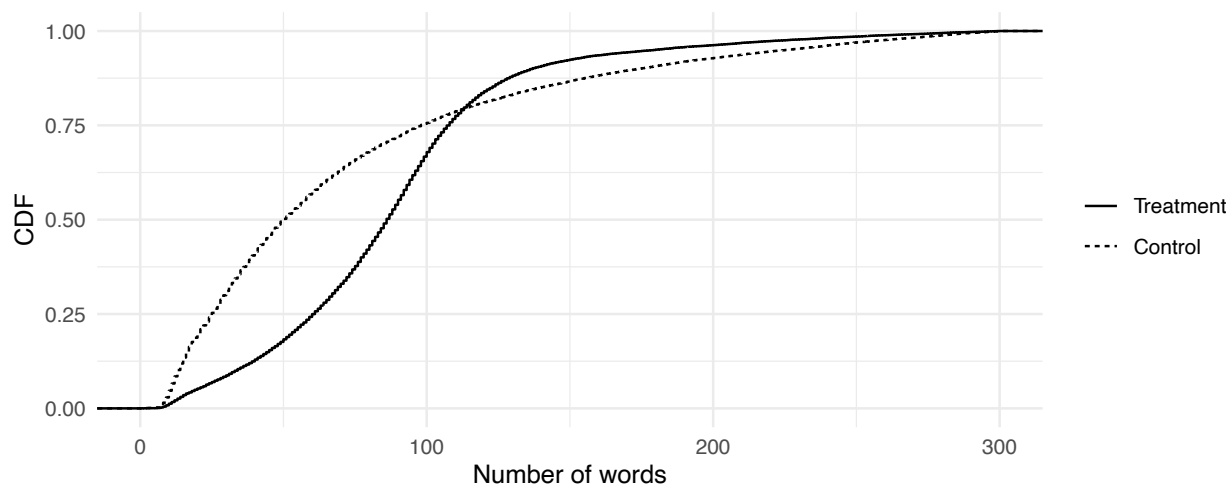
	<i>Dependent variable:</i>		
	Number of words in job post		
	OLS	Quantile Regression	
	(1)	(2)	(3)
GenAI Treatment Assigned (Trt)	1.942** (0.861)	2.253* (1.260)	32.000*** (0.555)
Anglophone		8.208*** (1.282)	
Anglophone X Trt		-0.066 (1.726)	
Constant	98.509*** (0.639)	94.010*** (0.949)	56.000*** (0.505)
Comparing Observations	Means 50,125	Means 50,125	Medians 50,125
R ²	0.0001	0.002	

Notes: This table analyzes the effect of the treatment on the number of words in the job posts. In Columns (1) and (2) are Ordinary Least Squares models. Column (3) compares the median number of words in job posts. “Anglophone” is 1 if the employer registers from an anglophone country, defined as the United States, Canada, United Kingdom, Ireland, or New Zealand. The sample is conditional on jobs which were posted. Significance indicators: $p \leq 0.10$: *, $p \leq 0.05$: ** and $p \leq .01$: ***.

On average, the treatment caused job posts to be longer. However, this masks an important distributional effect— the distribution of number of words in the treatment group was more compressed.

We start by looking at the mean. In Table D3 we see that job posts in the control group are on average 98 words long. Treated jobs are about 100 words long. However, if you look at the CDF of job post length in Figure 2-3 you see that the median job in the control group had 56 words, while the median number of words on job posts in the treatment group was 88.

Figure 2-3: Cumulative distribution function of the number of words in job posts



Notes: CDF of the number of words in employers job posts, by treatment status.

Table D4: Effects of generative AI on the number of skills requested by a job post

	<i>Dependent variable:</i>					
	Number of skills requested in job post					
	(1)	(2)	(3)	(4)	(5)	(6)
GenAI Treatment Assigned (Trt)	0.027 (0.020)	0.050* (0.029)	0.255*** (0.041)	0.298*** (0.036)	-0.218*** (0.068)	-0.438*** (0.069)
Anglophone		0.080*** (0.030)				
Anglophone X Trt		-0.040 (0.040)				
Constant	4.788*** (0.015)	4.745*** (0.022)	4.814*** (0.030)	4.468*** (0.027)	4.870*** (0.051)	4.842*** (0.054)
Category	All	All	Design	Software	Writing	Admin
Observations	50,125	50,125	11,918	13,100	4,068	4,558
R ²	0.00003	0.0002	0.003	0.005	0.003	0.009

Notes: This table analyzes the effect of the treatment on the number of skills requested in a job post. For jobs in the control group and those who opt-out of receiving an AI-written first draft, the skills are listed by the would-be employer as part of writing the job post. For jobs which get the post drafted by AI, the skills are pulled from the API call, although they can be overridden by the employer. “Anglophone” is 1 if the employer registers from an anglophone country, defined as the United States, Canada, United Kingdom, Ireland, or New Zealand. The sample is conditional on jobs which were posted. Significance indicators: $p \leq 0.10$: *, $p \leq 0.05$: ** and $p \leq .01$: ***.

2.4.4 Treated job posts listed different skill requirements

Every job post on the platform contains a list of skill requirements that jobseekers use to see if they are a good fit for a particular job. On average, there was no difference in the number of skills listed on each job post between the treatment and the control group, although this masks substantial heterogeneity. Table [D4](#) shows jobs in more technical categories (Design, Software) saw more skills listed on the job posts, while less technical categories (Admin, Writing) had fewer. This shows that the treatment had effects on how skills were conveyed on the job posts, but these effects were heterogeneous and not straightforward to summarise in a uniform way.

2.4.5 Treated job posts received more applications

Table D5: Effects of generative AI on number of applications a job post received

	<i>Dependent variable:</i>		
	Total apps		
	(1)	(2)	(3)
GenAI Treatment Assigned (Trt)	0.889*** (0.170)	1.754*** (0.249)	1.182*** (0.058)
Anglophone		3.109*** (0.253)	
Anglophone X Trt		-1.477*** (0.341)	
Constant	16.361*** (0.126)	14.657*** (0.188)	4.051*** (0.041)
Observations	50,125	50,125	181,962
R ²	0.001	0.005	0.002

Notes: This table analyzes the effect of the treatment on the number of applications the employer receives within 14 days of posting a job. “Anglophone” is 1 if the employer registers from an anglophone country, defined as the United States, Canada, United Kingdom, Ireland, or New Zealand. In Columns(1) and (2) the sample is made up of all employers who post a job post, and in Column (3) the sample is unconditional on whether or not the employer posted a job. Significance indicators: $p \leq 0.10$: *, $p \leq 0.05$: ** and $p \leq .01$: ***.

The treatment induced 28% more applications in the treatment group than in the control group. Job posts in the control group received 369,020 applications overall, while job posts in the treatment group received 475,585, a difference of 106,565 applications .

In Table D5 we will break down the effect of the treatment to the number of applications per job post—conditional on the employer publishing a job post. In Column (1) we see that job posts in the control group jobs received 16 applications on average. Across all job posts, treated jobs received almost 1 additional application. However, this masks significant heterogeneity by employer. In Column (2) we interact the treatment with a dummy variable for whether or not the employer was a native English speaker. First we notice that on average, workers prefer to apply to jobs from native English speaking employers—on average, job applications from native English speakers receive three more applications than those who were not. As for the effect of the treatment, there is no significant treatment effect to native English employers, while the treatment induces 1.75 additional applicants for jobs posted by

non native English speaking employers. While the treatment did not entirely close the gap between number of applications received by employers along this dimension, it significantly tightened it.

In Column (3) we look at the effect to applications unconditional on whether or not the employer posted a job. As we would expect looking at the entire sample, we see employers in the treatment group receive more applications— both because they are more likely to post a job and because the jobs posted in the treatment group draw more applications.

2.4.6 Treated employers job posts were less likely to make an offer, conditional on posting a job

Table D6: Effects of generative AI on number of hires

	<i>Dependent variable:</i>	
	Hire, conditional on posting a job	
	(1)	(2)
GenAI Treatment Assigned (Trt)	-0.035*** (0.003)	-0.035*** (0.005)
Anglophone		0.111*** (0.005)
Anglophone X Trt		0.006 (0.007)
Constant	0.192*** (0.003)	0.131*** (0.004)
Observations	50,125	50,125
R ²	0.002	0.025

Notes: This table analyzes the effect of the treatment on if the employer makes a hire. Hire, conditional on post is 1 if the job post that the employer was working on when they were allocated into the experiment makes a hire within 14 days. The sample is made up of all employers in the experimental sample. Significance indicators: $p \leq 0.10$: *, $p \leq 0.05$: ** and $p \leq .01$: ***.

Despite the large increase in employers propensity to post a job, and despite the increase in applications to those jobs, treated employers who post jobs are actually significantly less likely to make an offer or hire.

In Column (1) of Table I14 we can see that a posted job has around a 20% chance of making an offer on average. Treated jobs are 3 percentage points less likely to make an

offer. These results are generally consistent for both hires and offers⁵, so results to hiring are not overall driven by the employer making offers that are not accepted.

2.4.7 Treated non-native English speakers experienced more rejections after making an offer

Table D7: Effects of generative AI on the share of offers that are accepted

	<i>Dependent variable:</i>	
	Offer accepted	
	(1)	(2)
GenAI Treatment Assigned (Trt)	-0.029*** (0.008)	-0.060*** (0.013)
Anglophone		0.178*** (0.011)
Anglophone X Trt		0.052*** (0.016)
Constant	0.802*** (0.006)	0.688*** (0.009)
Observations	10,996	10,996
R ²	0.001	0.060

Notes: This table analyzes the effect of the treatment on the share of offers that are accepted. Offer accepted is 0 if an offer is made which does not lead to a hire and 1 if it does lead to a hire. The sample is made up of all employers in the experimental sample who post a job and make at least one offer. Significance indicators: $p \leq 0.10$: *, $p \leq 0.05$: ** and $p \leq .01$: ***.

Most offers are accepted. In our sample, 80% of job offers result in a hire. Table D7 Column (1) shows that workers given offers by employers in the treatment group were 3 percentage points less likely to accept. However, we can see from Column (2) that this result is driven entirely by non-native English speaking employers, for whom the treated group are 6 percentage points less likely to have an offer accepted.

2.4.8 Treated employers were no more likely to make a hire

One puzzle of this experiment is that despite the 20% increase in job posting, there is no overall increase in hires. Unconditional on whether or not the employers post a job, the

⁵See regression details on whether or not an employer made a hire in Table D6

Table D8: Effects of generative AI on number of hires, unconditional on posting a job

	<i>Dependent variable:</i>	
	Hire, unconditional on posting a job	
	(1)	(2)
GenAI Treatment Assigned (Trt)	0.0001 (0.001)	-0.001 (0.001)
Anglophone		0.049*** (0.001)
Anglophone X Trt		0.001 (0.002)
Constant	0.048*** (0.001)	0.026*** (0.001)
Observations	181,962	181,962
R ²	0.00000	0.013

Notes: This table analyzes the effect of the treatment on if employer makes an offer. Hire, unconditional on post is 1 if the employer makes any hire within 14 days of being allocated into the experiment. The sample is made up of all employers in the experimental sample. Significance indicators: $p \leq 0.10$: *, $p \leq 0.05$: ** and $p \leq .01$: ***.

likelihood of hiring in the control group is only 5% as we can see from Column (1) of Table D8. Among this entire experimental sample, treated employers are no less likely to make a hire.

These results imply that either none of the marginal jobs induced by the treatment made a hire, or the treatment actually made the inframarginal jobs worse.

2.5 Mechanisms

In Table D6 we showed that treated employers were 3 percentage points less likely to make a hire. In this section we provide evidence of underlying mechanisms. We first provide evidence that this effect is driven by the job posts induced by the treatment being on the margin of the cost benefit trade off. We provide evidence for this by showing that employers of treated job posts exhibit lower employer screening efforts. We next provide evidence that the job posts in the treatment group were more generic than those in the control group, as measured by both the text of the job posts and the similarity of their applicants. Lastly, we show that while the applicant pools were more similar for jobs in the treatment group, the issue is the applicant's fit, not their quality.

2.5.1 Employers exhibited lower search effort

Employers of treated job posts exhibited less search and screening efforts than those in the control group. Table E9 Column (1) shows that employers invite fewer would-be applicants to apply to treated jobs. In Column (2) the outcome is the number of applicants an employer puts on their short list. Employers shortlist fewer applicants to treated jobs. And in Column (3) the outcome is the number of interviews initiated by the employer, defined as a direct message from the employer to the applicant. Employers of treated job posts also interview fewer applicants. The magnitudes of these effects are all small but statistically significant, suggesting that the treatment induces more job posts, but that these job posts are relatively less beneficial to employers.

Table E9: Effects of generative AI on employer behavior

	<i>Dependent variable:</i>		
	Number of invites	Number of shortlists	Number of interviews
	(1)	(2)	(3)
GenAI Treatment Assigned	-0.008*** (0.002)	-0.005*** (0.001)	-0.251*** (0.029)
Constant	0.103*** (0.001)	0.032*** (0.001)	1.596*** (0.021)
Observations	50,125	50,125	50,125
R ²	0.0004	0.001	0.002

Notes: This table analyzes the impact of the treatment on employer behavior. Number of invites is the number of times a would-be employer reached out to a potential applicant and invited them to apply. Number of shortlists is the number of applications an employer put on their short list of potential hires. And number of interviews is defined as a 1 if the employer direct messaged a jobseeker after receiving their application. The sample is conditioned on employers who posted a job which received at least one application. Significance indicators: $p \leq 0.10$: *, $p \leq 0.05$: ** and $p \leq .01$: ***.

2.5.2 Treated job posts were more “generic”

Treated job posts had more generic text than job posts in the control group. To do this we use cosine similarity, which measures the distance between two texts language and content. To do this we first get the embeddings for each job post using OpenAI’s model “text-embedding-ada-00”. These embeddings are high-dimensional vectors that codify the semantic attributes

Table E10: Mean cosine similarity of job posts by treatment cell

	<i>Dependent variable:</i>	
	Mean cosine similarity	Rank
	(1)	(2)
GenAI Treatment Assigned	0.014*** (0.0002)	-6,647.600*** (98.939)
Constant	0.753*** (0.0002)	20,179.000*** (73.492)
Observations	33,022	33,022
R ²	0.107	0.120

Notes: This table analyzes the effect of the treatment on how different job posts are from each other. For each job post we get the embeddings using OpenAI’s ‘text-embedding-ada-002’ model, we then create a matrix of the cosine similarity between each job post and each other job post in the experiment. Then for each job post we take the mean of all of the cosine similarities, as a proxy for how generic a job post is. The outcome in column (1) is the mean cosine similarity between the ego and all other job posts in the experiment. The outcome in column (2) is the rank of those job posts in descending order. The sample consists of the subset of the experimental sample which post a job, and randomization occurs at the job post (and employer) level. Significance indicators: $p \leq 0.10$: *, $p \leq 0.05$: ** and $p \leq .01$: ***.

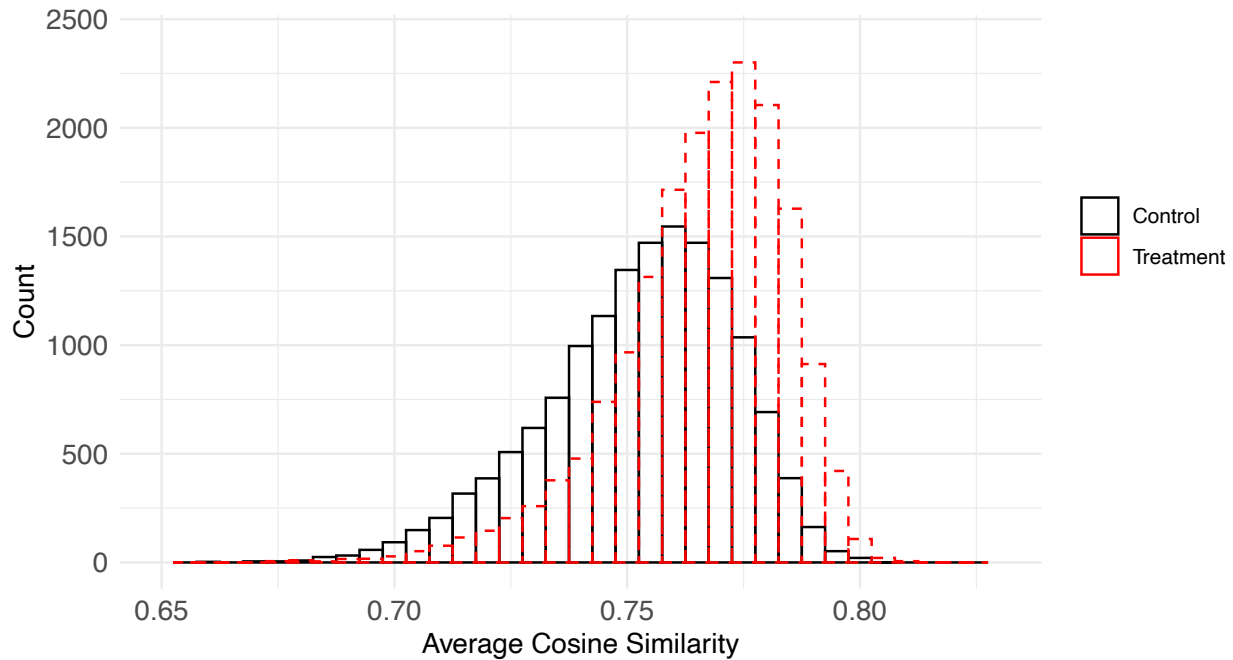
and content of the job descriptions, transforming the text into a numerical format that captures underlying meanings and themes. In Table E10 our outcome of interest is the mean of the cosine similarities between the embedding of job post i and the embeddings for each other job post $-i$. A cosine similarity of 1 means the texts are identical, and a cosine similarity of 0 means they are completely orthogonal. We give the definition of cosine similarity in Equation 2.1.

$$\cos(\mathbf{A}, \mathbf{B}) = \frac{\mathbf{A}\mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^n \mathbf{A}_i \mathbf{B}_i}{\sqrt{\sum_{i=1}^n (\mathbf{A}_i)^2} \sqrt{\sum_{i=1}^n (\mathbf{B}_i)^2}} \quad (2.1)$$

We find that job posts in the treatment group were on average closer to mean job post embedding than job posts in the control group. The treatment effect is small, only 0.014 on a base of 0.75. However, the range of average cosine similarities is very narrow—the lowest average cosine similarity is 0.67 while the highest is 0.81⁶. This is because despite the fact that these job posts can be in very different industries, they are all job posts. This

⁶See Appendix Section 2.9.2 to see examples of job posts with average cosine similarities at the min, max, and mean average cosine similarity.

Figure 2-4: Average cosine similarity by treatment status



Notes: This plot shows the average cosine similarity in the treatment and control cells for all employers in the sample who posted a job.

treatment effect covers 10% of the distance between the minimum and maximum average cosine similarity. Figure 2-4 shows how much this shifts the distribution of average cosine similarities.

Given the high dimensional nature of the embeddings, we cannot directly visualize them. To this end, we apply Principal Component Analysis (PCA) to reduce the dimensionality of the embeddings to two principal components, allowing us to visualize the embeddings in a 2D space. This reduction preserves as much of the variance in the data as possible in 2D. We plot the 2D embeddings for the treatment and control group in Figure 2-5. While the principal components themselves are not directly interpretable due to their composite nature, they still can facilitate a visual comparison of the job postings' embeddings. Most notably, the treatment appears to cause a shift in the distribution along the first principal component. We plot the 2D embeddings for the control group in Figure 2-5a. We investigate the treatment group in Figure 2-5b. Here we break down the job posts in the treatment group into those that opted-in to receive the AI written first draft, plotted in blue, and those that opted-out, plotted in red. For 75% of the job posts the employer opted-in to

receive the draft, and therefore the vast majority of embeddings are in blue. While the red embeddings for those that opted-out are placed more uniformly across the distribution of the first component, the ones that opted-in are clustered to the right.

2.5.3 Treated job posts had a higher fraction of their applications in common with other job posts

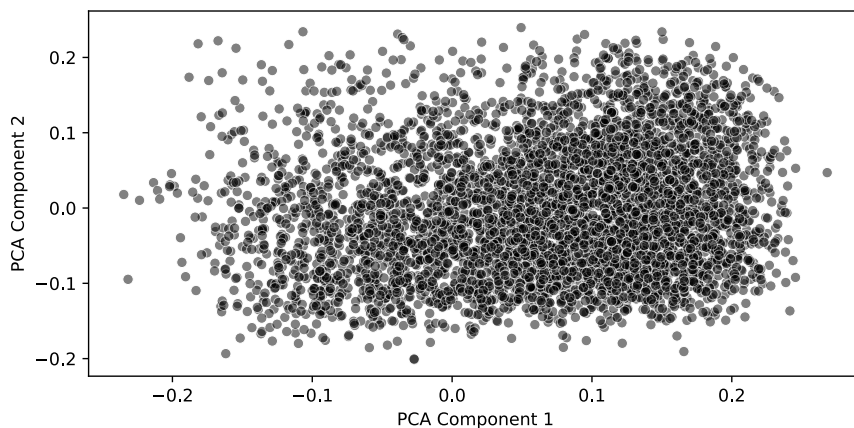
Table E11: Average share of applications in common with other job posts, times 100

	<i>Dependent variable:</i>
	Mean share of apps in common
GenAI Treatment Assigned	0.005*** (0.001)
Constant	0.055*** (0.001)
Observations	47,931
R ²	0.001

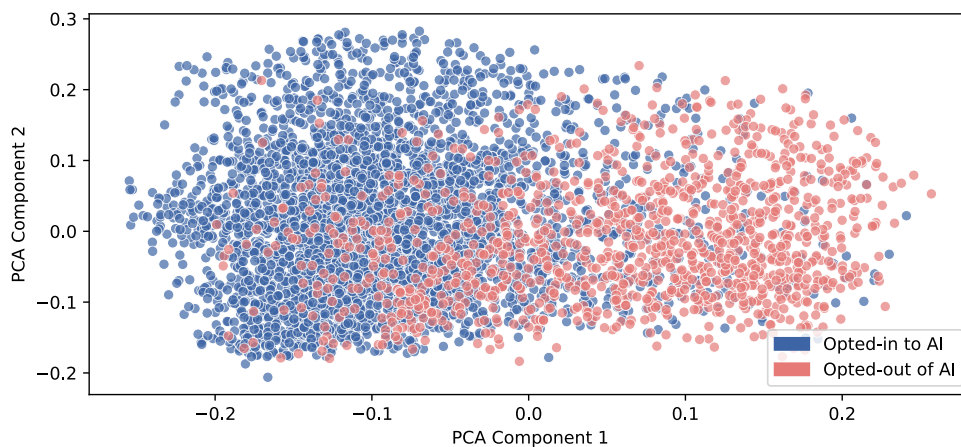
Notes: This table analyzes the effect of the treatment on how many applications job posts share with other job posts. We construct a matrix of all job posts, where the m by n th element is the fraction of the m th job posts' applications which come from a freelancer who also applies to the n th job. For each job post we take the mean of this measure across all other job posts. This is the independent variable. The sample is conditioned on employers who posted a job post which received at least one application. Significance indicators: $p \leq 0.10$: *, $p \leq 0.05$: ** and $p \leq .01$: ***.

We might imagine that if the treated job posts are in fact more generic, that they get applicants who are more similar to other job posts in the experiment. We then create a two dimensional matrix of job posts where the m by n th entry in the matrix is the cosine similarity of the n and m th job post in terms of the applications they received. For each job post i we then take the mean across all other job posts $-i$. This gives each job post an average measure of application overlap. In Table E11 we show that the mean share of applications in common is higher for jobs in the treatment group. In the control group, a job post's cosine similarity in terms of the applications it shares with other job posts in the experiment is 0.054. In the treatment group, it is 0.060. In both cases, the cosine similarity is very low but in the treatment group the similarity between job posts is higher. However, this might be because treated job posts get a larger number of applications overall.

Figure 2-5: Embeddings of Job Posts Reduced to 2 Dimensions



(a) Embeddings of Job Posts Reduced to 2 Dimensions for Control Group



(b) Embeddings of Job Posts Reduced to 2 Dimensions for Treatment Group

Notes: These plots shows the job posts' embeddings reduced to two dimensions. We use OpenAI's "text-embedding-ada-002" model to turn the text of job posts into embeddings, and then use PCA to reduce the dimensionality of the embeddings into two dimensions. We then take a random sample of 5,000 job posts in the treatment group and 5,000 job posts in the control group, for ease of visualization.

2.5.4 Applicant pools were not worse overall

First, we show that this effect is not driven by a worse applicant pool. It is possible that the job posts induced by the treatment had lower interest from applicants. However, we've already shown that treated jobs actually received more applications, in Table D5. Now we show that those applicants are no worse on average. When an employer collects applications for a job post, the platform recommends some applicants based on their wages, ratings, and employment history on the platform. In Table E12 the outcome of interest is the share of a jobs' applications which came recommended from the platform. Jobs in the treatment group saw a larger share of their applications come from recommended applicants.

Table E12: Effects of generative AI on quality of applicant pool

	<i>Dependent variable:</i>	
	Share of apps recommended	Number of recommended apps
	(1)	(2)
GenAI Treatment Assigned	0.009*** (0.003)	0.534*** (0.078)
Constant	0.296*** (0.002)	5.007*** (0.058)
Observations	50,125	50,125
R ²	0.0002	0.001

Notes: This table analyzes the impact of the treatment on the quality of a jobs' applicant pool. Share of apps recommended is the mean of applications a job post receives which the platform flags as recommended. Number of recommended apps is the number of application a job post receives which are recommended. The sample is conditioned on employers who posted a job which received at least one application. Significance indicators: $p \leq 0.10$: *, $p \leq 0.05$: ** and $p \leq .01$: ***.

2.6 Conceptual Framework

There is a unit mass of would-be employers (“employers”) considering posting one job each. There are two periods. In period 1, employers decide whether to post the job and whether to exert effort to be as specific as possible about the skill requirements and the details of the job. If they post the job, then in period 2 they receive applications and decide whether to hire a worker. If they do not post the job, nothing happens in period 2.

2.6.1 Period 2: The decision to hire

We first describe period 2. Each job j is defined by a location on a Hotelling line, $\theta_j \in (\underline{\theta}, \bar{\theta})$, which reflects the type of skills needed to complete the job. If the employer exerted effort in period 1, they receive N applications, with skills $\{\theta_i\}_{i=1}^N$, drawn iid from $U[\theta_j - \gamma, \theta_j + \gamma]$, where $\gamma > 0$ is a parameter that captures the fact that the employer cannot perfectly describe the skills needed in the job post.⁷ If the employer did not exert effort in period 1, they instead receive N applications drawn iid from $U[\theta_j - \rho\gamma, \theta_j + \rho\gamma]$, where $\rho > 1$ captures the fact that exerting no effort to specify the skills required results in a vague job post and thus draws applicants with a wider—and less relevant—set of skills.

Intuitively, exerting effort shrinks the support of the distribution of applicant skills and makes it more likely that the employer will receive an application close to θ_j . An employer is able to fill the job iff at least one application is within distance $m > 0$ of θ_j . If the employer is unable to fill the job—because they did not receive any application within distance $m > 0$ of θ_j , they receive period 2 utility of 0.

If the employer has at least one such application, they can choose whether to make a hire. If they make a hire, they receive value $v_j \sim G$ from completing the job and pay wage w .⁸ They also must pay idiosyncratic utility cost $\epsilon_j \sim U[0, 1]$, which reflects various hiring costs like search and screening. Therefore, conditional on being able to hire, they will hire iff $v_j - w - \epsilon_j \geq 0$.

⁷We define $\underline{\theta}$ and $\bar{\theta}$ such that this and subsequent ranges of applications are always interior to $(\underline{\theta}, \bar{\theta})$.

⁸We assume an exogenous and fixed wage because our experiment only affects a small subset of the market.

2.6.2 Period 1: The decision to post

In period 1, employers decide both whether to post the job, $p \in \{0, 1\}$ and, if they do post, whether to exert effort, $e \in \{0, 1\}$. Posting incurs cost $c > 0$ and effort incurs cost $c_e > 0$. They know v_j , but do not know ϵ_j nor whether they will receive an application sufficiently close to θ_j to be able to hire, so must form expectations over these objects when making their period 1 decisions. In particular, their utility if they post is given by

$$U(p = 1, e) = \pi(e, v_j)(v_j - w - \mathbb{E}[\epsilon_j | v_j - w - \epsilon_j \geq 0]) - c - ec_e,$$

where $\pi(e, v_j)$ is the probability of hiring, which happens if they are able to hire and ϵ_j is sufficiently low relative to v_j . If they do not post, they receive utility 0.

We now compute the objects $\mathbb{E}[\epsilon_j | v_j - w - \epsilon_j \geq 0]$ and $\pi(e, v_j)$. Since $\epsilon_j \sim U[0, 1]$, we can write $\mathbb{E}[\epsilon_j | v_j - w - \epsilon_j \geq 0] = (v_j - w)/2$.⁹ To obtain $\pi(e, v_j)$, note that this is given by $\Pr(\text{at least one application is within distance } m \text{ of } \theta_j | e) \cdot \Pr(v_j - w - \epsilon_j \geq 0)$. The latter term is just $v_j - w$. For the former term, denote an application as θ_i . Assume for now that $e = 1$. Then, this probability can be written as $\Pr_{e=1}(\min_i |\theta_j - \theta_i| < m) = 1 - \Pr_{e=1}(|\theta_j - \theta_i| > m)^N$. Since $\theta_i \sim U[\theta_j - \gamma, \theta_j + \gamma]$, this is $1 - (1 - \frac{m}{\gamma})^N$. Figure 2-6 shows the intuition for this: the probability of not being able to hire is simply the probability that all N draws fall outside of the shaded area, each of which occurs with probability $1 - \frac{2m}{2\gamma}$. If instead the employer did not exert effort in period 1, then this probability falls to $1 - (1 - \frac{m}{\rho\gamma})^N < 1 - (1 - \frac{m}{\gamma})^N$. Intuitively, if the support from which applications are drawn is wider, the probability of receiving an application within distance m of θ_j is lower.

Thus, plugging these objects in and simplifying, period 1 utility of posting is given by

$$U(p = 1, e) = \frac{1}{2} \left[\left(1 - \left(1 - \frac{m}{(1 + e(\rho - 1))\gamma} \right)^N \right) \right] (v_j - w)^2 - c - ec_e.$$

Note that effort and value of the job are complements: $\partial^2 U / \partial e \partial v_j > 0$. Intuitively, if v_j is higher, then the return to effort in terms of increased likelihood of finding a suitable

⁹We assume for simplicity that $v_j - w \in (0, 1)$. This is not a substantively important assumption—it merely simplifies the algebra. More generally, we could write $\epsilon_j \sim U[0, \bar{v} - w]$ where \bar{v} is the upper bound of v_j .

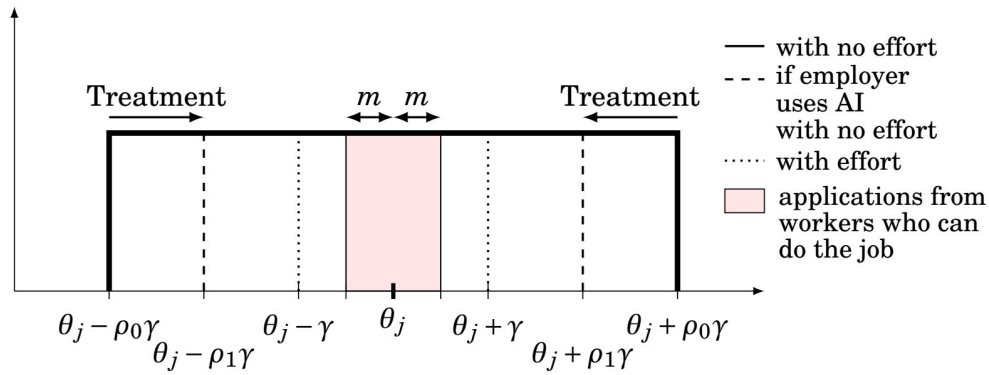


Figure 2-6: Stylized version of the distribution of applications job post j receives, and effect of the treatment

applicant is also higher.

The employer can choose one of three sets of actions: not post, post without effort, and post with effort. Their choice will be governed by v_j , as shown in Figure 2-7.¹⁰ For $v_j < \underline{v}_l$, they will not post, where \underline{v}_l is the unique value of v_j such that $U(p = 1, e = 0; v_j) = 0$. Intuitively, if the value of the job is low, it is not worthwhile for the employer to pay the posting cost c . For $v_j \in (\underline{v}_l, \underline{v}_h)$, they will post the job and not exert effort. Intuitively, for these workers the value of the job is high enough to justify the posting cost c , but not so high that the incremental gain from exerting effort to shrink the application pool exceeds the effort cost c_e . Finally, for $v_j > \underline{v}_h$, employers will post the job and exert effort. Intuitively, for very valuable jobs, the increased hiring probability from exerting effort is sufficient to justify the effort cost c_e .

2.6.3 Treatment

We now introduce a technology (AI) that does two things. First, it lowers the cost of posting a job from c_0 to c_1 , where $0 < c_1 < c_0$. Intuitively, AI writing software allows employers to spend less time writing a job post. Second, it shrinks the support of the application

¹⁰This depiction imposes a technical assumption that the first threshold for v_j is for the employers to post without effort, and the second threshold is that they will post with effort. This assumption is required for the effort choice to have bite: because effort and value are complements, if even the employer on the margin of posting preferred to exert effort, then all employers that post would exert effort (in which case the decision over effort would be irrelevant for the model). This assumption holds when c_e is sufficiently large—i.e., effort is costly enough that at least some employers that post prefer not to exert effort.

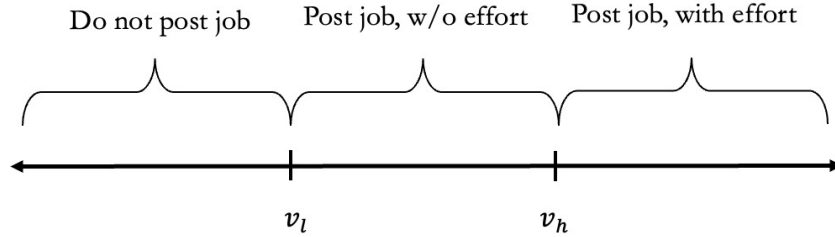


Figure 2-7: Possible values of v_j and what action the employer takes

distribution when the employer does not exert effort by lowering ρ from ρ_0 to ρ_1 , where $1 < \rho_1 < \rho_0$. Intuitively, AI writing software clarifies key elements of the job post if the employer’s original post was vague, but is still not as precise as the employer would be if they exerted effort to clearly specify the skills required.

Both of these effects cause \underline{v}_l to shift left. The lower cost of posting induces a previously-marginal employer to post as the cost has decreased. As the marginal employer was not exerting effort, the shift in ρ also increases their likelihood of being able to hire and thus further increases the return to posting. Intuitively, the cost of posting has decreased and the probability of hiring has increased, both of which cause employers with lower v_j to post who otherwise would not have.

The reduction in ρ causes \underline{v}_h to shift right. For an employer who was previously indifferent between exerting effort or not, the technology increases the probability that they will be able to hire if they do not exert effort, and thus they now prefer to not exert effort. Employers who have a very high value of v_j will still exert effort as $\rho_1 > 1$ —i.e., the incremental hiring probability is still worthwhile paying the effort cost for for very valuable jobs.

Treatment causes changes in the share of jobs that get posted, the likelihood of making a hire conditional on posting, and the unconditional likelihood of making a hire. We can see this in Figure 2-8, which shows that treatment causes a change for three groups. First, those with $v_j \in (\underline{v}_l^1, \underline{v}_l^0)$ post a job in treatment but not in control. These marginal jobs are less likely to hire than the inframarginal jobs because they are less valuable ($v_j < \underline{v}_l^0$) and so require even lower draws of the period 2 hiring cost ϵ_j .¹¹ Thus, for these jobs, the share that get posted increases, the probability of hiring conditional on posting decreases, and the

¹¹The probability that a job j posted without effort hires is $(1 - (1 - \frac{m}{\rho\gamma})^N)(v_j - w)$. As v_j is for these marginal jobs is lower than v_j for all inframarginal jobs, this probability decreases.

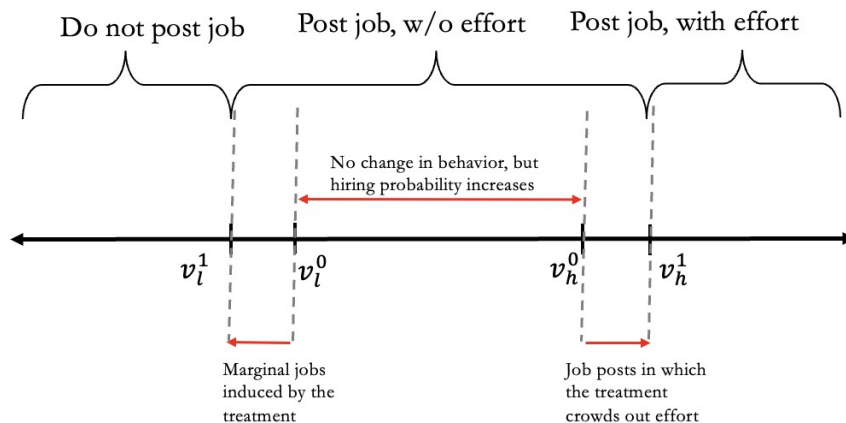


Figure 2-8: Impact of AI treatment on possible values of v_j and what action the employer takes

unconditional probability of hiring increases.

Second, those with $v_j \in (v_l^0, v_h^0)$ do not change their behavior—they post without effort in both treatment and control—but their probability of hiring increases as the shift in ρ from the technology increases their probability of finding a suitable applicant.

Third, those with $v_j \in (v_h^0, v_h^1)$ exert effort in control but not treatment. This does not affect the probability of posting because these jobs are always posted. It does reduce the probability for these jobs of making a hire, because the reduction in effort lowers the probability of finding an application with θ_i sufficiently close to θ_j . Thus, for these jobs, the share that gets posted is unaffected, and both the conditional and unconditional probability of hiring decreases.

Combining the previous three ranges of v_j , the model predicts that treatment increases the share of jobs that get posted. The effect to the probability of hiring conditional on posting is ambiguous, and will decrease if the effects to the first and third regions dominate the effects to the second. The effect of treatment on the unconditional probability of hiring is ambiguous. On the one hand, the increase in posted jobs increases the probability of a hire. On the other hand, the probability of hiring conditional on posting a job is lower for both marginal jobs (as they are less valuable) and inframarginal jobs (as some of them stop exerting effort). The net effect to the unconditional probability of hiring depends on which force dominates, which depends on the relative masses of v_j in the two regions as well as

the various parameters.¹²

2.6.4 Welfare

The treatment unambiguously increases employer welfare as they can always choose to ignore the technology. Marginals post jobs who not otherwise have done so, all inframarginals benefit from lower posting costs, and some inframarginal benefit from substituting costly effort towards using the technology instead.

2.7 Difference in differences analysis on near full roll out

After the conclusion of the experiment, the platform rolled out the policy to 95% of new employers, keeping 5% in the control group in perpetuity. This change was not publicly announced, and was likely a surprise to new entrants. Because there is a treatment and a control group both during the experiment and after, we can use a difference in differences analysis to provide evidence for what effects this treatment would have in equilibrium.

In Table G13 we report the difference in difference estimates for the primary outcomes. In Column (1) we find that prior to the roll out, the treatment had a 15% increase in the probability an employer completed their first job post. This is lower than the 20% increase during the primary experiment, and the decrease in the treatment effect might reflect learning in the control group as employers became more aware of AI. After the near full roll out, we see that the effect of the treatment substantially decreased. In the POST period, treated employers are only 0.12 percentage points more than control employers to post a job, on a base of 0.25, a 5% treatment effect. In Panel (a) of Figure 2-9 we show this outcome over time, for treated and control employers. The plot shows the shrinking treatment effect over

¹²Formally, the effect to the unconditional probability of hiring is given by $\int_{v_l^1}^{v_l^0} (1 - (1 - \frac{m}{\rho_1 \gamma})^N)(v - w) dG(v) + \int_{v_l^0}^{v_l^1} (1 - (1 - \frac{m}{\rho_0 \gamma})^N - (1 - \frac{m}{\rho_1 \gamma})^N) dG(v) - \int_{v_h^0}^{v_h^1} (1 - \frac{m}{\rho_1 \gamma})^N - (1 - \frac{m}{\gamma})^N) dG(v)$. The first two terms are positive and the third term is negative. This object could be either positive or negative. For example, if the mass of v in the first two ranges is small relative to the mass of v in the third range, this expression will be negative (and vice versa).

time, although the difference between the treatment and control group remain significant. We also observe that in September there is a large decrease in new posts in both the treatment and control group. This was due to a bug in the registration process that was fixed in October, but made it harder for new employers to register for the platform from desktop devices for two weeks. Fortunately it seems to have affected both treatment groups similarly.

In Table [G13](#) Column (2) we show that the treatment effect on the fraction of hires from posted jobs is no different before and after the policy change. And in Column (3) we show that the number of total hires in the treatment and control group is not significantly different after the policy roll out. Due to the smaller impact to number of posts, the point estimate for hires unconditional on posting is negative, although not different enough from zero to be significant. These two effects are consistent with the results we found during the main experiment.

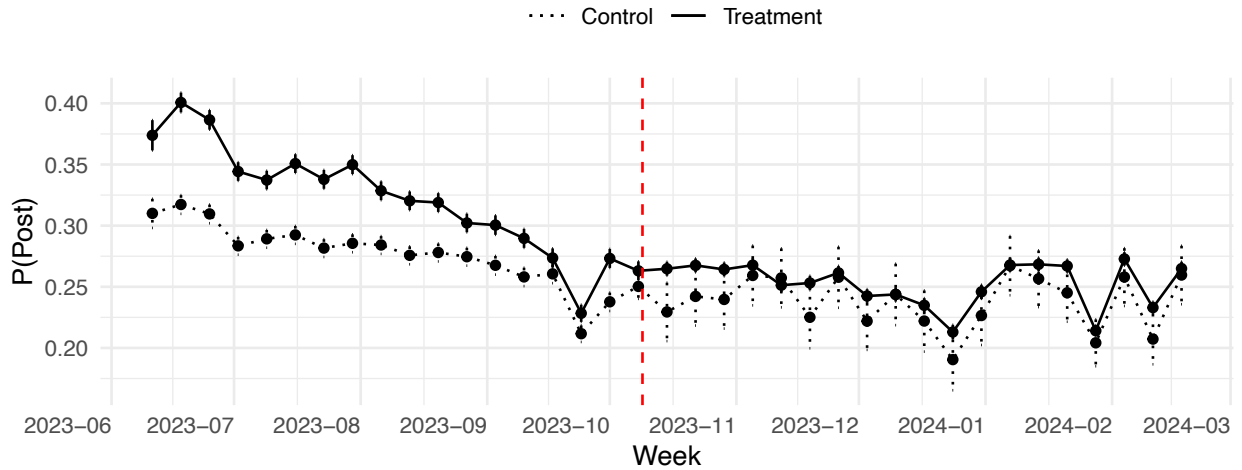
Overall, the difference in differences analysis suggests that the experimental results are consistent when rolled out to a large share of the market, although the magnitude of the treatment effect to posts is lower.

Table G13: Effect of AI treatment pre and post near full roll out

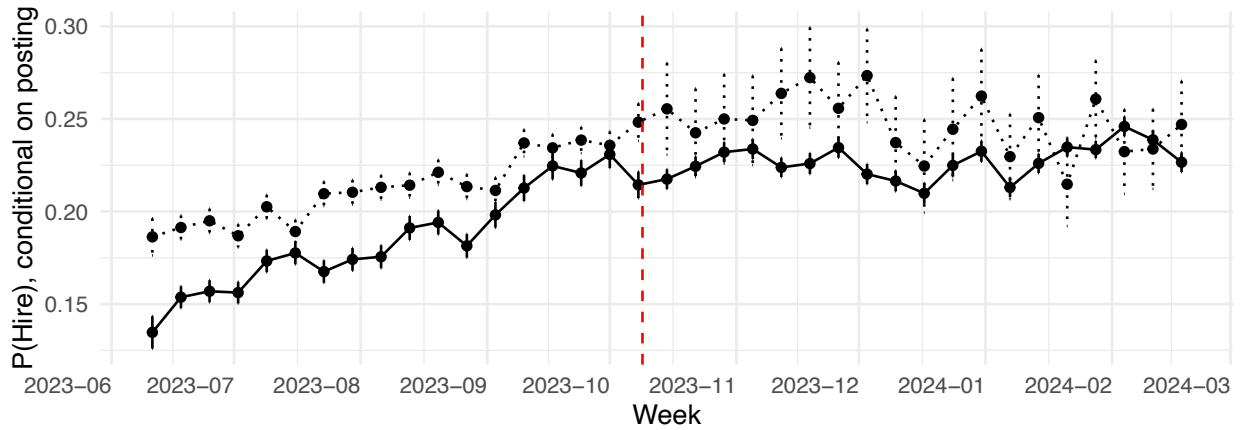
	<i>Dependent variable:</i>		
	Post (1)	Hire [cond] (2)	Hire [uncond] (3)
GenAI Treatment Assigned (Trt)	0.038*** (0.001)	-0.027*** (0.002)	0.0004 (0.001)
POST	-0.021*** (0.003)	0.036*** (0.005)	0.003** (0.001)
POST x Trt	-0.026*** (0.003)	0.008 (0.005)	-0.002 (0.001)
Constant	0.247*** (0.001)	0.215*** (0.002)	0.053*** (0.0004)
Observations	1,210,673	303,357	1,210,673
R ²	0.002	0.002	0.00001

Notes: This table analyzes the effect of the AI treatment during the experiment and after the treatment was rolled out to almost all new employers. Prior to the red line, 50% of employers starting their first job post are allocated into the treatment group. In the week of October 8th, 2023, the tool was rolled out to 95% of all employers starting their first job post, with 5% remaining in a control group. Allocation into treatment occurs when the employer begins a job post. Post is 1 if the employer allocated into the experiment posts a job. Hire [cond] is hires conditional on the employer posting a job. It is 1 if the job post that the employer was working on when they were allocated into the experiment makes a hire within 14 days. Hire [uncond] is hires unconditional on whether the employer posts. It is 1 if the employer allocated into the experiment makes a hire within 14 days, whether or not they complete a job post. The sample is made up of all new entrants to the platform and their outcomes during their first time starting a job post. Significance indicators: $p \leq 0.10$: *, $p \leq 0.05$: ** and $p \leq .01$: ***.

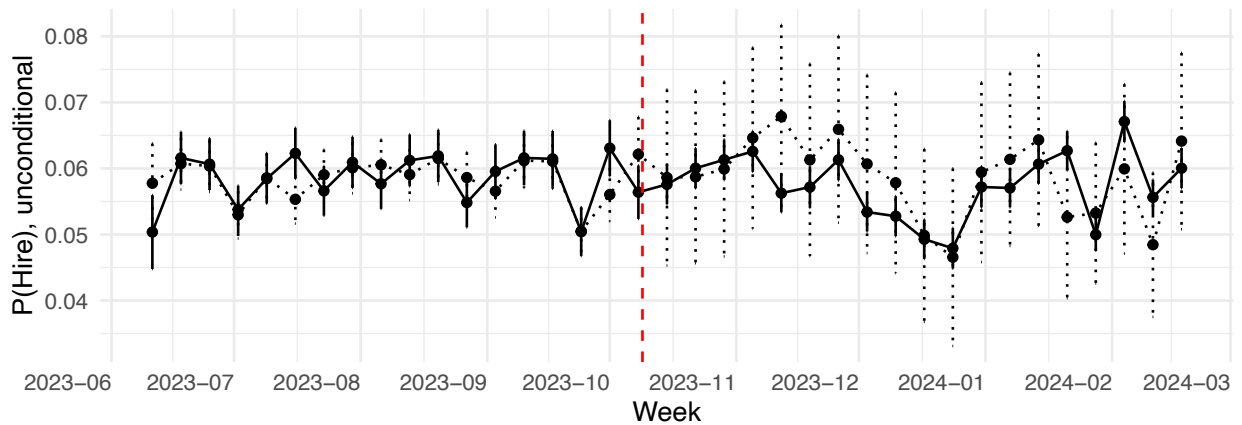
Figure 2-9: Posts and hires by treatment status during the experiment and after



(a) Fraction of started job posts that are completed, during the experiment and after



(b) Fraction of completed job posts that make a hire, during the experiment and after



(c) Fraction of employers that make a hire, unconditional on posting, during the experiment and after

Notes: Panel A of this plot shows the fraction of employers who post a job, by treatment status. Panel B shows the fraction of job posts that eventually make a hire. Panel C shows the fraction of employers that eventually make a hire, regardless of whether or not they post a job. The sample is of all first job posts by employers in the experiment. Before the dashed red line 50% of new employers were allocated to the treatment group. After the red line 95% of new employers were in the treatment group.

2.8 Conclusion

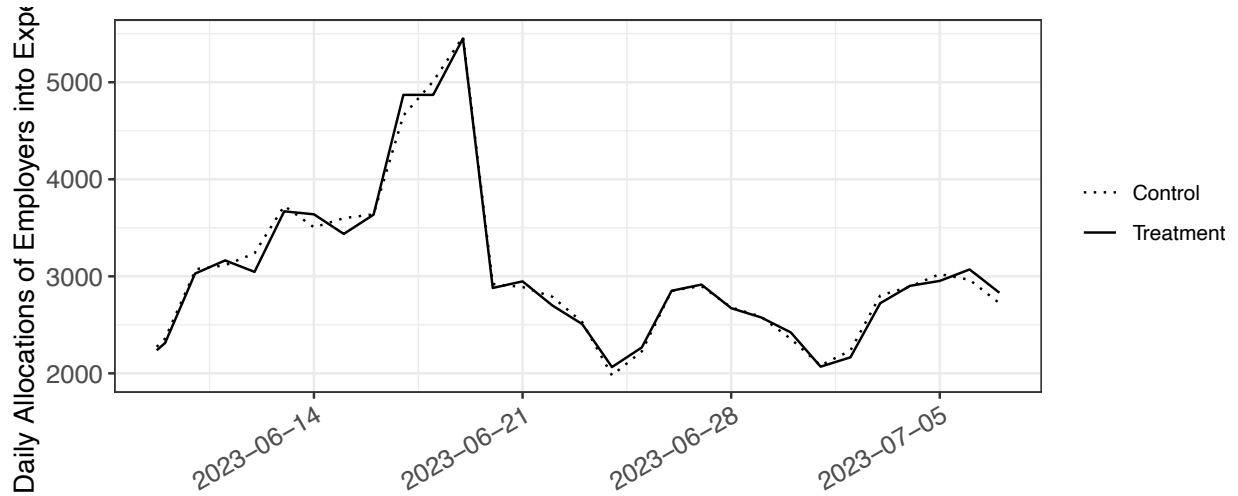
We show that job posting is costly— in an experiment run on an online labor market, treated employers who were offered to have an LLM write the first draft of their job post were 20% more likely to post a job.

We find the treatment benefited would-be employers. Treated employers spent over 40% less time to write their job posts, and those resulting job posts received at least as many applications with no worse applicant pools. These positive effects were significantly larger for employers who are not native English speakers. Nonetheless, treated job posts were less likely to make a hire.

Despite the large increase in job posts, the treatment group saw no more hires. We rationalize these results with a model where the treatment induces more job posts, but these marginal job posts are relatively less valuable to employers, and therefore less likely to result in a hire. Additionally, for the inframarginal job posts, the use of AI crowds out effort that employers would have put in themselves—resulting in more generic job posts.

After the conclusion of the experiment the treatment was rolled out to almost all new employers, and difference in difference estimates are consistent with the results from the experiment.

Figure 2-10: Daily allocations of employers into experimental cells



Notes: This plot shows the daily allocations into the treatment and control cells for the experimental sample of 181,962 employers.

2.9 Appendix

2.9.1 Additional tables and figures

<basicSystemPrompt> You are a(n) [platform] client posting a job.

<basicUserPrompt> Based on the following job requirements, write:

Title

Detailed job description:

Around 100 words in length

List relevant skills with bullet-points

Choose the most relevant size. Choose one of: 'small', 'medium', or 'large'

Choose the most relevant duration. Choose one of: 'under 1 month', '1 to 3 months', '3 to 6 months', or 'more than 6 months'

Choose the most relevant expertise level. Choose one of: 'entry', 'intermediate', or 'expert'

Respond with JSON! Keys should be ONLY 'title', 'description', 'size',

Table I14: Effects of generative AI on number of hires

	<i>Dependent variable:</i>	
	Offer, conditional on posting a job	
	(1)	(2)
GenAI Treatment Assigned (Trt)	-0.034*** (0.004)	-0.033*** (0.005)
Anglophone		0.078*** (0.005)
Anglophone X Trt		0.004 (0.007)
Constant	0.220*** (0.003)	0.177*** (0.004)
Observations	50,125	50,125
R ²	0.002	0.012

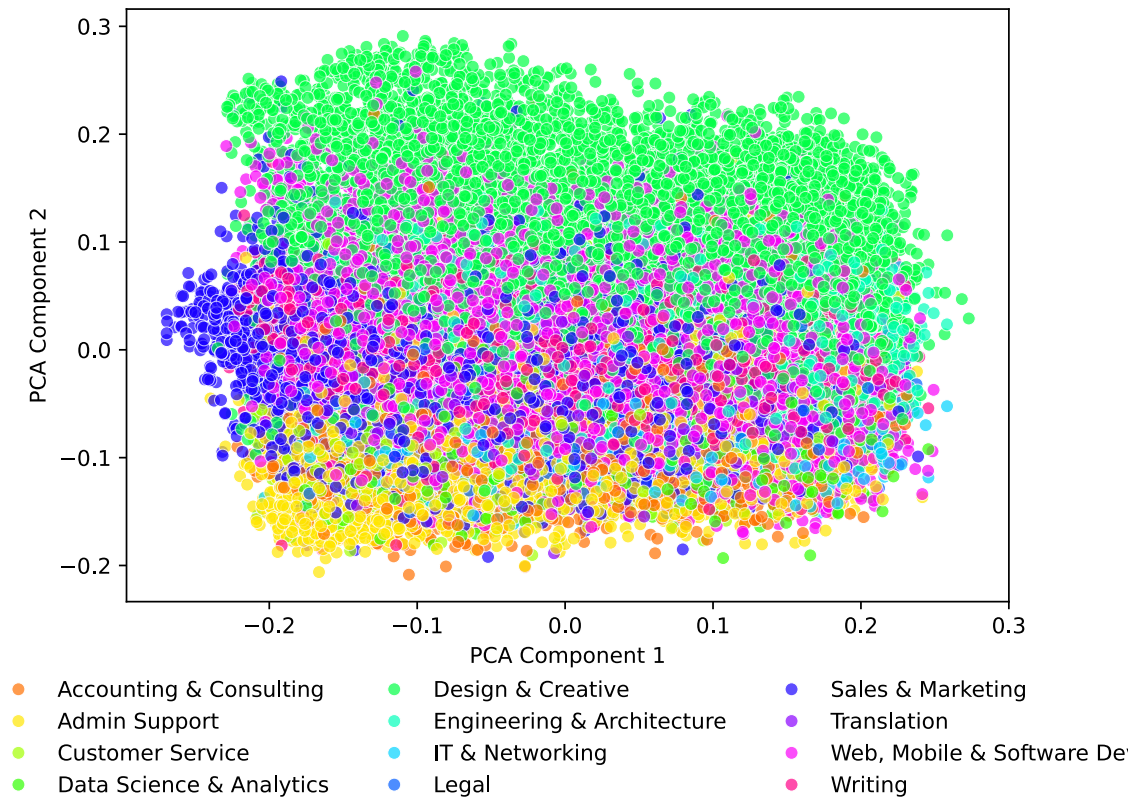
Notes: This table analyzes the effect of the treatment on if employer makes an offer. Ever offer, unconditional on post is 1 if the employer makes any offer within 14 days of being allocated into the experiment. Offer, conditional on post is 1 if the job post that the employer was working on when they were allocated into the experiment makes an offer within 14 days. Number of hires is the number of distinct contracts that form as a result of the job post. The sample is made up of all employers in the experimental sample. Significance indicators: $p \leq 0.10$: *, $p \leq 0.05$: ** and $p \leq .01$: ***.

□ ‘duration’, ‘expertise’.

Requirements: □ " " □ {{ requirements }}

2.9.2 Additional tables and figures

Figure 2-11: Embeddings of Job Posts Reduced to 2 Dimensions, by Job Category



Notes: This plot shows the text of all job posts reduced to two dimensions. We use OpenAI’s “text-embedding-ada-002” model to turn the text of job posts into embeddings, and then use PCA to reduce the dimensionality of the embeddings into two dimensions.

Range of average cosine similarities of job posts

The interpretation of the 0.014 effect of the AI treatment on the average cosine similarity of job posts does not have an obvious interpretation. A higher average cosine similarity means a job post is more similar to other job posts in the experiment, a lower average cosine similarity means a job post is more unique. The range of average cosine similarities amongst these job posts is 0.67 to 0.81, with a mean of 0.75. The following are examples of job posts at each of these points.

Most unique: Job post with average cosine similarity of 0.67

What needs to be done: in main.cpp file, I am calculating the force for all the particles. I want to optimize it. There are two linked lists in int main(). I want you to use them (list_cell and list_particle) basically, you have to fill the list_cell with list_particle's ID. and calculate force (in compute_acceleration) only for the particles which are in the same cell and neighbor cells, not for all. That's all. I would like to understand how you did it. Here is the code: (link removed) Run the project: get into the build folder...in terminal type Make then: ./md 100 10 0.01 (particles, Time, delta Time) For visualization, just download Paraview and open VTK file.

Mean uniqueness: Job post with average cosine similarity of 0.75

I'm looking for a logo for a small health and wellness company. I am a naturopathic doctor with an emphasis on weight loss counseling. I'm looking to have a design within the next month. I'm looking for a simple and clean logo that is modern yet whimsical and will transfer easily between Instagram, Facebook, a website, and other business materials like treatment plans, recipe books, etc.

Most generic: Job post with average cosine similarity of 0.81

We are looking for a skilled professional to assist us in creating a website and driving business growth. The ideal candidate will have expertise in web development and marketing strategies. The responsibilities include designing and developing a user-friendly and visually appealing website that aligns with our brand image and business objectives. The candidate should also possess knowledge of SEO techniques, social media marketing, and content creation to drive organic traffic and increase conversions. Excellent communication

and project management skills are essential for effectively collaborating with team members and delivering satisfactory results within the specified timeframe.

2.10 Second experiment to understand selection into receiving the AI generated first draft

In the previous experiment, employers could choose to opt out of receiving the AI generated job posts. Since these employers were all posting on the platform for the first time, we are not able to investigate which types of employers are selecting to receive help from AI. In order to investigate this selection, we look to another experiment run by the platform, this time run on a sample of employers who'd previously posted at least one job on the platform.

From April 20, 2023, through June 6, 2023, returning employers on the platform who posted a job were randomly allocated into a treatment and control group. The sample included all employers who had ever posted a job on the platform before. For treated employers, any job they post beginning at the time they are allocated into the experiment is considered treated.

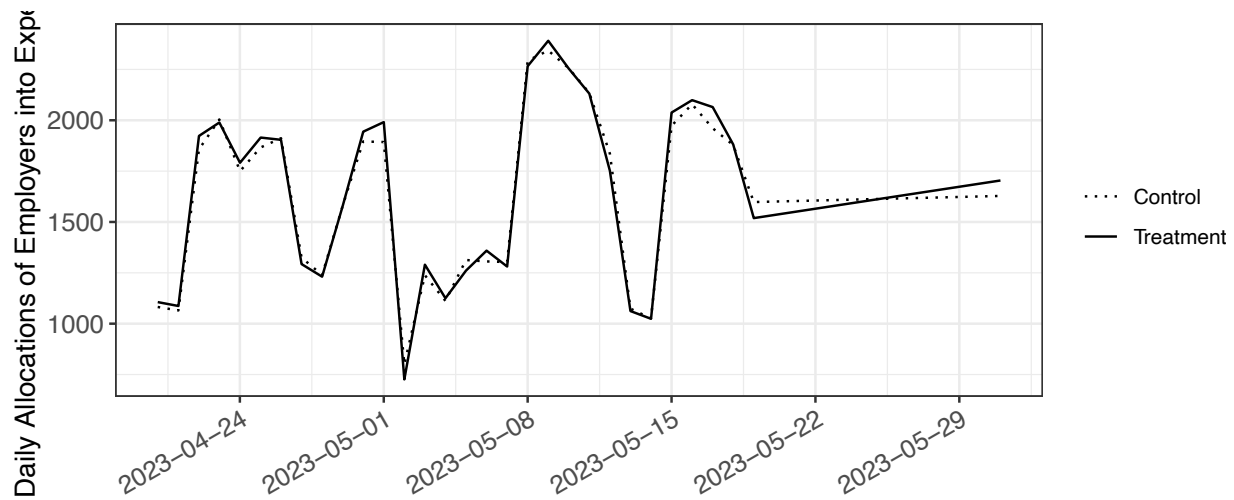
The experimental sample includes 101,601 employers who post 164,382 openings between them. Appendix Figure 2-12 shows the daily allocations of employers into the treatment and control groups. Table 2.10 reports pre-experiment attributes of these employers, and shows the sample of employers was well balanced in terms of the employers experience on the platform.

Table J15: Pre-randomization employer attributes by treatment status

Variable	Control	Treatment	P_value
From English-speaking country	0.59	0.59	0.61
US-based	0.42	0.41	0.09
Years since platform registration	3.13	3.13	0.89
Num posts, year before allocation	5.49	5.54	0.25
Num hires, year before allocation	3.40	3.43	0.36
Hourly wagebill, year before allocation	64,416.28	68,908.11	0.57
FP wagebill, year before allocation	44,184.90	51,323.58	0.67
Total hours demanded, year before allocation	3,925.41	3,241.50	0.45
Mean hourly wages, year before allocation	9.81	9.92	0.39

Notes: This table shows the difference between treatment and control workers for means of pre-experiment covariates, as well as a t-test comparing the difference between those means. Age is defined as the number of years between the employers' registration date and when they were allocated into the experiment. Mean hourly wages is conditional on the employer having made an hourly hire in the year prior to the experiment.

Figure 2-12: Daily allocations of employers into experimental cells, pilot experiment



Notes: This plot shows the daily allocations into the treatment and control cells for the experimental sample of 101,601 employers.

2.10.1 Description of data used in the analysis

The dataset we use in this analysis consists of all job posts posted by employers in the experimental sample between the moment they were allocated into the experiment and June 6, 2023 when allocation stopped. We construct job post-level data with all posts, applications, and hires they have within 14 days of posting. While in general we are interested in many outcomes related to posting and hiring, for these purposes we primarily want to 1) show that the take-up in this experiment was comparable with the main experiment and then 2) use the employer histories to understand if there is non-random selection into treatment.

2.10.2 Experimental intervention at the job description writing stage of job posting

When an employer on the platform wants to post a job, they go through a series of steps. First they provide a job title, the length of time they expect the job to last, and a list of skills required or demanded of the job. After they provide this information, they report some information on their expected budget and then move on to a page where they can input a job description. For employers in the control group, here they type in their job description and then submit the job to be posted.

For employers in the treatment group, after they input the basic information about the job and complete the budget step, they encounter an additional page that asks if they'd like help describing their job. If they select "yes" they have the option to click "Generate job post." The information they have entered so far is incorporated into a prompt, calling a popular generative AI service. The exact prompt is listed below.

```
# Given a job title of '{{title}}'  
# Given a job length of '{{duration}}'  
# Given job skills of {{skillNames}}  
# Write a detailed job description, without a title  
# Ask the candidate to submit a proposal  
# The candidate should describe how they can help with the project  
# The candidate should include some links to past completed projects
```

If the employer is not interested in the service, they click a button that says "I'll write it myself," and they are sent to the basic page employers in the control group would see.

2.11 First stage

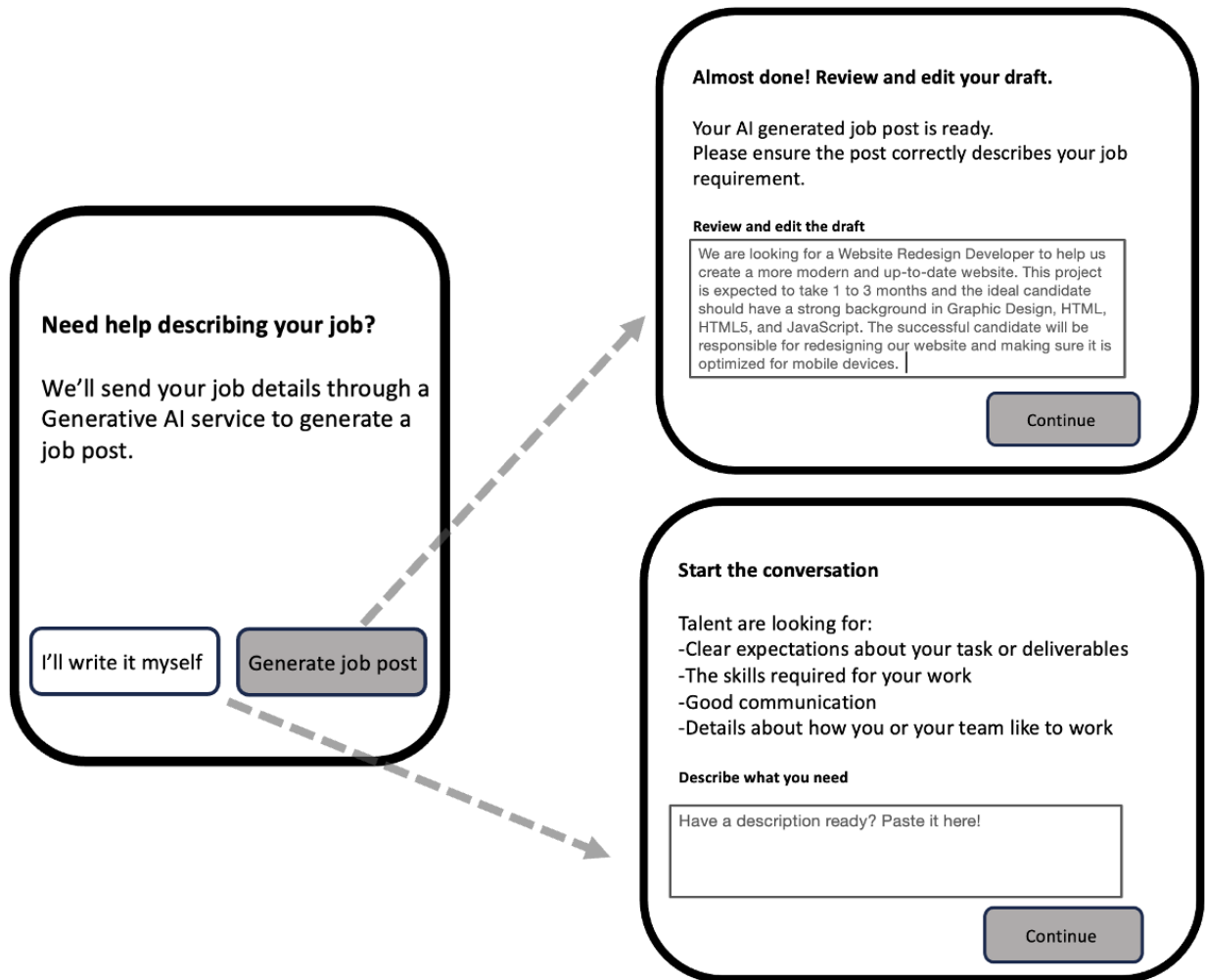
Most employers used the AI-generated job posts at least once. The platform records every action and even click taken by each user to the microsecond. Appendix Table [K16](#) helps us to see the 'first-stage' of the treatment. Of all employers in the treatment group, 53% opted-in to having the generative AI write their first job post. Of employers who made it through this stage, 62% opted-in. Of the employers who opted in, 78% edited the proposed job description, meaning 22% of employers posted the job without changing anything themselves.

2.12 Results

2.12.1 Treated employers were more likely to post a job

Treated employers were 10% more likely to post a job. In Table [L17](#) Column (1) we see that on this sample of returning employers, 92% who start a job post end up finishing it.

Figure 2-13: The “describe your job post” page in the job post process for the treatment group



Notes: This is a stylized version of the page of the job post process where employers write their job post for employers in the treatment group. For employers in the control group, they only see the bottom page titled “Start the conversation.”

Table K16: Treatment take up

Opted In	Count	Percent	Edited job after opting in
Yes	27,192	53%	78%
No	16,707	33%	NA
Never got this far	7,081	14%	NA

Notes: This table provides summary statistics on employers in the treatment group. “Opting in” means the employer chose to have GenAI generate at least one job post for them. Some employers drop off the job post process before getting to that step, these are labeled ‘Never got this far.’

This 10% increase is only about half of the size of the treatment effect we saw in the main experiment, which may be because it was run in April 2023, when employers may have been less familiar with LLMs.

Table L17: Effects of generative AI on employer proclivity to post jobs

	<i>Dependent variable:</i>	
	Indicator for if first job is posted (1)	Number of job postings (2)
GenAI Treatment Assigned	0.024*** (0.002)	0.056*** (0.012)
Constant	0.921*** (0.001)	1.570*** (0.008)
Observations	101,601	101,601
R ²	0.002	0.0002

Notes: This table analyzes the effect of the treatment on the number of jobs the employer posts over the experimental period. Likelihood of completing first job post is a binary variable for the job post that the employer was working on when they were allocated into the experiment. The sample is made up of all employers in the experimental sample. Number of job posts excludes any spam postings. Significance indicators: $p \leq 0.10$: *, $p \leq 0.05$: ** and $p \leq .01$: ***.

2.12.2 Employers who opt-in to treatment are slightly positively selected

In Table L18 we compare employers who opted in to the treatment with those who opted out on pre-experiment platform experience. We find that the employers who opted in to receive the AI-written draft are slightly positive selected on observables. This suggests that negative treatment effects to likelihood of hiring in the first experiment are unlikely to be due to the selection of “worse” employers taking up the treatment.

Table L18: Selection into opt-ing into the treatment, from the treatment group

	<i>Dependent variable:</i>			
	Hourly earnings	Fixed price earnings	Hours demanded	Hourly wages
	(1)	(2)	(3)	(4)
Opted-In to GenAI	17,871.920* (9,304.224)	18,859.110 (34,388.340)	1,043.654** (437.399)	0.177 (0.204)
Constant	60,942.970*** (7,322.737)	44,858.690* (27,064.780)	2,844.011*** (344.248)	9.947*** (0.160)
Observations	43,899	43,899	43,899	43,899
R ²	0.0001	0.00001	0.0001	0.00002

Notes: This table compares pre-experiment observable characteristics of employers in the treatment group who opt-ed in to the treatment to those who opt-ed out of it. Earnings, hours, and hourly rates are averages calculated from the month prior to when they were allocated into the experiment. Significance indicators: $p \leq 0.10$: *, $p \leq 0.05$: ** and $p \leq .01$: ***.

Chapter 3

Using AI to Upskill Non-Technical Workers into Data Science: A Field Experiment

WITH BCG COAUTHORS—MOHAMED ABBADI, FRANCOIS CANDELON,
DANIEL SACK, LISA KRAYER, URVI AWASTHI, RYAN KENNEDY,
CRISTIAN ARNOLDS

3.1 Introduction

The rapid advances in Generative Artificial Intelligence (GenAI) and its widespread deployment has sparked both excitement and concern about its potential impact on the workforce. These models increasing capabilities of automating complex tasks and knowledge work raises concerns about job displacement and the need for workers to adapt to the changing demands of the labor market. Reskilling or upskilling workers into new skills which complement emerging technology are key strategies for mitigating the negative effects of automation ([Acemoglu and Restrepo, 2018](#); [Djankov and Saliola, 2018](#)). While recent studies have explored the effects of GenAI on worker performance in tasks within their existing skill set ([Brynjolfsson et al., 2023b](#); [Dell’Acqua et al., 2023](#)) there remains a gap in understanding how GenAI can be used to help workers acquire new skills and adapt to changing

job demands. In this paper, we investigate the potential of GenAI as a tool for upskilling non-technical workers in the domain of data science.

Through a randomized experiment at a large management consulting firm, we demonstrate that providing non-technical workers with access to and training in GPT-4 can significantly improve their productivity and accuracy on data science tasks, even for problems that the AI alone cannot solve. We measure treated and control group workers on the accuracy of their output on a series of data science tasks. And we also compare their output to a benchmark set by the performance of data scientists at the firm for whom these tasks are a regular part of their job. This allows us to directly test how much non-technical workers armed with GenAI can do relative to workers in data science roles at their firm.

There is evidence from the literature that on-the-job training can help benefit both workers and firms. [Heckman et al. \(1998\)](#) shows that policies promoting skill formation, including on-the-job training, can significantly impact workers' earnings and skill development throughout their careers. [Acemoglu and Pischke \(1998\)](#) give evidence that when the structure of wages within a firm are distorted away from the competitive benchmark to the benefit of lower skilled workers, employers have incentives to provide general skills training to their workers, as they can benefit from the resulting higher productivity. This suggests that as automation from AI increase, firms in affected industries may benefit by increasing their training efforts to help workers adapt to new skills or even new roles.

However, in a landscape where the demands for skills are changing this rapidly, workers and firms may be weary about learning a new skill that will become obsolete as soon as the next generation of some language model is released. If GenAI is a general purpose technology ([Eloundou et al., 2023](#)) which can be used to allow workers to flexibly solve new types of problems as they emerge. There is evidence that GenAI is an effective and patient teacher ([Mollick and Mollick, 2022](#)). We therefore believe it is worth investigating whether GenAI itself can be used as a tool for on-the-job training in learning how to gain new skills.

We conducted a randomized controlled trial involving almost a thousand associates and consultants at the Boston Consulting Group (BCG), a large management consulting firm. Participants were randomly assigned to either a treatment group, which received access to and training in ChatGPT's most powerful model GPT-4, or a control group, which received

training on using Stack Overflow and other resources commonly used by data scientists. We surveyed workers both before and after the experiment. The experiment consisted a 20 minute interactive training session tailored to each treatment group followed by a series of data science tasks designed to be more technical than the participants' current roles. The tasks, developed in collaboration with OpenAI, were specifically designed to be challenging for GPT-4 to solve independently, requiring human input and reasoning. Each participant was randomly assigned two out of the three tasks, which included a statistics understanding task, prediction task, and a coding task. Each task tests a different skillset. The coding task is a very practical data cleaning task in python, which should be relevant for most data science jobs. The prediction task is the closest to the sort of projects BCG data scientists are often assigned—they must use historical data on past soccer games and provide the predictability of future soccer games, a common task in sports analytics. And the statistical understanding task is meant to test how workers make decisions about what types of statistics and machine learning tools to implement, requiring the deepest level of understanding of the three tasks. After grading the tasks, we also compare the performance of the participants was compared to a benchmark set by BCG data scientists who completed the same tasks without the use of ChatGPT.

Treated workers perform better on all three tasks. They are more likely to submit answers for the coding and statistics problems, and they take less time to submit answers on the prediction and coding problems. On grades normalized to between (0, 1), treated workers perform 43 percentage points better than the control group on the coding problem, receiving scores that were statistically indistinguishable from those of the data scientists. On the prediction problem, the treated workers performed better than control group workers, but left a large gap between their performance and the benchmark set by the data scientists. Lastly, treated workers performed only marginally better than the workers in the control group, suggesting that access to ChatGPT was least helpful on the tasks requiring deeper understanding. Taking all of these results together suggests that ChatGPT has the largest benefits to performing coding tasks, although it improves performance on all three.

Following the experiment we survey workers on their beliefs about their own and ChatGPT's abilities. We find that workers in the treatment group are no more able to answer

technical questions on sections where they are not allowed to use ChatGPT. However, we do find that treated workers are more confident in their ability to contribute to data science projects with the help of ChatGPT. While they are more confident in their technical abilities, they are not more confident in their ability to catch when ChatGPT is wrong, nor are they more trusting of the results the ChatGPT come up alone. Finally, we find that the treated workers exhibit strong overconfidence in ChatGPT's current capabilities—performing worse than the control group at guessing what types of problem GPT-4 can and cannot solve.

We contribute to the new literature on the effects of Generative AI on worker productivity. In one study, customer service agents given access to LLM suggestions are able to resolve more customer complaints more efficiently (Brynjolfsson et al., 2023b). In another example, consultants using GenAI on tasks that was in the model's range of ability were 25% faster at completing tasks, completed 12% more tasks and produced 40% higher quality output on average compared to their counterparts that did not use GenAI (Dell'Acqua et al., 2023). In this case, the tasks were within the skillset of the users (comparing interviews to excel data, and writing a Harvard Business Review style article), but they were able to use GenAI to produce output faster and on average of higher quality. However, on a task that the model reliably got the wrong answer on, the treated group was 25 percentage points less likely to come to the correct answer themselves. Our findings suggest that GenAI can improve the productivity of non-technical workers on complex tasks outside of their skillset, even for problems GenAI can not solve on its own.

Second, we contribute to a literature on job training and upskilling. Automation from AI is a serious concerns for academics and policy makers, with reskilling workers one of the primary strategies for workers to take to keep from being displaced (Djankov and Saliola, 2018; Acemoglu and Restrepo, 2018). In Deming and Noray (2020) the returns to work experience are a race between on-the-job learning and skill obsolescence. They show that the earnings premium for technology-intensive college subjects decline faster than more general subjects. This highlights the need for flexibility in skill acquisition, and show the benefit of training from a general purpose technology like GPT-4 (Eloundou et al., 2023) by making it easier for workers to adapt to changing job demands.

And lastly, we contribute to a literature on limitations of human-AI interaction. Prior

work has shown that people given access to AI often are not able to judge the quality of AI's outputs. Radiologists paired with AI are worse than when AI does diagnostics alone, because the radiologists rely on the AI when they are most uncertain, even though when they are uncertain is tightly correlated with when the AI is uncertain (Agarwal et al., 2023). In an online labor market, employers given access to AI-written first drafts of job posts produce more generic job posts which are less likely to make a hire (Wiles and Horton, 2024). Dell'Acqua (2022) finds that when recruiters have access to applicant recommendations by very high-quality AI, that they take the AI's suggestions, even when it's not correct. Our result that the workers with training in ChatGPT are overconfident in GPT-4 complement these findings, and we show that they even get worse than before at predicting the boundaries of GPT-4's abilities.

The rest of the paper proceeds as follows. Section 3.2 describes the experimental design and tasks we administer to workers. Section 3.3 describes our methods and analysis. Section 3.4 reports the experimental results of the ChatGPT training on workers' abilities to complete data science tasks as well as their perceptions about the technology and their own technical skills. Section 3.5 concludes.

3.2 Experimental Design

We report the results from a large randomized control trial run on associates and consultants of the Boston Consulting Group, a large managerial consulting firm, to test whether high skilled but non-technical workers can do data science work with the help of GPT-4¹.

The experiment took place in March and April of 2024. In the recruitment phase, all BCG associates and consultants and were sent an email inviting them to participate in a study on upskilling and GenAI. We indicate that participation in the study is voluntary, can be done during work hours, and the time will count as an "office contribution" to their career development committee, which has financial implications to their annual bonuses. We also provide additional incentives to the top 50% of performers in each treatment group

¹We pre-registered our study with the AEA RCT Registry on March 13, 2024 detailing the design structure, the experimental conditions, the dependent variables, and our main analytical approaches.

to encourage an ‘honest effort’ in the tasks². Those who registered were given a survey on their demographics, programming and ChatGPT skills, technology openness, creativity, and learning orientation (Agarwal and Prasad, 1998; Miron et al., 2004; Jha and Bhattacharyya, 2013). Demographic and other variables were later used for stratified random assignment, as described below. Details of the registration survey are available in the Appendix .2.

Simultaneously, BCG data scientists were also invited to participated in a a similar exercise, where they simply completed the tasks used in the experiment. Their output from these tasks served as the benchmark³ for the “typical performance of a data scientist.” 40 data scientists submitted tasks to serve as this benchmark.

After the registration survey, consultants were then randomly assigned to either be in treatment group or control group. Treatment was stratified across gender, location, role (i.e., associate or consultant), coding skills, college degree (i.e., bachelors, masters, Ph.D.), and experience with ChatGPT for coding. The experimental sample consisted of 986 constultants, with 493 each in the treatment and control group. The experiment contained four phases, a pre-experiment survey, a 15-20 minute training on effectively using ChatGPT (in the treatment group) or Stack Overflow (in the control group), a series of tasks more technical than their current role, and then a post-experiment survey (Figure 3-1).

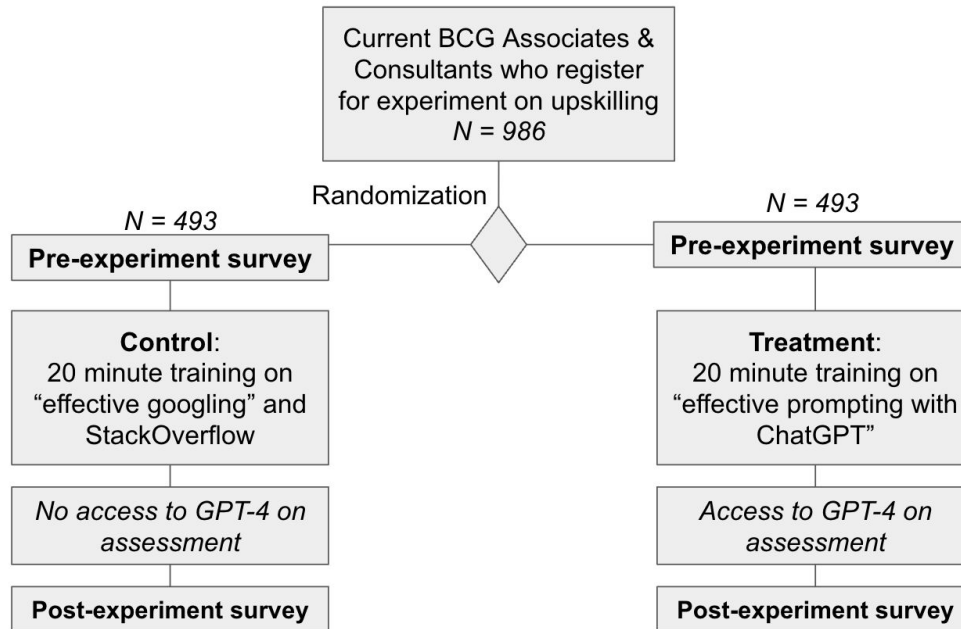
The pre-experiment survey consisted of questions of the participants subjective coding skills, GenAI usage, professional identity, and career aspirations. Next, participants were provided with a 15-20 minute tailored training to each experimental condition – the treatment group will receive specialized instructions on ChatGPT prompting and using it for data science, while the control group will be get training on effective googling and how to leverage websites commonly utilized by data scientists like StackOverflow and Khan Academy. Both trainings involved a combination of videos and interactive practice to prove competence. Details of the pre-experiment survey and both trainings can be found in Appendix Section .4.1.

The main experiment involved participants completing three complicated tasks repre-

²Top performers in each treatment group received recognition among BCG leadership as well as an invitation to a small group chat with offers for OpenAI and OpenAI merchandise.

³Data scientists were told not to use ChatGPT in their completion of the tasks. Therefore when comparing the results of the participants to the data scientists benchmark, it is likely the benchmark of a data scientist without access to AI.

Figure 3-1: Overview of experimental design



Notes: The registration survey is in Appendix Section .2. The pre-experiment survey and training for can be found in Appendix Section .4.1. The text of all three tasks can be found in Appendix Section .4.2. The post-experiment survey can be found in Appendix Section .4.3.

sentative of work done by BCG data scientists. There were three possible tasks participants completed, to test their ability to do different types of data science work. The first task was on statistical understanding—participants were given data on home buyers and had to use this data to predict whether a couple will take a mortgage out on a house, and were given a series of graphs and machine learning output they had to interpret. The second was a predictions task, which requires the participants to use historical data on men’s international football games to develop a strategy for sports investing. The third was a coding task, where the participants had to write and submit python code to clean and merge data, and then answer questions about the data.

The tasks were created in collaboration an analyst from OpenAI to be unable for GPT-4 to correctly answer. For all three tasks, if the participant let ChatGPT answer the question on it’s own, the answer was incorrect. Each task was intended to take 90 minutes to complete, so to avoid fatigue, we gave each participant a random two of the three tasks, with the task order randomized. Details on the tasks can be found in Appendix Section .4.2.

Following the task completion, a post-experiment survey similar to the pre-survey was sent to participants to measure any change in participants’ perceptions about GenAI. Details on the post-experiment survey can be found in Appendix Section [4.3](#).

We consider the main experimental sample to be those workers who submitted something for both the first and second task. In the first panel of Table [B1](#) we show the drop of at each stage of the experiment. While there was attrition at each stage, it was not significantly different in the treatment and control group⁴. With attrition, this sample leaves 487 across the treatment and control groups. In the second panel of Table [B1](#) we show the remaining sample is well balanced⁵.

3.3 Methods

The main analysis is comparing scores on the three tasks across the treatment and control group. We compare the output of the tasks across an objective benchmark when possible, and for each task we also compare their output to the benchmark set by the data scientists.

3.3.1 Primary outcomes

Our first set of outcomes are related to how the workers perform on a statistics task, a predictions task, and a coding task. We will measure workers’ likelihood of finishing each problem, how long it took them to complete the problems⁶, and how they performed on the problems. For the coding task, there is a conservatively defined correct answer, which receives a ‘1,’ while anything else receives ‘0’ as well as a series of 32 correct steps the worker must take to get to the correct answer. For the predictions task, the output is a vector of probabilities without an objectively correct answer. For this prediction problem, we compare the answer from each worker to the benchmark from the data scientists’ answers, with details in Section [3.3.2](#). The statistical understanding task has both multiple choice

⁴We show that those who finished the experiment and attritors are not different in the treatment and control group in Appendix Table [A14](#).

⁵To account for differential attrition, we report Lee Bounds on our main results in Appendix Table [A15](#) (Lee, 2009) and find all main results retain their significance.

⁶Overall time to complete each task will be measured and logged by Qualtrics.

Table B1: Comparison of worker covariates, by treatment assignment

<i>Flow from initial allocation into analysis sample</i>				
	<i>Total (N)</i>	<i>Treatment (N)</i>	<i>Control (N)</i>	<i>P-value</i>
Total workers allocated	983	493	493	
↪ began survey	573	298	275	0.33
↪ completed first task	511	270	241	0.19
↪ completed second task	487	260	227	0.13
<i>Pre-allocation attributes of the final analysis sample: N = 487</i>				
	Treatment mean: \bar{X}_{TRT}	Control mean: \bar{X}_{CTL}	P-value	
Female	0.369	0.37	0.985	
Bachelors Degree	0.238	0.291	0.192	
Masters Degree	0.677	0.604	0.092	
Doctorate	0.085	0.106	0.428	
Consultant	0.515	0.493	0.361	
Office in Africa	0.019	0.018	0.896	
Office in Asia Pacific	0.135	0.115	0.505	
Office in Central or South America	0.019	0.004	0.14	
Office in Europe or Middle East	0.492	0.52	0.546	
Office in North America	0.335	0.344	0.835	

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

questions with correct and incorrect answers, as well as open ended questions which we have graded by ChatGPT.

Our second set of outcomes test whether or not workers appeared to retain knowledge without ChatGPT. We measure this with a set of questions in the post-experiment survey that both groups are not allowed to use ChatGPT to answer. It is possible that in applying GPT-4 to technical problems the treated workers will retain some knowledge. However it is also possible that even if the treated workers are better at performing data science tasks,

they may not have more knowledge about data science after the tool they learned with is taken away.

Third, we test whether the experience using ChatGPT makes workers better at gauging the bounds of its abilities. In the pre and post experiment survey we provide the workers a series of problems and ask them “How likely is GPT-4 to solve this problem correctly?” We hypothesize that after completing these tasks, the workers in the treatment group will be better at forecasting which types of problems AI is good at solving.

Lastly, we see if this experience increases workers’ confidence in their ability to do data science. We hypothesize that workers who are in the treatment group will have more confidence in their ability to do data science and are more likely to consider moving into more data science heavy roles. We also hypothesize that the treatment group will have higher confidence in their ability to use ChatGPT to help them learn new skills.

3.3.2 Task grading

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

Statistics and machine learning tasks

Each question in the statistics task will be graded against the rubric (shown in Appendix Section 4.4). The rubric scores are a weighted correctness score such that the final score will be determined by a weighted sum across all answers:

$$\text{Total correctness} = \sum_{i=1}^n (\text{Correctness of answer}_i \times \text{Complexity weight}_i) \quad (3.1)$$

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

Predictions tasks

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

$$\text{Loss Score} = \frac{1}{n} \sum_{i=0}^n |\text{score}_i - \text{DS score}_i| \quad (3.2)$$

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

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

Coding task

There is one distinct correct answer for the coding assignment. Correctness is a binary measure where its 0 if wrong and 1 if correct. Second, we will compare the output from the workers and data scientists to a rubric we created with 10 steps, where each step is necessary to get the correct score. As with the statistics task, we grade this against the rubric (shown in Appendix Section .4.4). The rubric scores are a weighted correctness score such that the final score will be determined by a weighted sum across all answers:

$$\text{Total correctness} = \sum_{i=1}^n (\text{Correctness of answer}_i \times \text{Complexity weight}_i) \quad (3.3)$$

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

3.3.3 Estimating treatment effects

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

$$y_i = \beta_0 + \beta_{\text{GPT}}T_{\text{GPT}} + \gamma X_i + \varepsilon_i \quad (3.4)$$

3.3.4 Heterogeneous treatment effects

In addition to the main outcomes above, we plan to explore various factors that may influence the performance and outcomes of the consultants. For example, we can see if consultants with more technical backgrounds or those who are better at guessing what types of problems are within GPT-4’s range of ability have larger treatment effects.

3.4 Results

3.4.1 Treated workers performed better on data science tasks

In Figure 3-2 we show that treated workers were able to more correctly solve all three data science problems than their counterparts in the control group, conditional on submitting anything. We normalize the scores for each to (0,1). Details on each regression can be found in Tables A5 and A6.

Treated workers performed 44 percentage points better than the control group on the coding problem, and received scores which were indistinguishable from those of the data scientists. In the control group no data scientists achieved a perfect score, and in the treatment group only five did. Because of this low level of variation, we will make the primary outcome the percentage of correct steps that the worker took. On average, control group workers took 19% of the correct steps, while the treatment group took 63% of the correct

steps. Not only did the treatment group perform much better, but their score is statistically indistinguishable from the average score of the data scientists.

Treated workers also performed better at the prediction problem. In this problem, their output is a vector of probabilities about the predictability of soccer games. Since there is no ground truth for how predictable a soccer game is, we use the performance of the data scientists as the benchmark for a “correct” answer. The workers’ scores for the prediction problem is the absolute error between their vector and the benchmark vector. For ease of interpretation we make the workers’ score on the prediction problem 1 minus their absolute error, so that 1 is exactly the correct vector of probabilities and 0 is a vector which is orthogonal to the answer. The control group had an average score of 0.43. The treatment group performed better, with an average score of 0.59, however, they were still far from the performance of the data scientists.

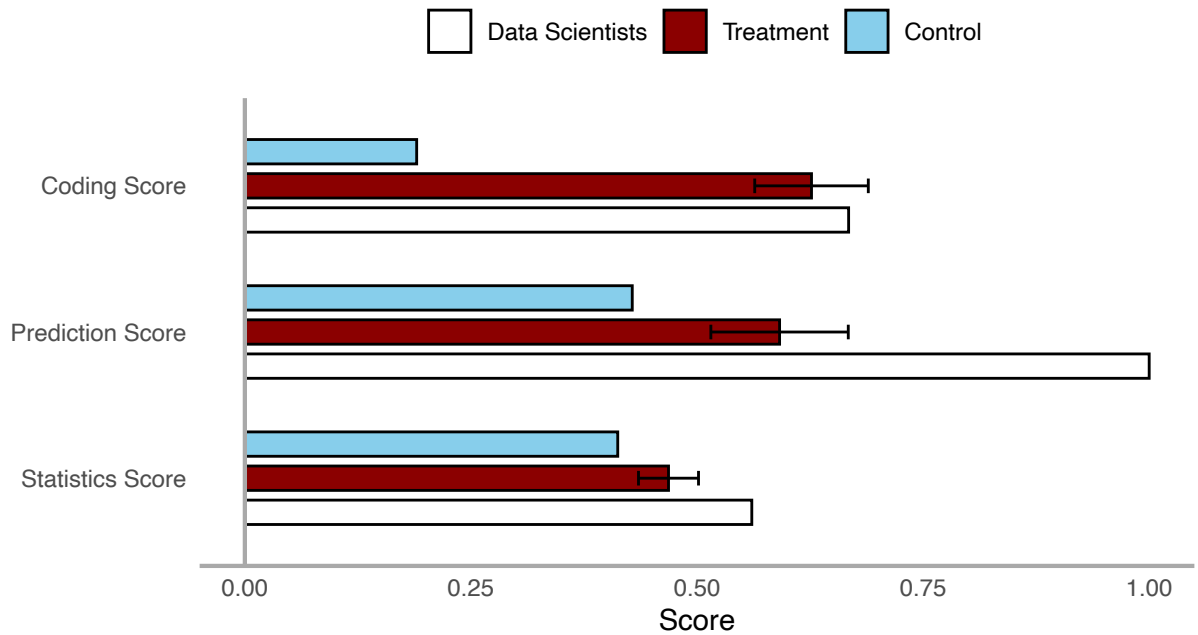
Treated workers performed better on the statistical understanding problem. The statistics problem set included 10 multiple choice and “select all that apply” problems that were graded extremely conservatively for correctness. In the control group, workers got an average score of 41%. The treatment group performed a bit better, with an average score of 46%, and got close to but did not reach the performance of the data scientists.

3.4.2 Treated workers are slightly more likely to complete tasks

In Table [A10](#) we show the effect of the treatment on workers’ probability of submitting any answer for each tasks. This analysis is agnostic to the quality of the answers submitted. Most workers submitted something—78% of workers in the control group assigned the coding tasks submitted it, 85% of workers assigned the statistics problem submitted something, and 87% of workers assigned the prediction problem submitted something. For the statistics and coding problems, the treatment group’s probability of submitting is 6 percentage points higher, although it is no different from the control group on the prediction problem.

In Appendix Table [A14](#) we show that workers who submit something are positively selected—they are more likely to be proficient coders, more likely to have advanced degrees, and are more likely to code for work. However, the those who submit something versus those who attrit look very similar on observables in the treatment and control group. Despite this

Figure 3-2: Effect of AI treatment on workers' ability to solve data science problems



Notes: This plot reports the effect of the treatment on the consultants self reported confidence in their own data science skills. The x-axis is the mean score for each treatment group on each set of problems, where 1 is a perfect score and 0 is the lowest possible score. The first outcome is the sum of the consultant's score on each statistics question, divided by the total number of possible points. The second outcome is the score they got on the prediction problem, which is 1 minus their mean absolute error. The third outcome is the percentage of correct steps they take in answering the coding question. A 95% confidence interval is plotted around the treatment group's mean as compared to the control group mean. The benchmark set by the data scientists is also plotted. The sample includes all experimental participants who submitted something for grading on each task. Text of problems can be found in Appendix Section .4.2. Regression details can be found in Table A5 and Table A6.

balance across treatment and control, for our main outcomes in Figure 3-2 we compute Lee Bounds on the estimates in Appendix Table A15 and find that under conservative estimates, all of our results hold and remain significant (Lee, 2009).

Table D2: Effects of AI to whether or not they get submit any answer on each task

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

Notes: This table analyzes the effect of the treatment on the consultants submitting any answer to each question. Text of problems can be found in Appendix Section 4.2. All regressions include controls for gender, location, native english status, and low tenure. Significance indicators: $p \leq 0.10$: *, $p \leq 0.05$: ** and $p \leq .01$: ***.

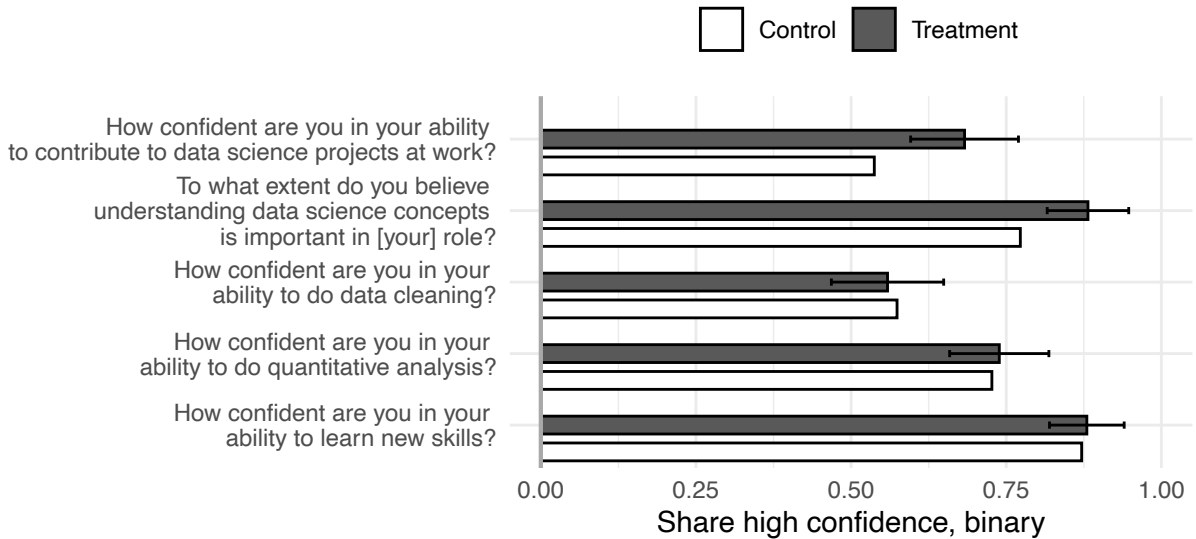
3.4.3 Treated workers completed tasks faster

Table D3: Effects of AI on number of minutes to complete each task

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

Notes: This table analyzes the effect of the treatment on the length of time it took for consultants to finish each task, conditional on completion. The outcome in Column (1) is the number of minutes they spent on the Statistics task. The outcome in Column (2) is the number of minutes spent on the Problem Solving and Prediction task. And the outcome in Column (3) is the number of minutes spent on the Coding task. Consultants were randomly assigned two of the three tasks, and given 90 minutes maximum to complete each. All regressions include controls for gender, location, native english status, and low tenure. Significance indicators: $p \leq 0.10$: *, $p \leq 0.05$: ** and $p \leq .01$: ***.

Figure 3-3: Effect of AI treatment on workers confidence in data science skills



Notes: This plot reports the effect of the treatment on the workers self reported confidence in their own data science skills. Text of questions can be found in Appendix Section .4.3. Regression details can be found in Table A7.

Table D3 shows the effect of the treatment on the length of time workers took completing each task, conditional on submitting something. Workers were allowed to take up to 90 minutes on each task. On average it took workers 63 minutes to complete the statistics task, 68 minutes to complete the prediction task, and 78 minutes to complete the coding task. The treated workers took just as long on the statistical understanding task, however they took 20% less time or 14 fewer minutes on the prediction task, and 12% or 9 fewer minutes on the coding task.

3.4.4 Treated workers are more confident in their data science skills

We find that after submitting their solutions, treated workers are 28% more likely to say that with ChatGPT they are confident in their ability to contribute to data science projects at work. On a likert scale measured from 0 to 7, 54% of workers in the control group say they have “confidence” or “high confidence” (a 5 or above out of 7) on their ability to contribute to data science projects at work with the help of ChatGPT. Treated workers are confident in these abilities 15 percentage points more often, a 28% increase.

3.4.5 No impact to workers ability to answer technical problems without help of ChatGPT

Despite evidence that use of AI made the workers significantly better at solving data science problems, and their confidence in their own data science skills when aided with ChatGPT, after the experiment they were no more likely to be able to answer questions about probabilities, machine learning, or coding without the use of ChatGPT. In the post experiment survey, workers were asked five questions on topics related to their tasks, for example⁷, “Distance-based algorithms are not affected by scaling” and “Which of these following code snippets will give us a dataframe filtered to only rows which correspond to ‘treatment?’” Workers in both groups are instructed not to use ChatGPT to answer these questions.

In Figure 3-4 we report the effect of the treatment on whether the worker correctly answered each question. In this table we define a correct answer conservatively— for “Select all that apply” questions, to receive a correct answer they must select all that are true and none that are false. The treated group performs no better than the control group on these questions⁸.

3.4.6 Treated workers exhibit overconfidence in AI’s current capabilities

Workers in the treatment group perform worse at guessing whether something is within GPT-4’s capabilities after the conclusion of the experiment. Before and after the experiment, workers are posed seven problems and ask to give their opinion on the likelihood that GPT-4 is able to come to the correct conclusion for each problem⁹. Four of these problems GPT-4 reliably gets the answer wrong, while the other three it most often can solve. Prior to the experiment, workers have comparable levels of confidence in GPT-4’s ability to correctly answer each question.

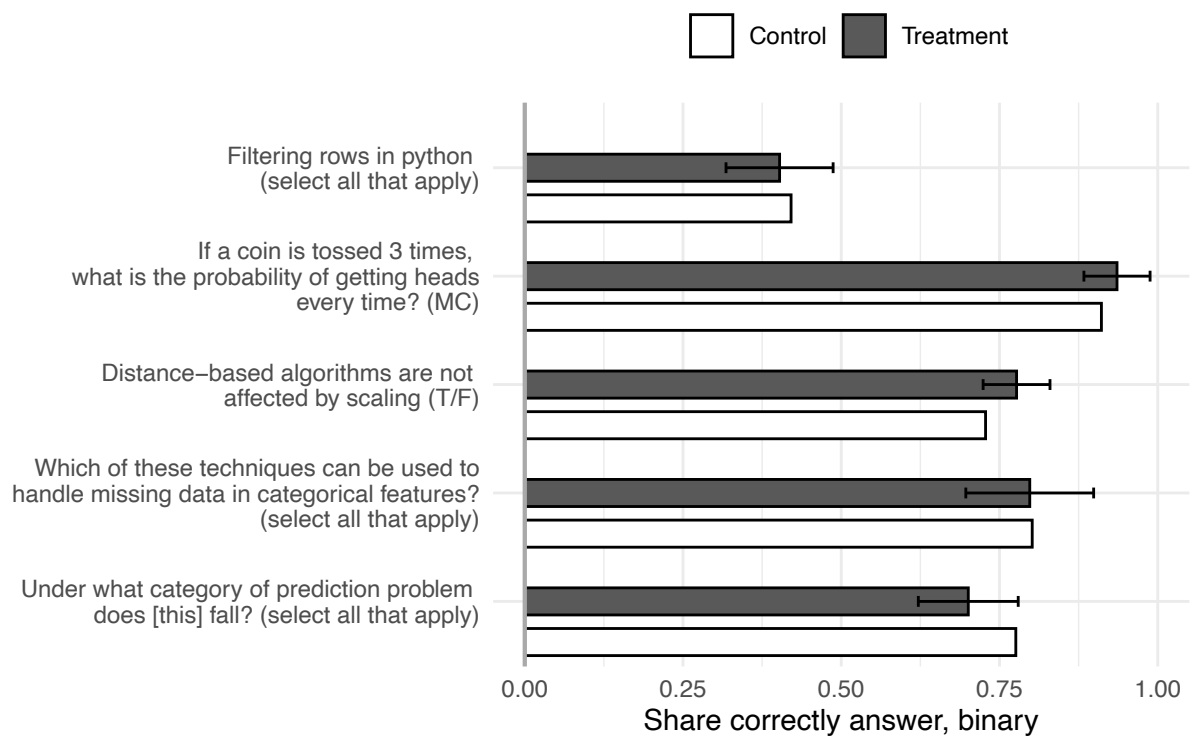
After the experiment, however, workers in the treatment group become significantly more optimistic and significantly more wrong about GPT-4’s capabilities. In Figure 3-5

⁷Full text of questions can be found in Appendix Section .4.3.

⁸Under less conservative definitions, the results to learning are even more precise nulls.

⁹Text of questions can be found in Appendix Section .4.3.

Figure 3-4: Effects of AI treatment to post experiment data science knowledge without use of GenAI



Notes: This analysis looks at the effect of treatment on workers ability to correctly answer data science questions after the conclusion of the experiment. The x-axis is the mean probability of getting the correct answer for each treatment group. The y-axis has the text of each question, with the format of the answer. A 95% confidence interval is plotted around each estimate. Text of questions can be found in Appendix Section .4.3. Regression details can be found in Table A8.

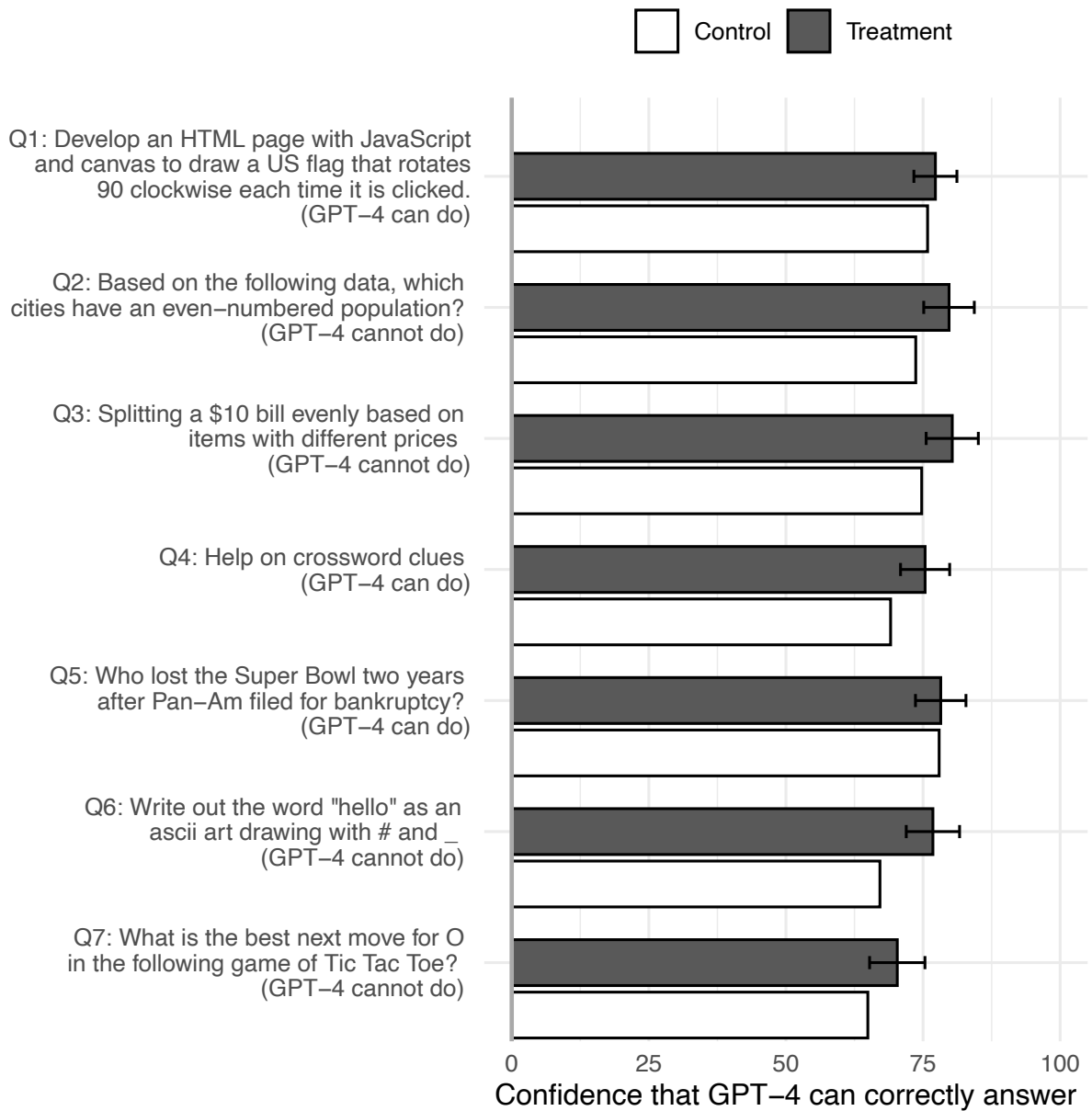
we show the difference between the treatment and control group in percentage points for workers belief each problem can be correctly answered by GPT-4. Workers in both groups were optimistic about GPT-4’s capabilities—the base rate for each problem is between 64 and 78%. For all four of the problems which it cannot solve, treated workers report 5 to 10 percentage points higher likelihood’s that GPT-4 can correctly answer each problem. The only two problems which treated workers are not more optimistic about GPT-4’s capabilities are two of the three problems it actually can correctly solve.

3.4.7 Heterogeneous treatment effects

We find some evidence that the effects to scores are largest for workers with less technical backgrounds. In Table [A12](#) we look at the effect of the treatment interacted with the workers’ pre-experiment coding experience. We find that on the statistics and coding problems, workers who report having no coding experience see the largest treatment effects. Since these workers are also the ones with the lowest scores in the control group, the performance of the treated workers without any coding experience is indistinguishable from the performance of the treated workers who report being proficient coders prior to the experiment.

It is possible that the only workers who can successfully use ChatGPT to learn new skills are the one’s who know what it’s useful for. To test this hypothesis, we interact the treatment with worker’s pre-experiment knowledge of GPT-4’s strengths and weaknesses in Table [A13](#). Using their performance on questions where they have to guess whether or not GPT-4 can correctly answer a question, we find that people who perform well on these questions see no larger treatment effects than those perform poorly.

Figure 3-5: Effect AI treatment on predictions about GPT-4’s capabilities



Notes: This analysis looks at the effect of treatment on worker’s confidence in AI’s capabilities. The x-axis is the difference in the worker’s confidence out of 100 that GPT-4 would be able to answer the question correctly. The text, or a summary of the text, of the question is on the y axis, with whether or not GPT-4 can actually get the correct answer. A 95% confidence interval is plotted around each estimate. Text of questions can be found in Appendix Section 4.3. Regression details can be found in Table A9.

3.5 Conclusion

We run a randomized control trial in the setting of a large managerial consulting firm to understand to what extent GenAI can be used to help non-technical workers perform data science tasks. We find that workers given access to and training in ChatGPT significantly outperform the control group on a series of data science problems, with the largest effects on the coding task where treated workers' performance was statistically indistinguishable from actual data scientists. These findings highlight the potential for AI itself to be used as an on-the-job training tool to help workers adapt to the changing skill demands of the labor market. In this way, AI-enabled upskilling could be an important strategy for workers and firms to avoid job displacement from AI and automation ([Acemoglu and Restrepo, 2018](#); [Djankov and Saliola, 2018](#)).

However, our results also point to some important limitations and concerns around using GenAI to complete work outside of one's skillset. While treated workers could complete the data science tasks with the aid of ChatGPT, they did not demonstrate any greater retention of data science knowledge in post-tests without use of the ChatGPT. This suggests there may be limits to the depth of genuine skill acquisition, at least in the short-term. Moreover, we find that exposure to ChatGPT induced overconfidence in the AI's abilities, with treated workers more likely to believe ChatGPT could solve problems that it in fact could not. This echoes findings from other recent studies on human susceptibility to AI errors and over-reliance on AI assistance ([Agarwal et al., 2023](#); [Wiles and Horton, 2024](#)). As firms seek to implement AI tools like ChatGPT for workforce upskilling, it will be important to develop training approaches that instill a proper understanding of the AI's limitations.

We believe this paper provides a first piece of evidence that GenAI can be used to widen the scope of work that is within workers' skillsets. However, fully realizing GenAI's potential as a tool for upskilling while mitigating the risks of overconfidence and over-reliance will require ongoing research and responsible implementation.

.1 Appendix Tables & Figures

Table A4: P-value of difference between pre-allocations covariates, by treatment assignment

	P-value
Tenure years at BCG	0.797
English proficiency	0.120
Frequency of code for work	0.698
Know how to code	0.814
Python familiarity	0.505
Number of programming languages	0.189
ChatGPT for coding familiarity	0.618
Use ChatGPT or other LLM for work	0.599
Use ChatCGPTor other LLM for personal	0.766

Notes: This table reports means and standard errors of various pre-treatment covariates for the treatment group and the control group, in the final experimental sample where $N = 487$. The reported p-values are for two-sided t-tests of the null hypothesis of no difference in means across groups.

Table A5: Effects of AI to whether or not they get the correct answer on various tasks

	<i>Dependent variable:</i>		
	Stats Task Score	Prediction Task Score	Coding Task Score
	(1)	(2)	(3)
GenAI Treatment Assigned (Trt)	0.056*** (0.017)	0.163*** (0.039)	0.437*** (0.032)
Mean Y in Control Group	0.41	0.43	0.19
Observations	333	318	303
R ²	0.051	0.062	0.392

Notes: This table analyzes the effect of the treatment on the consultants ability to correctly answer questions. The first outcome is the sum of the consultant's score on each statistics question, divided by the total number of possible points. The second outcome is the score they got on the prediction problem, which is 1 minus their mean absolute error from the correct answer. The third outcome is the percentage of correct steps they take in answering the coding question. Exact definition of grading for each problem can be found in Appendix Section .4.2. All regressions include controls for gender, location, native english status, and low tenure. Significance indicators: $p \leq 0.10$: *, $p \leq 0.05$: ** and $p \leq .01$: ***.

Table A6: Effects of AI on distance from workers’ output to data scientists without AI

	<i>Dependent variable:</i>		
	Stats Distance	Prediction Distance	Coding Distance
	(1)	(2)	(3)
GenAI Treatment Assigned (Trt)	-0.056*** (0.017)	-0.163*** (0.039)	-0.437*** (0.032)
Mean Y in Control Group	0.15	0.57	0.48
Observations	333	318	303
R ²	0.051	0.062	0.392

Notes: This table analyzes the effect of the treatment on the consultants ability to correctly answer questions, relative to a benchmark set by data scientists without AI. The first outcome is the points the worker got on the statistics question, divided by the maximum points. The second outcome is their mean absolute error of the distance from their answer to the data scientist’s answer. The third outcome is the percentage of correct steps they took on the coding problem, weighted by difficulty. Exact definition of grading for each problem can be found in Appendix Section .4.2. All regressions include controls for gender, location, native english status, and low tenure. Significance indicators: $p \leq 0.10$: *, $p \leq 0.05$: ** and $p \leq .01$: ***.

Table A7: Effects of AI to self reported confidence in data science skills, binary outcome

	<i>Dependent variable:</i>				
	Own DS Skills	Importance of DS	Various tech abilities		
	(1)	(2)	(3)	(4)	(5)
GenAI Treatment Assigned (Trt)	0.146*** (0.044)	0.109*** (0.034)	-0.015 (0.046)	0.011 (0.041)	0.008 (0.031)
Mean Y in Control Group	0.54	0.77	0.57	0.73	0.87
Observations	465	479	461	466	470
R ²	0.057	0.055	0.030	0.038	0.006

Notes: This table analyzes the effect of the treatment on the share of consultants who report being confidence in their technical abilities. The outcome in Column (1) is a positive answer to the question “How confident are you in your ability to contribute to data science projects?” The outcome in Column (2) is their answer to “To what extent do you believe understanding data science concepts is important in the role of a BCG consultant?” The outcomes in Column’s (3) - (5) are on their confidence in their ability to do data cleaning, quantitative analysis, and learn new skills, respectively. Outcomes are on a scale from 1 to 7, with “High confidence” defined as 1 if they answer 5,6, or 7, and 0 otherwise. All regressions include controls for gender, location, native english status, and low tenure. Significance indicators: $p \leq 0.10$: *, $p \leq 0.05$: ** and $p \leq .01$: ***.

Table A8: Effects of AI treatment to post experiment data science knowledge without use of AI

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

Notes: This table analyzes the effect of the treatment on the consultants ability to answer data science and coding questions, after the conclusion of the experiment. Text of questions can be found in Appendix Section 4.3. All regressions include controls for gender, location, native english status, and low tenure. Significance indicators: $p \leq 0.10$: *, $p \leq 0.05$: ** and $p \leq .01$: ***.

Table A9: Effects of AI treatment to post experiment questions about GPT-4’s capabilities

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

Notes: This table analyzes the effect of the treatment on the consultants confidence in GPT-4’s ability to get the right answer on various questions, after the conclusion of the experiment. For each question, the consultant gave a percentage confidence in GPT-4’s ability to answer the question correctly. The question in Columns 1 and 4, and 5 GPT-4 usually get correct. Questions 2,3,6 and7, GPT-4 almost never get correct. Text of questions can be found in Appendix Section .4.3. All regressions include controls for gender, location, native english status, and low tenure. Significance indicators: $p \leq 0.10$: *, $p \leq 0.05$: ** and $p \leq .01$: ***.

Table A10: Effects of AI to whether or not they get submit any answer on each task

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

Notes: This table analyzes the effect of the treatment on the consultants submitting any answer to each question. Text of problems can be found in Appendix Section 4.2. All regressions include controls for gender, location, native english status, and low tenure. Significance indicators: $p \leq 0.10$: *, $p \leq 0.05$: ** and $p \leq .01$: ***.

Table A11: Effects of AI to whether or not they get submit any answer on each task

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

Notes: This table analyzes the effect of the treatment on the consultants submitting any answer to each question. Text of problems can be found in Appendix Section 4.2. All regressions include controls for gender, location, native english status, and low tenure. Significance indicators: $p \leq 0.10$: *, $p \leq 0.05$: ** and $p \leq .01$: ***.

Table A12: Effects of AI to whether or not they get the correct answer on various tasks

	<i>Dependent variable:</i>		
	Stats Task Score	Prediction Task Score	Coding Task Score
	(1)	(2)	(3)
GenAI Treatment Assigned (Trt)	0.127*** (0.029)	0.077 (0.073)	0.519*** (0.056)
Coding basics	0.069** (0.030)	0.012 (0.068)	0.083 (0.060)
Proficient coder	0.144*** (0.029)	0.010 (0.075)	0.148** (0.059)
Coding basics x Trt	-0.095** (0.041)	0.105 (0.095)	-0.170** (0.080)
Proficient coder x Trt	-0.098** (0.040)	0.127 (0.100)	-0.082 (0.078)
Mean Y in Control Group	0.41	0.43	0.19
Observations	333	318	303
R ²	0.136	0.076	0.420

Notes: This table analyzes the effect of the treatment on the consultants ability to correctly answer questions, by their pre-experiment coding knowledge. The omitted variable is “No prior coding experience.” The first outcome is the sum of the consultant’s score on each statistics question, divided by the total number of possible points. The second outcome is the score they got on the prediction problem, which is 1- their mean absolute error. The third outcome is the percentage of correct steps they take in answering the coding question. Exact definition of grading for each problem can be found in Appendix Section .4.2. All regressions include controls for gender, location, native english status, and low tenure. Significance indicators: $p \leq 0.10$: *, $p \leq 0.05$: ** and $p \leq .01$: ***.

Table A13: Effects of AI to whether or not they get the correct answer on various tasks

	<i>Dependent variable:</i>		
	Stats Task Score	Prediction Task Score	Coding Task Score
	(1)	(2)	(3)
GenAI Treatment Assigned (Trt)	0.064*** (0.021)	0.161*** (0.049)	0.417*** (0.039)
Knowledge of GPT's strengths	0.033 (0.026)	0.005 (0.057)	0.060 (0.050)
Knowledge of GPT's strengths x Trt	-0.023 (0.036)	0.009 (0.080)	0.049 (0.067)
Mean Y in Control Group	0.41	0.43	0.19
Observations	333	318	303
R ²	0.056	0.063	0.407

Notes: This table analyzes the effect of the treatment on the consultants ability to correctly answer questions, by their pre-experiment coding knowledge. “Knowledge of GPT’s strengths” is 1 if the consultant got 4 out of 7, or more of questions correct on the pre-experiment survey asking about their guesses of whether or not GPT-4 can correctly answer a question. The omitted variable is consultants who got fewer than 4 out of 7 correct. The first outcome is the sum of the consultant’s score on each statistics question, divided by the total number of possible points. The second outcome is the score they got on the prediction problem, which is 1 minus their mean absolute error. The third outcome is the percentage of correct steps they take in answering the coding question. Exact definition of grading for each problem can be found in Appendix Section .4.2. All regressions include controls for gender, location, native english status, and low tenure. Significance indicators: $p \leq 0.10$: *, $p \leq 0.05$: ** and $p \leq .01$: ***.

Table A14: Comparing those who submit something for both tasks (primary analysis sample) to attritors

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

Notes: This table reports the mean of various pre-experiment covariates amongst the primary analysis sample with those who attrit. The primary analysis sample is made up of workers who submitted anything to be graded for both of their two tasks. The sample here is made up of all workers who started the pre-experiment survey, $N = 573$. We run a Welch Two Sample t-test on each covariate for attritors and non-attritors, within each treatment group. Significance indicators: $p \leq 0.10$: *, $p \leq 0.05$: ** and $p \leq .01$: ***.

Table A15: Lee Bounds on Treatment Effects for Main Results

	Treatment effect	Lee Lower Bound	Lee Upper Bound
Statistics Correctness Score	0.065*** (0.017)	0.056** (0.024)	0.076*** (0.034)
Coding Process Score	0.45*** (0.033)	0.43*** (0.042)	0.48*** (0.056)
Prediction Task Score	0.15*** (0.038)	0.15*** (0.05)	0.16*** (0.05)

Notes: This table analyzes the effect of the treatment on the consultants ability to correctly answer questions. The first outcome is the sum of the consultant’s score on each statistics question, divided by the total number of possible points. The second outcome is the percentage of correct steps they take in answering the coding question. The third outcome is the score they got on the prediction problem, which is their mean absolute error multiplied by -1 for ease of interpretation. Exact definition of grading can be found in Appendix Section .4.2. All regressions include no controls. Significance indicators: $p \leq 0.10$: *, $p \leq 0.05$: ** and $p \leq .01$: ***.

.2 Pre-Experiment Survey: Registration for Generative AI Experiment with OpenAI

Thank you again for taking part in the Generative AI Experiment! The following questionnaire will take roughly 30 minutes to complete and contains questions about your background and your experiences. Please take the time to thoroughly and thoughtfully respond to these questions, as it is a crucial part of the overall experiment. We ask that you please take this questionnaire in one sitting, before February 16, 2024.

Please note that by submitting this questionnaire, you agree to not discuss the contents of the experiment to anyone, inside or outside of BCG. This is crucial for experimental integrity, to ensure robustness of the results for scientific publication.

Data Use and Collection:

All data collected in this questionnaire will NOT be used for any other purposes other than this Generative AI experiment. Any data that is published internally to BCG, in scientific journals or alike will only be done so in aggregate, and personal information will never be released. This data will also only be shared with OpenAI in aggregate and personal information will not be released outside of BCG/BHI. Within the scope of this questionnaire, we will only collect your personal data, listed below.

- Name
- Email
- Location
- Gender
- Tenure
- Title
- English proficiency
- Education
- Proficiency and orientation towards tech

Your personal data will only be used for testing the hypotheses of this Generative AI experiment, within the scope of your employment contract. We will process your personal data in accordance with applicable data protection laws and BCG's Privacy Policy [Link to internal policy]

CDC Contribution:

As mentioned in the email, successful completion of participation in the study will count as an "office contribution" to your CDC to reflect our appreciation for your efforts. You will have the opportunity to provide your CDA details after completing the study. However, to avail of this opportunity, you must put in an "honest effort" throughout, as judged by the quality of your responses.

If there are any questions at all, please contact Lisa Kraymer (kramer.lisa@bcg.com)

Survey

Demographics (Role and Location)

1. Please Provide your Name First Name _____

Last Name _____

2. Please Provide your BCG Email Address Below

3. Please Select Your Home BCG Office Location

- Africa
- Asia Pacific
- Central & South America
- Europe & The Middle East
- North America

4. Please Select Your Home BCG Office (Conditionally Shown if: (2 = Africa))

- Cairo
- Casablanca
- Johannesburg
- Lagos
- Luanda
- Nairobi
- Other (Please Elaborate) _____

5. Please Select Your Home BCG Office (Conditionally Shown if: (2 = Asia Pacific))

- Auckland
- Bangkok
- Beijing
- Bengaluru
- Canberra

- Chennai
- Fukuoka
- Ho Chi Minh City
- Hong Kong
- Jakarta
- Kuala Lumpur
- Kyoto
- Manila
- Melbourne
- Mumbai
- Nagoya
- Gurugram
- New Delhi
- Osaka
- Perth
- Seoul
- Shanghai
- Shenzhen
- Singapore
- Sydney
- Taipei
- Tokyo
- Other (Please Elaborate) _____

6. Please Select Your Home BCG Office (Conditionally Shown if: (2 = Central & South America))

- Bogota
- Buenos Aires
- Lima
- Panama City
- Rio De Janeiro
- Santiago
- San Jose
- Sao Paulo
- Other (Please Elaborate) _____

7. Please Select Your Home BCG Office (Conditionally Shown if: (2 = Europe & The Middle East))

- Abu Dhabi
- Amsterdam
- Athens
- Baku
- Barcelona
- Berlin
- Brussels
- Budapest
- Cologne
- Copenhagen
- Doha
- Dubai
- Dusseldorf
- Frankfurt

- Geneva
- Hamburg
- Helsinki
- Istanbul
- Lisbon
- London
- Madrid
- Milan
- Munich
- Oslo
- Paris
- Prague
- Riyadh
- Rome
- Stockholm
- Stuttgart
- Tel Aviv
- Vienna
- Warsaw
- Zurich
- Other (Please Elaborate) _____

8. Please Select Your BCG Affiliation Below

- Traditional BCG Consulting Team
- BCG X
- BCG Platinion

- Other (Please Specify) _____

9. Please Select Your Official Title at BCG

- Associate
- Consultant
- BCG X Data Scientist
- BCG X Senior Data Scientist
- Other (Please Specify) _____

10. Please Select Your Total Tenure at BCG (in Years)

- 0 to 1 Years
- 1 Years to 2 Years
- 2 Years to 3 Years
- 3 Years to 4 Years
- 4 Years to 5 Years
- 5+ Years

Demographics (Education and Language)

1. What is your gender?

- Female
- Male
- Prefer Not to Say
- Other

2. What is your English proficiency? (Reading, Written, and Spoken Combined)

- Beginner
- Intermediate

- Advanced
- Native

3. What is your highest education level?

- Bachelors
- Masters
- Professional Degree (e.g., MD, JD etc.)
- Doctorate

4. If you have a Bachelors degree, what was your major? Select the applicable categories and specify your degree in the text box.

- Science and Mathematics _____
- Engineering and Technology _____
- Health Sciences _____
- Social Sciences _____
- Business and Economics _____
- Arts and Humanities _____
- Education _____
- Agriculture and Environmental Studies _____
- Other _____

5. If you have a Masters degree, what was your major? Select the applicable categories and specify your degree in the text box. (Conditionally Hidden if: (12 = Bachelors))

- Science and Mathematics _____
- Engineering and Technology _____
- Health Sciences _____
- Social Sciences _____

- Business and Economics _____
- Arts and Humanities _____
- Education _____
- Agriculture and Environmental Studies _____
- Other _____

6. If you have a Professional degree, what was your major? Select the applicable categories and specify your degree in the text box. (Conditionally Hidden if: (12 = Bachelors OR 12 = Masters))

- Science and Mathematics _____
- Engineering and Technology _____
- Health Sciences _____
- Social Sciences _____
- Business and Economics _____
- Arts and Humanities _____
- Education _____
- Agriculture and Environmental Studies _____
- Other _____

7. If you have a Doctorate degree, what was your major? Select the applicable categories and specify your degree in the text box. (Conditionally Hidden if: (12 = Bachelors OR 12 = Masters))

- Science and Mathematics _____
- Engineering and Technology _____
- Health Sciences _____
- Social Sciences _____
- Business and Economics _____

- Arts and Humanities _____
- Education _____
- Agriculture and Environmental Studies _____
- Other _____

Programming Proficiency

17. What tools do you currently use for data analysis?

- Excel
- Tableau
- Alteryx
- Programming (e.g. Python)
- ChatGPT
- Other LLMs / LLM based tools
- Other non-LLM based tools

18. Do you know how to code?

- Yes, I am an expert level coder
- I know how to code, but am not an expert
- I only know the basics of coding
- No, I do not know how to code

19. How often do you code for work? (Conditionally Hidden if: (17 = I know how to code, but am not an expert))

- I never code for work
- I code occasionally, but usually use other analytics tools
- I code every time I work on analytical projects, but this is only occasionally

- I code every time I work on analytical projects and I frequently am staffed on analytical projects
- Coding is a core part of my job

20. How many years of programming experience do you have? (Conditionally Hidden if: (17 = I know how to code, but am not an expert))

- 0-1
- 2-3
- 3-5
- 5-8
- 8+

21. How many programming languages are you familiar with? (Conditionally Hidden if: (17 = I know how to code, but am not an expert))

- 0
- 1
- 2-3
- 4+

22. How familiar are you with Python? (Conditionally Hidden if: (17 = I know how to code, but am not an expert))

- 0 = I Do Not Program
- 1 = Low Familiarity/Novice
- 2
- 3
- 4
- 5 = High Familiarity/Expert

ChatGPT Proficiency

23. How often do you use ChatGPT or other LLMs for work?

- I have never used ChatGPT
- I have tried ChatGPT once or twice
- I use ChatGPT less than once per week
- I use ChatGPT at least once per week
- I use ChatGPT every day

24. How often do you use ChatGPT or other LLMs in your personal life?

- I have never used ChatGPT
- I have tried ChatGPT once or twice
- I use ChatGPT less than once per week
- I use ChatGPT at least once per week
- I use ChatGPT every day

25. Please rate the extent to which you agree or disagree with the following statements

(1-7 Rating)

- I am familiar with GenAI for writing
- I am familiar with using GenAI for coding
- I am familiar with prompt engineering (i.e., crafting prompts to get a better answer from an AI model)
- I am familiar with more than 2 prompting strategies
- ChatGPT helps me become a better consultant
- I understand how large language models (LLMs), which underpin generative AI tools for writing, work
- I believe I can tell when ChatGPT is hallucinating

- I have created a specialized GPTs for my purposes
- I have used ChatGPT with Code Interpreter / Advanced Data Analytics
- I use ChatGPT for writing code

26. Please rate the extent to which you agree or disagree with the following statements
(1-7 Rating)

- ChatGPT is primarily a Data Science tool
- ChatGPT is primarily a tool for writing
- ChatGPT helps me be more proficient at problem solving
- ChatGPT helps me be more efficient at creating slides

Here's the LaTeX version of the provided text:

Tech Openness and Playfulness

27. Please rate the extent to which you agree or disagree with the following statements
(1-7 Rating)

- If I hear about a new technology product or service, I will look for ways to experiment with it
- Among my peers, I am usually the first to try out new technology products and services
- In general, I am hesitant to try out new technology products and services
- I like to experiment with new technology products and services
- I am spontaneous when I interact with new technology products or services
- I am unimaginative when I interact with new technology products or services
- I am playful when I interact with new technology products or services
- I am flexible when I interact with new technology products or services
- I am uninventive when I interact with new technology products or services

- I am creative when I interact with new technology products or services
- I am unoriginal when I interact with new technology products or services

Creativity

28. Please rate the extent to which you agree or disagree with the following statements
(1-7 Rating)

- I try not to oppose team members
- I adapt myself to the system
- I adhere to accepted rules in my area of work
- I avoid cutting corners
- I am thorough when solving problems
- I address small details needed to perform the task
- I perform the task precisely over a long time
- I am good in tasks that require dealing with details
- I have a lot of creative ideas
- I prefer tasks that enable me to think creatively
- I am innovative
- I like to do things in an original way

Learning Orientation

29. Please indicate the extent to which you agree or disagree with the following statements (1-7 Rating)

- I enjoy learning new topics
- I like to read diverse topics
- I find pleasure in learning
- I get intrinsically motivated to constantly expand my knowledge

- I seek deep-seated conceptual knowledge for the task assigned to me
- I spend a lot of time thinking about how my performance is in comparison to others
- I like to seek rewards in short term for my efforts
- I prefer to see tangible output as a reward for my effort
- I generally perform and undertake those tasks for which I get rewarded soon
- I feel very good when I know I have outperformed other colleagues
- I always try to communicate my achievements to my friends and supervisors

Concluding Remarks

30. Do you agree not to discuss the contents of this experiment with anyone, inside or outside of BCG? This is crucial for experimental integrity, to ensure robustness of the results for scientific publication.

.3 Main experiment

.3.1 Survey

Introduction - consent

Welcome to the ChatGPT / Generative AI Experiment!

We are thrilled to have you begin this study. Before you begin, please read through the following notes:

Goal of Study:

This is a scientific study conducted in collaboration with researchers from BCG and other institutions. We hope to publish the results from this experiment in a leading academic journal.

Due to the rigorous nature of academia, and the high standard needed for peer-reviewed publications, we ask for your full engagement and feedback. Please see the note about CDC contribution below for those that put in an honest effort.

This experiment will take you roughly 4 hours (or less) to complete.

Confidentiality:

Please DO NOT discuss the details of this study with anyone, either among your peers or anyone inside or outside of BCG, even after they may have completed their participation. This seriously compromises the integrity of the full study. We want to absolutely avoid this.

Data Collected during Study:

During the study, you will be given a short survey, series of tasks to perform and another short survey towards the end. For each task, you will type your answers in response. All data and information are fictional.

Your responses will be evaluated by a combination of humans and algorithms. All personal or identifying information will be scrubbed prior to this evaluation process.

Data Usage:

Aggregate and deidentified information collected from this survey will be used for research purposes. All efforts will be made to keep your study-related information confidential. In particular, we will work to make sure that your responses are not accessed by anyone outside the research team.

Your personal data will ONLY be used to communicate your office contribution with your CDA and in case we need to have a follow-up interview or survey.

CDC Office Contribution and Other Incentives:

As a token of our appreciation for your commitment, we are offering the following incentives for successful completion of this experiment:

1. CDC office contribution “recognition for anyone who puts in an honest effort” into all aspects of the experiment (including the follow-up interview, to be scheduled), as judged by the quality of their answers
2. In addition to the above, participants in the top 50th percentile as judged by quality of answers, with access to similar resources, will be noted as such to their CDA
3. In addition to the above, participants with access to similar resources with extraordinary performance will be commemorated with a BCG leadership recognition, a small group chat with OpenAI and OpenAI merchandise.

Note that participation is totally voluntary and there’s no repercussions in case you decide to end your participation before finishing it. However, this would not count as an office contribution.

Once you have blocked 4 hours of uninterrupted time, you may start by continuing.

1. You cannot participate in this study on a phone, tablet etc. Please only proceed when you are logged in via your laptop/computer with a stable internet connection
 - Yes, I am logged in via my laptop or my computer

Please type your BCG email address below to proceed _____

CDC Office Contribution

If you’d like for your participation in this study to count as an “office contribution” as described above, please type in your CDA’s BCG email address below.

If you do not want this, please type “N/A” _____

Overall flow of the study and expectations

Your participation in this study will take approximately 4 hours and will consist of 7 sections:

- Filling out a short pre-survey (~10 min)

- Going over a short training (~15 min)
- Optional break (~10 min)
- Complete the first task (~90 min)
- Break (~10 min)
- Completing the second task (~90 min)
- Filling out post-survey (~15 min)

The tasks you will complete are independent of each other and unrelated to other survey components.

We highly encourage you to do your best to complete these tasks and while it might be challenging sometimes, we truly appreciate the effort. Don't forget that top 50th percentile, as judged by quality of answers, with access to similar resources, will be noted as such to their CDA.

.4 Main experiment

.4.1 Pre-Experiment Survey

GenAI_DataScience_Prod_GPT

Start of Block: Welcome

Welcome **Welcome to the Upskilling Study!**

Thank you so much for taking your time to support this project. Your participation is critical to BCG's success as a thought leader in Generative AI.

Page Break

Consent **Goal of Study:**

This is a scientific study conducted in collaboration with researchers from BCG, OpenAI, and other institutions. We hope to publish the results from this study in a leading academic journal.

Due to the rigorous nature of academia, and the high standard needed for peer-reviewed publications, we ask for your full engagement and feedback. Please see the note about CDC contribution below for those that put in an honest effort.

We anticipate that participation will take you roughly 4 hours (or less) to complete.

Confidentiality:

Please DO NOT discuss the details of this study with anyone, either among your peers or anyone inside or outside of BCG, even after they may have completed their participation. This seriously compromises the integrity of the full study. We want to absolutely avoid this.

Data Collected during Study:

During the study, you will be given a short survey, series of tasks to perform and another short survey towards the end. For each task, you will type your answers in response.

Your responses will be evaluated by a combination of humans and algorithms. All personal or identifying information will be scrubbed prior to this evaluation process.

All data will be aggregated and any personal identifiable information will be removed before sharing with any external collaborators, including OpenAI.

Data Usage:

Aggregate and deidentified information collected from this survey will be used for research purposes. All efforts will be made to keep your study-related information confidential. In particular, we will work to make sure that your responses are not accessed by anyone outside the research team.

Your personal data will ONLY be used to communicate your office contribution with your CDA and in case we need to have a follow-up interview or survey.

All data collected in this questionnaire will NOT be used for any other purposes other than this Generative AI experiment. Any data that is published internally to BCG, in scientific journals or alike will only be presented in aggregate, and personal information will never be released. This data will also only be shared with OpenAI after it is aggregated and personal information will

not be released outside of BCG/BHI. Within the scope of this questionnaire, we will only collect your name, email and technical background.

Your personal data will only be used for testing the hypotheses of this Generative AI experiment, within the scope of your employment contract. We will process your personal data in accordance with applicable data protection laws and [BCG's Privacy Policy](#).

CDC Office Contribution and Other Incentives:

As a token of our appreciation for your commitment, we are offering the following incentives for successful completion of your participation: CDC "office contribution" recognition for anyone who puts in an "honest effort" into all aspects of the study (including the follow-up interview, to be scheduled), as judged by the quality of their answers. In addition to the above, participants in the **top 50th percentile** as judged by quality of answers, with access to similar resources, will be noted as such to their CDA. In addition to the above, participants with access to similar resources **with extraordinary performance** will be commemorated with a BCG leadership recognition, and a small group chat with OpenAI. Note that participation is totally voluntary and there are no repercussions in case you decide to end your participation before finishing it. However, this would not count as an office contribution.

Once you have blocked 4 hours of uninterrupted time, you may start by continuing.

Page Break

LaptopUse You cannot participate in this study on a phone, tablet etc. Please only proceed when you are logged in via your laptop/computer with a stable internet connection

Yes, I am logged in via my laptop or computer (1)

Email Please type your BCG email address below to proceed

CdcContribution CDC Office Contribution

If you'd like for your participation in this study to count as an "office contribution" as described on the previous page, please type in your CDA's BCG email address below. If you do not want this, please type "N/A"

Page Break

Overview **Approximate flow of the study and what to expect.**

You can expect this study to take approximately 4 hours. It consists of 7 distinct sections:

Pre-survey (~10 min) Training (~15 min) Optional break (~10 min) First task (~90 min) Break (~10 min) Second task (~90 min) Post-survey (~15 min)

Note that the tasks are completely independent of each other and unrelated to other survey components.

Please keep in mind that you cannot go backwards in this survey. Once you hit next, you will not be able to return. Please complete each page before moving on.

We highly encourage you to do your best to complete these tasks and while it might be challenging sometimes, we truly appreciate the effort. Don't forget that **top 50th percentile** will be noted as such to their CDA.

End of Block: Welcome

Start of Block: Pre-Survey



PreSurTaskOnLoadTime PreSurveyTaskOnLoadTimeTracker

PreSurvey **Pre-Survey**

First, we would like you to answer a few survey questions



NeedCognition **Please indicate the extent to which you agree or disagree with the following statements:**

	Strongly disagree (0)	Disagree (1)	Neither agree nor disagree (2)	Agree (3)	Strongly agree (4)
I would prefer complex to simple problems (1)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I like to have the responsibility of handling a situation that requires a lot of thinking (2)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Thinking is not my idea of fun (4)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I would rather do something that requires little thought than something that is sure to challenge my thinking abilities (5)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I really enjoy a task that involves coming up with new solutions to problems (6)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I would prefer a task that is intellectual, difficult, and important to one that is somewhat important but does not require much thought (7)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>



ConsistencyInterest **Please indicate the extent to which you agree or disagree with the following statements:**

	Strongly disagree (0)	Disagree (1)	Neither agree nor disagree (2)	Agree (3)	Strongly agree (4)
I often set a goal but later choose to pursue a different one (1)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I have been obsessed with a certain idea or project for a short time but later lost interest (8)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I have difficulty maintaining my focus on projects that take more than a few months to complete (2)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
New ideas and projects sometimes distract me from previous ones (4)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
My interests change from year to year. I become interested in new pursuits every few months (5)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>



Perseverance Please indicate the extent to which you agree or disagree with the following statements:

	Strongly disagree (0)	Disagree (1)	Neither agree nor disagree (2)	Agree (3)	Strongly agree (4)
I finish whatever I begin (1)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Setbacks don't discourage me (8)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I am diligent (2)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I am a hard worker (4)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I have achieved a goal that took years of work (5)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I have overcome setbacks to conquer an important challenge (7)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>



DataScienceSkills In what aspects of data science do you have experience?

	No experience (0)	Somewhat experienced (1)	Very experienced (2)
Data visualization (1)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Machine learning models (2)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Statistical analysis (3)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Data cleaning and preparation (4)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>



DSCconfidence On a scale of 1-7, where 1 = "Not at all" and 7 = "Extremely", please rate the following.

	1	2	3	4	5	6	7
How confident are you in your ability to contribute to data science projects? ()							
To what extent do you believe understanding data science concepts is important in the role of a BCG A/C? ()							

DataScienceTools **What tools do you currently use for data analysis? Select all that apply.**

- Excel (1)
 - Tableau (2)
 - Alteryx (3)
 - Programming (4)
 - ChatGPT (5)
 - Other LLMs / LLM based tools (6)
 - Other non-LLM based tools (7)
-



ExcelFrequency **How frequently do you use Excel, Tableau or Alteryx in your day-to-day work?**

- Daily (5)
 - Several times a week (4)
 - Once a week (3)
 - A few times a month (2)
 - Rarely (once a month or slightly less) (1)
 - Never (0)
-



QuantExpertise **Please indicate the extent to which you agree or disagree with the following statements:**

	Strongly agree (4)	Somewhat agree (3)	Neutral (2)	Somewhat disagree (1)	Strongly disagree (0)
I consider myself an expert in Excel (1)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I consider myself an expert in Tableau (2)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I consider myself an expert in Alteryx (3)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Based on prior CDC reviews, PSI (problem solving and insights) has been a core strength (4)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>



CodingPre **Do you know how to code?**

- Yes, I am an expert level coder (3)
- I know how to code, but am not an expert (2)
- I only know the basics of coding (1)
- No, I do not know how to code (-1)

Page Break



ProfessionalIdPre1 **Please indicate the extent to which you agree or disagree with the following statements:**

	Strongly Agree (4)	Somewhat Agree (3)	Neutral (2)	Somewhat Disagree (1)	Strongly Disagree (0)
Generative AI helps me feel valuable in my role (1)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Generative AI elevates how important I feel my job is for society (2)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Generative AI elevates my professional status and level of influence within my organization (3)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Generative AI helps me feel more competent in my role (4)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Generative AI enables my ability to execute tasks and reach desired outcomes (5)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Generative AI enables my ability to execute data analytics tasks and reach desired outcomes (6)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Generative AI increases the value I place on my expertise and skill cultivation (7)

Generative AI increases my level of autonomy in making individual decisions in my role (8)

Generative AI helps me be more confident that I will meet my project managers expectations (9)

Generative AI enables me to do what I really want to do in my role (11)

Generative AI will change the dynamic in my team (14)

Generative AI improved how I perceive my role in the organization (15)



ProfessionalIdPre2 **Please indicate the extent to which you agree or disagree with the following statements:**

	Strongly Agree (4)	Somewhat Agree (3)	Neutral (2)	Somewhat Disagree (1)	Strongly Disagree (0)
Using Generative AI helps me stay aligned with my project managers expectations (10)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I believe using Generative AI will contribute to the betterment of others in my work (12)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I see Generative AI as my coworker (13)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I would recommend Generative AI to other consultants (16)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I am proud of BCG's approach to Generative AI adoption within the firm (17)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I believe BCG is at the leading edge of the Generative AI revolution (18)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
My managers and supervisors will expect more output from me because of Gen AI (19)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Sustained use of ChatGPT for data science would have the potential to make me a better consultant in the 'Problem solving and insights' dimension (20)

Sustained use of ChatGPT for data science would have the potential to make me a better consultant in the 'Communication and Presence' dimension (21)

Sustained use of ChatGPT for data science would have the potential to make me a better consultant in the 'Practicality and Effectiveness' dimension (22)



**GenAIUsage Rate how helpful you think Generative AI tools are for these use cases
(Rating 1-7, where 1 = Not at all helpful; 7 = Extremely helpful; with ability to say "I don't know")**

	1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	7 (7)	I don't know (-1)
Write a first draft for simple texts (e.g., emails) (1)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Write a first draft for complex texts (e.g., reports) (2)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Write my final version for simple texts (e.g., emails) (3)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Write my final version for complex texts (e.g., reports) (4)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Review my writing (grammar, typos, etc.) (5)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Be more persuasive (6)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Brainstorm ideas (7)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Be more creative (8)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Writing code for data analytics (9)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Writing code for data visualizations (10)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Learning how to use excel for data analysis and visualizations (11)

Identifying which machine learning models to use for a project (12)

Understanding the statistical significance of a result (13)

Writing code for data cleaning and preparation (14)



GenAllImpactPre **Since implementing Generative AI, how have your project teams been affected? Mark the position of the team relative to the description on the left and the description on the right**

1 2 3 4 5 6 7

	1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	7 (7)	
Decreased collaboration	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Increased collaboration
Decreased efficiency	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Increased efficiency
Decreased clarity of responsibilities	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Increased clarity of responsibilities
Decreased learning	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Increased learning
Decreased decision quality	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Increased decision quality
Reduced team morale	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Improved team morale

GenAIBenefitsPre In a few words, what do you think will be the biggest benefits of Generative AI for you?

GenAIRisksPre In a few words, what do you think will be the biggest risks of Generative AI for you?

GenAIRolePre Given the capabilities of Generative AI, do you see the role of associates and consultants evolving in the next 5 years? If so, how?

Page Break

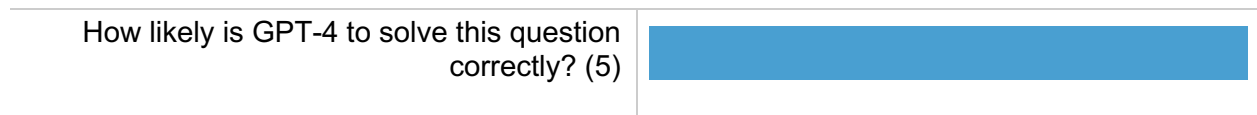
GenAICalPre0 Rate how likely you think it is that ChatGPT will give a correct answer to the following prompts.

PLEASE DO NOT USE ChatGPT, OTHER LLMs OR ANY OTHER SEARCH ENGINE (e.g., Google) TO ANSWER THESE QUESTIONS

GenAICalPre1 Develop an HTML page with JavaScript and canvas to draw a representation of the US flag that rotates 90 degrees clockwise each time it is clicked.

Extremely unlikely Somewhat unlikely Neither likely nor unlikely Somewhat likely Extremely likely I don't know

0 10 20 30 40 50 60 70 80 90 100



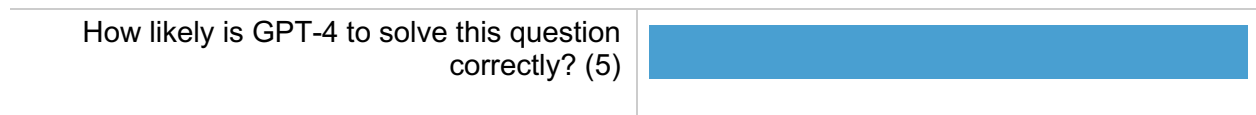
GenAICalPre2 Here is some data about Australian cities that I copied from Wikipedia. Based on this data, which cities had an odd-numbered population in 2011?

Australian Capital City Statistical Areas Population Table

City Statistical Area	Pop. June 2022	Pop. June 2011	Growth	Included SUAs
Greater Sydney	5,297,089	4,608,949	+14.93%	Sydney, Central Coast
Greater Melbourne	5,031,195	4,169,366	+20.67%	Melbourne, Bacchus Mars
Greater Brisbane	2,628,083	2,147,436	+22.38%	Brisbane
Greater Perth	2,224,475	1,833,567	+21.32%	Perth
Greater Adelaide	1,418,455	1,264,091	+12.21%	Adelaide
Australian Capital Territory	456,692	367,985	+24.11%	Canberra, Queanbeyan (A
Greater Hobart	252,693	216,273	+16.84%	Hobart
Greater Darwin	149,582	129,106	+15.86%	Darwin

Extremely unlikely Somewhat unlikely Neither likely nor unlikely Somewhat likely Extremely likely I don't know

0 10 20 30 40 50 60 70 80 90 100



GenAICalPre3 Imagine you have a large box filled with small identical cubes. The box is completely full, and the dimensions of the box are 10 cubes long, 5 cubes wide, and 2 cubes high. You decide to take out all the cubes and rearrange them to form a new box that is 5 cubes long, 4 cubes wide, and 4 cubes high. How many cubes do you have left over after filling the new box?

Extremely unlikely Somewhat unlikely Neither likely nor unlikely Somewhat likely Extremely likely I don't know

0 10 20 30 40 50 60 70 80 90 100

How likely is GPT-4 to solve this question correctly? (5)



GenAICalPre4 I'm playing wordle. My guesses so far are 1. CRANE (only the last E present, but in the wrong location) 2. POURS (only the first P present, but in the wrong location) 3. MIGHT (no letters present) 4. DEARY (only the E present, but in the wrong location) What do you think the word actually is?

ExtremelySomewhat Neither SomewhatExtremely I don't
unlikely unlikely likely likely likely know
nor
unlikely

0 10 20 30 40 50 60 70 80 90 100

How likely is GPT-4 to solve this question correctly? (5)



GenAICalPre5 Write a webpage that shows a drawing of a cake and plays "happy birthday" when the page loads. Both should be generated with javascript. Make sure the cake looks right and the melody and note duration are correct in the music.

ExtremelySomewhat Neither SomewhatExtremely I don't
unlikely unlikely likely likely likely know
nor
unlikely

0 10 20 30 40 50 60 70 80 90 100

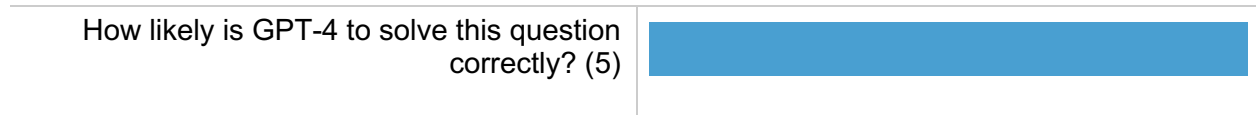
How likely is GPT-4 to solve this question correctly? (5)



GenAICalPre6 Write a single HTML file that has a javascript program that uses a canvas2d to draw "hello" with individual lines and curves. Do not use fillText.

ExtremelySomewhat Neither SomewhatExtremely I don't
unlikely unlikely likely likely likely know
nor
unlikely

0 10 20 30 40 50 60 70 80 90 100



GenAICalPre7 Capitalize each sentence beginning with ""Input:""

Input: darcy, she left Elizabeth to walk by herself.

Output: Darcy, she left Elizabeth to walk by herself.

Input: funny little Roo, said Kanga, as she got the bath-water ready.

Output: Funny little Roo, said Kanga, as she got the bath-water ready.

Input: hello this is a string.

Output: Hello this is a string.

Thank you for your help with this. From now on you will count the number of words in a sentence.

Input: This is an example sentence.

Output: 5

Input: Now another sentence.

Output: 3

Input: How long is this much longer sentence that has many words?

ExtremelySomewhat Neither SomewhatExtremely I don't
unlikely unlikely likely likely likely know
nor
unlikely

0 10 20 30 40 50 60 70 80 90 100

How likely is GPT-4 to solve this question correctly? (5)



Page Break



FatiguePre **How would you rate your current level of focus and energy for completing this survey?**

- Very high – I'm fully focused and ready (5)
- Somewhat high – I feel quite prepared and alert (4)
- Neutral – I'm neither tired nor particularly energized (3)
- Somewhat low – I'm a bit tired or distracted (2)
- Very low – I'm already feeling quite fatigued or unfocused (1)

End of Block: Pre-Survey

Start of Block: Training_GPT

TimerTrainingGPT Timing

First Click (1)

Last Click (2)

Page Submit (3)

Click Count (4)

TrainingGPTInt **Training**

Introduction to ChatGPT Enterprise

This training should take you 15-20 minutes and will auto-advance to the next section in 30 minutes.

This training program is designed to equip you with advanced skills in talking to ChatGPT. Through a series of interactive modules, you will learn how to effectively use ChatGPT to your advantage.

TrainerGPTVideo 1. Please start by watching the 2 videos below
Getting ChatGPT to do what you want

Signing Into ChatGPT

Page Break

TrainingGPT2 2. ChatGPT is a guide, not an individual contributor

Throughout this assignment and in general when working with ChatGPT, you may be tempted to have ChatGPT do all your work. The outputs look very convincing!

In our study last year, we saw that ChatGPT hurt performance by 23% for individuals who over-relied on it for problem solving. Therefore, we encourage you to do the assignments alongside ChatGPT – using ChatGPT as your guide.

Use your own rigor and intuition to quality check ChatGPT's output.

Page Break

TrainingGPT3 3. Introduction to talking to ChatGPT

This training will describe process of designing and refining instructions (i.e. prompts) given to a large language model (e.g. ChatGPT) to get better results and elicit desired behaviors. Note that combining these methods can sometimes have greater effect.

TrainingGPT3.1 **Standard Prompting**

Most users of ChatGPT use standard prompting (also known as “naïve” prompting or “zero-shot” prompting). This is when the model is given a task without prior examples; it must deduce what to do from the prompt and its existing training. For example, just asking ChatGPT “What are the best practices for talking to ChatGPT?”.

Standard prompting is often sufficient if you are following a few best practices:

Be clear and concise with common language. Avoid confusing consulting jargon!

Provide context such as the purpose of the ask and details behind the instructions

Be specific by clearly stating what you are trying to accomplish Clarify the output format – e.g. bullets, tables, paragraphs, etc

Worse

Better

Analyze this data a CSV file containing sales data from the last quarter, including columns for date, product ID, and sales volume. Can you provide a Python script using pandas to read this file and calculate the total sales volume for each product? Please include comments in the script explaining each step of the process.	I have What's the best machine learning model? Given a dataset with 1000 rows of customer demographic information (age, income, number of purchases) and a binary target variable indicating whether each customer subscribed to a service, which machine learning model would be most appropriate for predicting subscription likelihood? Please explain your recommendation based on the data characteristics.
---	---

Make a graph from this data series data showing the daily number of visitors to my website over the past year. Could you guide me on how to use Python to plot this data, including a moving average line to highlight trends? Additionally, could you explain how to customize the plot to add labels for the x-axis ("Date") and y-axis ("Number of Visitors"), and a title for the chart ("Daily Website Visitors")?	I have time-
--	--------------

TrainingGPT3.2

Ask the model to adopt a persona

Asking the model to adopt a persona can allow you to get more specific answers compared to

just asking the question. This can either be a certain individual (e.g. Elon Musk), or a specific qualification. Adapting the concept of adopting a persona for data science tasks can help in obtaining more specialized and nuanced responses. Let's try it!

TrainingGPTPrompt3.3

Type the following into ChatGPT: "What's the best way to analyze large datasets?" and copy the answer below:

TrainingGPTPrompt3.4 Now tell it to adopt a persona: "Acting as a data scientist, tell me the best way to analyze large datasets." Now copy the answer below and take a mental note of how the answer has changed:

TrainingGPTPrompt3.5 Finally, get more specific in your persona: "You'll act as a data scientist who specializes in big data analytics, with extensive experience in Python and Spark. Explain the most efficient method to process and analyze multi-terabyte datasets, including step-by-step instructions on setting up the environment, loading the data, and performing

exploratory data analysis.”

Copy the answer below and take a mental note of how the answer has changed again:

TrainingGPTExample Provide examples (i.e. one-shot or few shot prompting)

Providing general instructions that apply to all examples is generally more efficient than demonstrating all permutations of a task by example, but in some cases providing examples may be easier. For example, if you intend for the model to copy a particular style of responding to user queries which is difficult to describe explicitly. Incorporating the concept of one-shot or few-shot prompting into data science or data cleaning tasks can effectively guide the model to understand and replicate a specific answering or problem-solving style. This is known as "few-shot" prompting. Let's try it!

TrainingGPTExample1 Type the following into ChatGPT: “How do I extract key phrases from text?” and copy the answer below

TrainingGPTExample2 Now try giving it an example of how to respond: “Answer in a consistent style as this example. Question: How can I identify the sentiment of user reviews? Answer: Sentiment analysis of user reviews can be efficiently performed using Natural

Language Processing (NLP) techniques. The first step involves preprocessing the text by removing stop words and punctuation, followed by tokenization. Next, applying a pretrained model like VADER (Valence Aware Dictionary and Sentiment Reasoner) or a fine-tuned BERT model can classify the sentiment of each review into categories such as positive, negative, or neutral. This process enables an automated and scalable way to gauge customer sentiment from textual data. Now that I have given you an example – How do I extract key phrases from text?” Now copy the answer below and take a mental note of how the answer has changed:

TrainingGPTSteps

Specify the steps required to complete a task

Some tasks are best specified as a sequence of steps. Writing the steps out explicitly can make it easier for the model to follow them. This is known as “Chain-of-thought” prompting. Let’s try it!

TrainingGPTSteps1 Ask ChatGPT: “How do I classify images using a deep learning model?” and copy the answer below:

TrainingGPTSteps2 Now try a chain-of-thought approach and type the following into ChatGPT:

“Consider and include the following elements in your project to classify images using a deep learning model:

Data Collection: Identify and gather images for your dataset. Mention the source of your images and how many categories or classes of images you plan to classify.

Data Preprocessing: Describe the steps for resizing images to a uniform size, normalizing pixel values, and splitting the dataset into training, validation, and test sets.

Model Architecture Design: Choose a deep learning model architecture suitable for image classification. Consider whether to use a pre-trained model for transfer learning or to design a model from scratch.

Model Training: Outline the process for compiling the model with an appropriate optimizer and loss function. Mention how you will use data augmentation to improve model generalization.

Hyperparameter Tuning: Discuss the approach for tuning hyperparameters, such as learning rate, batch size, and the number of epochs, to improve model performance.

Model Evaluation: Explain how to evaluate the model's performance on the test set using metrics such as accuracy and precision. Consider plotting a confusion matrix to understand the model's classification behavior across different classes.

Model Deployment: Describe how you would deploy the model for real-world use, including converting the model to a suitable format for deployment and integrating it with an application for image classification.

Performance Monitoring and Updating: Consider how you will monitor the model's performance in production and the steps for retraining the model with new data or adjusting it based on performance feedback.

Let us think step by step. How do I tackle this image classification project with a deep learning model?” Now copy the answer below and take a mental note of how the answer has changed:

TrainingGPTTime

Give the model time to “think”

If asked to multiply 17 by 28, you might not know it instantly, but can still work it out with time. Similarly, models make more reasoning errors when trying to answer right away, rather than taking time to work out an answer. Asking for a "chain of thought" before an answer can help the model reason its way toward correct answers more reliably.

TrainingGPTTime1

Suppose for example we want a model to evaluate a student's solution to a math problem. The most obvious way to approach this is to simply ask the model if the student's solution is correct or not. However, this can sometimes lead to ChatGPT giving you the wrong answer. The following is an example prompt that is more likely to give an accurate answer: "First work out your own solution to the problem. Then compare your solution to the student's solution and evaluate if the student's solution is correct or not. Don't decide if the student's solution is correct until you have done the problem yourself."

TrainingGPTTime2 You can also ask the model to check it's own work or identify if it missed anything.

For example, suppose that we are using a model to list excerpts from a json file. If the file is very large, it is common for a model to stop too early and fail to list all relevant excerpts. In that case, you can ask the model to find any excerpts it missed on previous passes.

Are there more relevant excerpts? Take care not to repeat excerpts. Also ensure that excerpts contain all relevant context needed to interpret them - in other words don't extract small snippets that are missing important context.

Page Break

TrainingGPTDataA 4. Analyzing Data with ChatGPT's Data Analyst When working with data, especially if using code like Python, using ChatGPT's Data Analyst can significantly enhance your ability to analyze and interpret data directly within the chat. The Data Analyst allows for the execution of Python code, enabling data analysis, visualization, and more. **Keep in mind that ChatGPT's Data Analyst is still being refined and it sometimes makes mistakes! Its critical to do the work alongside ChatGPT to check your work!**

Here's an example of how to get started with Data Analyst and how to make sure it shows you the code it uses so you can put that code into your own notebooks for checking its work:

TrainingGPTDataA1 Best Practices for Using ChatGPT's Data Analyst: **Do the work alongside ChatGPT:** ChatGPT can make errors, even when using Data Analyst! If you are asking ChatGPT to do analysis for you, test what it is doing somewhere outside of ChatGPT. For example, if you ask it to write code for you, copy and paste the code it uses into a Jupyter notebook or other IDE. Run the code and test that it is doing what you expect!

Clear Definition of the Task: Before asking ChatGPT to use Data Analyst, clearly define what you aim to achieve with your data. Whether it's data cleaning, visualization, statistical analysis, or machine learning - having a clear objective will guide the code you write and the questions you ask the ChatGPT Data Analyst.

Break Down the Task: Divide your overall task into smaller, manageable steps. This could include data importation, preprocessing, exploratory data analysis (EDA), model building, and evaluation. Addressing each step individually can simplify complex analyses.

Provide Context: When prompting the ChatGPT Data Analyst, provide as much context as possible. This includes the structure of your dataset, the libraries you wish to use, and any specific methods or techniques you're interested in.

Specify the Output Format: Indicate how you'd like the results to be presented. For instance, if you're visualizing data, specify the type of plot you need. For statistical analyses, mention how you'd like the results to be summarized.

Ask ChatGPT about the errors you see when testing its work: Always test ChatGPT's work outside of ChatGPT! If you run into errors, ask ChatGPT to help you solve the problem.

Iterative Exploration: Data analysis is often exploratory and iterative. Don't hesitate to refine your questions/prompts based on the output you receive. If an analysis doesn't provide the insight you were hoping for, adjust your approach and try again.

Use ChatGPT to Help: Try asking ChatGPT for help refining your prompts to get the outcome you are looking for!

TrainingGPTDataA2 **Try some prompts that will get the data analysis process started:**

TrainingGPTDataA3 **Example: Time Series Analysis** **Initial Task:** Analyze seasonal patterns in time series data. **Prompt:** "I have a time series dataset stored in a pandas DataFrame df with two columns: 'Date' (datetime) and 'Daily_Sales'. I'd like to analyze seasonal patterns in daily sales over the year. Write Python code using pandas and statsmodels to decompose the time series into trend, seasonal, and residual components. Then, plot these components using matplotlib to visualize the seasonal patterns."

TrainingGPTDataA4 **Example 2: Natural Language Processing (NLP) for Sentiment Analysis** **Initial Task:** Perform sentiment analysis on customer reviews. **Prompt:** "Given a list of customer reviews stored in a pandas DataFrame reviews_df with a column 'Review_Text', use Python's Natural Language Toolkit (NLTK) or another NLP library to preprocess the text (tokenization, removing stopwords, and lemmatization). Then, apply a pre-trained sentiment analysis model from the library to classify each review as positive, negative, or neutral. Summarize the overall sentiment distribution among the reviews."

Page Break

TrainingGPTIssues 5. Common issues and their solutions

Context memory is overloaded – ChatGPT can only remember so much! If you find that your chat is getting stuck in a concept you don't want– make a new chat and restart with more specific instructions up front. If you do this during the study, make sure to provide us with links to every chat you create to answer the question. **A file is overloading the context** – If you load a file into ChatGPT it will sometimes read the file directly into the chat context memory and overload the chat creating weird results. When uploading large files, you tell ChatGPT not to read it into the chat's memory by saying something like: "Only open this file when running python code using Data Analyst and don't store it in the context memory of this chat".

Alternatively, you can tell ChatGPT to only read a certain portion of the file, e.g. "Only read the first 10 rows of this file." **ChatGPT's Data Analyst keeps having errors** – Ask ChatGPT not to run the code and instead just generate the code. Then run the code yourself – e.g, by using Jupyter.

TrainingGPTOtherRes 6. Other Resources:

Stack Overflow (<https://stackoverflow.com>) is one of the largest, most trusted online communities for developers to learn, share their programming knowledge, and build their careers. It features a vast repository of questions and answers on a wide array of programming and data science topics.

Feel free to watch this video on how to leverage Stack Overflow for coding and problem solving.

TrainingGPTOAI Finally, read through OpenAI's prompt engineering documentation if you want some additional examples of techniques and strategies. Or even if you just want to reference some of these strategies again during the study:

<https://platform.openai.com/docs/guides/prompt-engineering/strategy-use-external-tools>

Page Break

.4.2 Task details

TrainingGPTOffDoc Please download the following pdf to have access to all of these resources while working on the rest of the survey

[Training Document](#)

End of Block: Training_GPT

Start of Block: Begin Tasks - Optional Break

TimerOptionalBreak Timing

First Click (1)

Last Click (2)

Page Submit (3)

Click Count (4)

OptionalBreak1 **Please feel free to take an optional break if you desire. Please keep breaks to under 10 minutes, if possible, to not disturb the flow of your engagement. The next section (once you advance) will be a timed 90-minute section focusing on a data science task.**

Once you are ready to proceed to the next task, you may continue by using the arrow below.

End of Block: Begin Tasks - Optional Break

Start of Block: Coding Task - Instructions

CTInstruct1 **Coding Task**

This next Task is a coding assignment. You will be asked to use a Google Colab notebook to write Python code to process some data and arrive at a solution. Python is a common programming language used for data science tasks. We will walk you through what all this means.

You will be limited to 90 minutes to complete the assignment. But BEFORE the 90-minute timer starts, you will need to watch some Colab introduction videos and verify Colab is setup properly.

1. Colab Instructions (~15 minutes)

Setup Google & Colab

Setup Coding Task Notebook

Colab Videos to help you with setting up the Coding Task Notebook
2. Coding Task (90 minutes)

CTInstruct2 **1. Colab Instructions (~15 minutes)**

CTInstruct3 **Setup Google & Colab**

You should have received instructions to set up a Google Account with your BCG Email and Install Google Colab on your Google Account. The email was titled “[GenAI Experiment with OpenAI] Getting Started Details”. If you have not gone through the instructions in that email, please do that now.

In case you can't find the email, the instructions are in this file for your convenience: [Setup Google & Colab](#)

If you are having issues, please check this Troubleshooting Guide, with solutions to common issues that come up during this process: [Troubleshooting](#)

CTInstruct4 **Setup Coding Task Notebook:**

For the Coding Task, you will be working in a shared Google Folder located at the following link. You will need to copy and paste this URL into a new tab: `{e://Field/google_drive_link}`

The instructions for opening the Colab document and getting started can be found in the "Colab_Setup_Instructions.docx" file located within this shared folder.

Once you have the Colab document open, you need to follow the instructions inside the notebook itself, to update and run the "Setup Logging Before The Task" Cell.

If these instructions are confusing, don't worry! The videos below will also walk you through these instructions step-by-step.

CTInstruct5 Please Confirm that you have Colab installed on your Google Account, and are able to open the shared notebook file in Colab ([\\${e://Field/google_drive_link}](#)):

Yes, I have installed Colab on Google Account and am able to open the shared notebook. (1)

CTInstruct6 **Colab Videos:** Watch the following videos to familiarize yourself with Colab.

CTInstruct7 Confirm you have watched the Colab videos above.

Yes (1)

CTInstruct8 Once you have successfully run the "Setup Logging Before The Task" cell, you are ready to advance to the Task instructions on the next page. You will know that it is successful if you see the words "Successfully mounted your Drive! Continue below to the task" below the cell.

Here is what it will look like if you are successful:

Once you advance to the next page, your 90 minute task timer will start.

CTInstruct9 **In Case of Errors** If you run into errors during the setup, you will not be able to complete the Coding Assignment. In this case, please refer to the [Troubleshooting](#) and see if any of the solutions apply to your case.

Here is a common error message. If you see this - DO NOT proceed!

If you get an "Access Blocked : Authorization Error" while running the setup cell, please switch to a personal Google account to complete this task.

If you have tried everything and you still see an error, please reach out to Ryan Kennedy (kennedy.ryan@bcgfd.com) on Slack. Please do not reach out with any questions regarding how to complete the task itself, but rather only questions regarding setup errors.

DO NOT Proceed to the Task if you have not completed this step or if you are seeing any errors

CTInstruct10 Confirm that you have gone to the shared Google Folder, were able to Mount your Google Drive using the "Setup Logging Before The Task" cell, and see "Successfully mounted your Drive! Continue below to the task" in the space below the "Setup Logging Before The Task" cell.

Yes (1)

CTInstruct11 You are now ready to move on to the Coding Task. The official 90-minute task timer will start once you advance.

End of Block: Coding Task - Instructions

Start of Block: Coding Task

CodingInstructionGPT **Please read through the instructions to complete the next task!**

You are **not expected to have any prior coding experience**. Please try your best to complete the assignment **with code in Colab**, and get as far as you can.

You may spend **up to 90 minutes** on this task. At the 90-minute mark, the form will **automatically submit and move you to the next portion of the study, regardless of your progress**. You may choose to advance prior to the 90 minutes if you have fully completed the task.

Use ChatGPT Enterprise to perform this task **in whatever way you like (uploading files, copying and pasting the questions or errors and results)**. Access ChatGPT Enterprise by going to <https://chat.openai.com/> and log in using your BCG login.

We highly encourage you to do your best to complete this task. It might be challenging sometimes! But you're going to get CDC credit just for putting in an honest effort and if you are

in the **top 50th percentile of your group** we'll notify your CDA about how well you did! We truly appreciate your effort.

Page Break

CodingTaskTimer Timing

First Click (1)

Last Click (2)

Page Submit (3)

Click Count (4)



CodingTaskOnLoadTime CodingTaskOnLoadTimeTracker

CodingTaskIntro **The clock has started! You now have 90 minutes to complete the coding task. In case you need it again, here's the link to the Google Colab**

Documents: [\\${e://Field/google_drive_link}](#)

Use ChatGPT enterprise to perform this task **in whatever way you like (uploading files, copying and pasting the questions or errors and results)**. Access ChatGPT Enterprise by going to <https://chat.openai.com/> and log in using your BCG login.

No matter how you use ChatGPT or other resources to complete this task, **make sure any code you use is run in your Google Colab notebook so that we can review your solution.** If you used ChatGPT to help generate and execute your code, please copy the code back to your Colab notebook, run it, and troubleshoot for errors. We will not be able to review your solution if the code is not run in Colab so you will receive no credit if your notebook file is empty. Return to this survey when you are finished with the task in Colab.

Here are the Intro Colab Video Links provided again for reference:

[Intro to Colab](#)

[Using Colab Features](#)

CodingTaskInstruct Below you will find the details of the question. You will be provided with the overview of the data sets and an overview of the data cleaning steps you will need to take. The steps you need to take are also noted in your Colab Notebook.

CodingTask Assignment

Use the datasets found in your Google Folder to answer the question: Which 5 customer IDs had the highest average order by total price in May 2022?

Overview of Datasets

Dataset 1: Orders Data (orders.csv)

File Type: CSV

Contents:

customer_id: The unique identifier for each customer.

order_info: Information about each order in the format order number ;

date and time. The order number in this dataset (once decoupled from the date and time) corresponds with that in the next.

Dataset 2: Products Data (products.csv)

File Type: CSV

Contents:

customer_id: The unique identifier for each customer associated with an order.

order_id: ID of each order in the format order number. The order number in this dataset corresponds with that in the previous once decoupled from the date and time.

order_products: Details about the products in each order in the following format: {product_id: [product_price, product_quantity], ...}.

Each product is sold either at its original price or a 20% discount.

Commonalities

Both datasets share the customer_id field and order_id information with the order number.

Each combination of order ID and customer ID is unique. This is because each order ID is unique, whereas customer IDs may be repeated across multiple orders.

Note that the order and customer IDs across the two files are consistent.

Whenever you have information about either one of the IDs, it is correct.

Overview of Data Cleaning Steps

Data Quality and Cleaning Guidelines

Order and Customer IDs: Entries are always correct when not NULL, and NULL values should be tried to be filled in wherever possible using data from elsewhere.

Date and Time Fields: Entries with incorrect values should be removed.

Product Quantities and Product IDs: Always correct unless marked as NULL, which indicates missing values.

Product Prices: Each product ID is associated with a unique price. For some orders, the original unique price for a given product ID is discounted at 20% so that the discounted price is what is shown for those orders. However, for orders where the price is not discounted, sometimes there are junk or NULL values instead of the correct original price. Junk or NULL values in the product

prices should be replaced with the original price (the discounted price should be left as is wherever it is shown but not added in elsewhere).

Tips for Handling Junk and NULL Values; duplicates

Examine common values in each column to identify patterns and potential corrections.

Attempt to fix junk or NULL values using information elsewhere in the data before considering row deletion.

Date time fields can have incorrect fields that are not correctable, discard the affected rows and values to maintain data integrity.

Check for duplicates at every stage

ConfirmColabCode **Please confirm that you used Colab to complete this assignment and that all the code is in the Colab Notebook.**

Yes, all my code for this assignment is in the Colab Notebook. (1)

CodingTaskAnswers **Enter your answer from the task in a comma separated list here (Ex : "193738, 129490, 102948, 109812, 892738") or leave blank if you did not finish**

Page Break

Display This Question:

If If Enter your answer from the task in a comma separated list here (Ex : "193738, 129490, 102948, 109812, 892738") or leave blank if you did not finish Text Response Is Empty

CodingTaskPostAnswer If you were not able to enter your answer before the timer, please enter your answer from the task in a comma separated list here (Ex : "193738, 129490, 102948, 109812, 892738") or leave blank if you did not finish

Page Break



CodingGPTConvConf **Confirm you used ChatGPT to help complete your task**

- Yes I used ChatGPT (1)
- No I did not use ChatGPT (0)

Display This Question:

If Confirm you used ChatGPT to help complete your task = No I did not use ChatGPT

CodingGPTConvWhy Explain why you did not use ChatGPT when you were instructed to. Make sure to use it for any remaining task where you are instructed to.

CodingGPTConvLink **"Share" your ChatGPT conversations with us, so we can better understand how ChatGPT assisted you with the tasks. Only include conversations that you used for the task you just completed. If you used multiple ChatGPT conversations for the task, please share a link to each one.**

NOTE: Please do not delete your conversations for a week or so, so we can make sure to collect the data we need to from them. If you delete the conversation on your side, we will no longer be able to view your shared links.

The screenshot below shows how you can Share your conversation. Select the 3 dots on the Individual Conversation tab, and select "Share", then copy the link and paste it below.



CodingGoogleConf **Did you use Google to help complete your task?**

- Yes (1)
 - No (0)
-

CodingOtherTools **Please explain any other tools you used to complete your tasks. Include the name of the tool used, and how you used it to assist you.**

Be as specific as you can.

End of Block: Coding Task

Start of Block: Break1

TimerTaskBreak Timing

First Click (1)

Last Click (2)

Page Submit (3)

Click Count (4)

Display This Question:

If Task_Counter = 1

TimerBreakText **Please take at least a 10 minute break. Note that the survey will auto-advance to the next section in 30 minutes. When you are ready to start the next task,**

click the arrow below to continue.

The next Task should be completed within 90 minutes

End of Block: Break1

Start of Block: Statistics Task

StatsInstructGPT **Please read through the instructions to complete the next task!**

You may spend **up to 90 minutes** on this task. At the 90-minute mark, the form will **automatically submit and move you to the next portion of the study, regardless of your progress**. You may choose to advance prior to the 90 minutes if you have fully completed the task.

Use ChatGPT Enterprise to perform this task **in whatever way you like (uploading files, copying and pasting the questions or errors and results)**. Access ChatGPT Enterprise by going to <https://chat.openai.com/> and log in using your BCG login.

We highly encourage you to do your best to complete this task. It might be challenging sometimes! But you're going to get CDC credit just for putting in an honest effort and if you are in the **top 50th percentile of your group** we'll notify your CDA about how well you did! We truly appreciate your effort.

Page Break

TimerStats Timing

First Click (1)

Last Click (2)

Page Submit (3)

Click Count (4)



StatsTaskOnLoadTime StatsTaskOnLoadTimeTracker

StatsInstructions **Instructions:**

For this task **USE ChatGPT** enterprise version by accessing <https://chat.openai.com/> using your BCG login to perform the task below and answer all the questions. **However, do not send any images to GPT and refrain from copying and pasting the exact question.**

StatsQ1 **Question 1**

The following is the first five rows of data containing financial and demographic information about domestic partners who have co-purchased a home in the last several years.

Please note that the following table is illustrative and represents a snapshot sample of the data to solve this problem. All the information you need to solve the problem is contained within this snapshot.

Age 1	Age 2	Income 1	Income 2	Borough	ZIP Code	Date	Price	Mortgage
39	37	270000	180000	Manhattan	10076	1 January 2016	1,125,000	Yes
NULL	38	445000	670000	Manhattan	10025	1 January 2016	2,249,000	Yes
27	29	145000	225000	Queens	11106	2 January 2016	900,000	Yes
33	NULL	90000	76000	Brooklyn	11203	2 January 2016	415,000	Yes
68	55	78000	450000	Bronx	10474	2 January 2016	3,399,000	No

StatsQ1.1 You have been tasked with predicting based on demographics and price whether a mortgage was taken out to by the house. You prompt ChatGPT for detailed instructions on how to do this, and ChatGPT recommend using a logistic regression model. It recommends the following steps (the text in blue is the ChatGPT output we are referring to).



StatsQ1.1.A 1. **Understand Your Dataset Explore and Preprocess:** Start by exploring your dataset to understand the features available and their types (numerical, categorical). Clean the data by handling outliers and possibly irrelevant features. Preprocessing steps like encoding techniques (e.g., one-hot encoding) might be necessary for categorical data. Ensure that your dataset does not have missing values. You can either fill them in with a strategy (mean, median, mode) or remove the rows/columns with missing values, depending on the situation.

**Which of the following are among the steps you could take to address this point?
Select all that apply.**

- Plot the distribution of each of the numerical variables and remove rows with outliers from this dataset (1)
- One-hot encode the "Borough" variable (2)
- Investigate relationships between variables (3)
- Convert date to a numerical variable (4)
- One-hot encode the ZIP code variable (5)
- One-hot encode the age variables (6)
- Bin the ZIP codes by neighborhoods and do not process further (7)
- Bin the ZIP codes by neighborhoods and one-hot encode (8)
- Check columns with null values and remove those with >80% missing values (9)
- Impute NULL values by using a summary statistics or by developing a simple model that predicts those values based (10)

StatsQ1.1.B 2. Split the Data **Train-Test Split:** Divide your dataset into a training set and a testing set (commonly a 70-30 or 80-20 split) to evaluate the model's performance on unseen data. **3. Train the Model** **Training:** Use the training dataset to train your model, adjusting parameters as needed. For complex models, consider using cross-validation to fine-tune hyperparameters and prevent overfitting. **4. Evaluate the Model** **Performance Metrics:** Evaluate your model on the test set using appropriate metrics such as accuracy, precision, recall, F1 score, and the ROC-AUC curve. These metrics will help you understand how well your model is performing in terms of both its ability to predict mortgages correctly and its robustness against false positives or negatives. **What issue necessitates**

using all these metrics? Which of the above steps is affected by this issue and how?
(Answer in 100 words or less – bullet points ok)

StatsQ1.1.C **Would you change the order of any of the above steps (i.e., steps 1-4)? Why or why not?**

(Answer in 100 words or less – bullet points ok)



StatsQ1.2 **You want to try a k-Nearest Neighbors model. Which of the following are not required (although recommended) for logistic regression, but absolutely necessary for k-Nearest Neighbors? Select all that apply.**

- Transform numerical variables (e.g. log) (4)
- Make sure there are only two classes to predict (5)
- Convert Mortgage column from string to binary (6)
- Standardize numerical variables (7)
- Impute the missing age with the other age in the same row (8)
- One-hot encode the appropriate variables (9)



StatsQ1.3 **You also try a decision tree model for the same classification problem, to compare performance. You realize your model is performing quite poorly on both**

training and validation sets. You double-check the code and there are no bugs. What could be causing this problem? Select all that apply

- Your model is underfit (4)
- Your model is overfit (5)
- The learning rate hyperparameter is too small (6)
- The learning rate hyperparameter is too large (7)
- The decision tree is too shallow (8)
- The decision tree is too deep (9)
- None of the above (10)



StatsQ1.4 Next, you have been instructed to predict the price based on the other variables, and this time you have been instructed to use linear regression. Following instructions from ChatGPT, you perform a basic linear regression. You notice that your R2 value is too low. You prompt ChatGPT for suggestions on how to diagnose the

problem, and it is recommended that you check the residual plots. You notice that the residual plot does not appear random. What could this mean? Select all that apply.

- The observed values of your dependent variable are independent from each other (4)
 - Your model is missing an important variable (5)
 - There is some interaction between your variables (6)
 - A higher order term might be required in your regression (7)
 - Variance of the residual is the same for any value of X (8)
-

StatsQ1.5 For the following residual plots, what could be the characteristics of or issues with the data or model that are corresponding with these results (choose from the list provided for each image)? It is possible that more than one characteristic or issue applies to any given image, and it is possible that a characteristic or issue may apply to more than one image.



StatsQ1.5.A **Plot A**

Characteristic or issues choices:

- No characteristic or issue is apparent (1)
 - Heteroscedastic data (2)
 - Outliers (3)
 - Response variable requires transformation (4)
 - A higher order variable might be required (5)
-

StatsQ1.5.B **Plot B**

Characteristic or issues choices:

- No characteristic or issue is apparent (1)
 - Heteroscedastic data (2)
 - Outliers (3)
 - Response variable requires transformation (4)
 - A higher order variable might be required (5)
-

StatsQ1.5.C **Plot C**

Characteristic or issues choices:

- No characteristic or issue is apparent (1)
 - Heteroscedastic data (2)
 - Outliers (3)
 - Response variable requires transformation (4)
 - A higher order variable might be required (5)
-

StatsQ1.5.D **Plot D**

Characteristic or issues choices:

- No characteristic or issue is apparent (1)
 - Heteroscedastic data (2)
 - Outliers (3)
 - Response variable requires transformation (4)
 - A higher order variable might be required (5)
-

StatsQ1.5.E **Plot E**

Characteristic or issues choices:

- No characteristic or issue is apparent (1)
 - Heteroscedastic data (2)
 - Outliers (3)
 - Response variable requires transformation (4)
 - A higher order variable might be required (5)
-

StatsQ1.5.F **Plot F**

Characteristic or issues choices:

- No characteristic or issue is apparent (1)
 - Heteroscedastic data (2)
 - Outliers (3)
 - Response variable requires transformation (4)
 - A higher order variable might be required (5)
-

StatsQ1.5.G **Plot G**

Characteristic or issues choices:

- No characteristic or issue is apparent (1)
 - Heteroscedastic data (2)
 - Outliers (3)
 - Response variable requires transformation (4)
 - A higher order variable might be required (5)
-

StatsQ1.6 **You are asked to train a new model to predict price on the newest version of the dataset. In this version, there are several more fields collected with demographic information and financial information of the couples. However, this data is only from the last month. Which of the following steps recommended by ChatGPT could be beneficial to take to address some of the issues that are likely to arise because of this? Select all that apply.**

- Perform PCA (4)
 - Use a neural network instead of linear regression (5)
 - Use a regularized model instead of linear regression (6)
 - None of the above (7)
-

StatsQ2.A **Question 2**

You are asked to prepare a simple linear model to classify the following points into class 1 (black dots) and class 2 (white dots). What is the best empirical risk of this model that you can achieve with 0-1 loss?

Justify your answer and show your working steps.

StatsQ2.B ChatGPT has run 3 classifiers on your data and provided a visual output, but not specified which models yielded which output. For each of the three images, name a classifier that could create the boundary represented by the solid black line, and one that could not (class 1 is the orange dots, and class 2 is the blue dots). You can ignore the dashed line, you can use the metrics on the bottom-left but you do not need them.

Justify your answer.

StatsQ3 Question 3:

Imagine you're a logistics manager and one of your delivery trucks has gone missing. You believe it lost its signal while on either Route A or Route B, with a 65% and 35% chance of being on each route respectively. Based on the coverage area of these routes, if the truck is on Route A and you search for a day, there's a 45% chance you'll find it. However, if it's on Route B and you search for a day, the probability of locating it is 75%.

StatsQ3.A If you only had one day to search for the truck, on which route would you focus your search efforts in order to maximize your chances of finding it?

Explain your choice and break down your calculations.

StatsQ3.B Assume that you made the rational decision on the first day, but didn't manage to locate the truck. The truck remains at the position that it was originally lost at and has not been moved. You have another day committed for search - has your initial idea of which route the truck is on changed? Where should you search now?

Explain your choice and break down your calculations.

Page Break



StatsGPTConvConf **Confirm you used ChatGPT to help complete your task**

- Yes I used ChatGPT (1)
- No I did not use ChatGPT (0)

Display This Question:

If Confirm you used ChatGPT to help complete your task = No I did not use ChatGPT

StatsGPTConvWhy Explain why you did not use ChatGPT when you were instructed to. Make sure to use it for any remaining task where you are instructed to.

StatsGPTConvLink **"Share" your ChatGPT conversations with us, so we can better understand how ChatGPT assisted you with the tasks. Only include conversations that you used for the task you just completed. If you used multiple ChatGPT conversations for the task, please share a link to each one.**

NOTE: Please do not delete your conversations for a week or so, so we can make sure to collect the data we need to from them. If you delete the conversation on your side, we will no longer be able to view your shared links.

The screenshot below shows how you can Share your conversation. Select the 3 dots on the Individual Conversation tab, and select "Share", then copy the link and paste it below.



StatsGoogleConf **Did you use Google to help complete your task?**

Yes (1)

No (0)

StatsOtherTools **Please explain any other tools you used to complete your tasks. Include the name of the tool used, and how you used it to assist you.**

Be as specific as you can.

End of Block: Statistics Task

Start of Block: Problem Solving Task

PSInstructionsGPT **Please read through the instructions to complete the next task!**

You may spend **up to 90 minutes** on this task. At the 90-minute mark, the form will **automatically submit and move you to the next portion of the study, regardless of your progress**. You may choose to advance prior to the 90 minutes if you have fully completed the task.

Use ChatGPT Enterprise to perform this task **in whatever way you like (uploading files, copying and pasting the questions or errors and results)**. Access ChatGPT Enterprise by going to <https://chat.openai.com/> and log in using your BCG login.

We highly encourage you to do your best to complete this task. It might be challenging

sometimes! But you're going to get CDC credit just for putting in an honest effort and if you are in the **top 50th percentile of your group** we'll notify your CDA about how well you did! We truly appreciate your effort.



Page Break



TimerPS Timing
First Click (1)
Last Click (2)
Page Submit (3)
Click Count (4)

JS

ProbsTaskOnLoadTime ProbsTaskOnLoadTimeTracker

PSInstructions **Instructions:**

For this task, **USE ChatGPT** enterprise version by accessing <https://chat.openai.com/> using your BCG login. Feel free to work with ChatGPT in whatever way you like (uploading files, copying and pasting the question or results).

PSQuestion **Problem Solving Task**

QUESTION

Imagine you are staffed on a case where you must implement a data-driven strategy for sports investing. You are given a dataset containing records of 45,360 international football matches, spanning from the inaugural official match in 1872 through to the year 2024. The competitions range from the FIFA World Cup, FIFI Wild Cup, to ordinary friendly games. All matches are men's senior internationals, excluding Olympic Games, matches involving B-teams, U-23, or league select teams.

Your task (make sure to describe and document your approach and your findings):

Develop and implement a method for quantifying how predictable each match result was. *You can solve this problem however you like, using any analytics platforms at your disposal (e.g. Excel, Alteryx, Python).* Explain in detail each step you took for your approach and justify. What was the most surprising match result in this dataset, based on your method? Return a .csv or Excel file containing four columns – the match date, the home team, the away team, and **your numerically determined match result predictability using the above method** for each match in the dataset

Keep in mind you have 90 minutes to complete this task. Time box and make sure you return a final answer.

DATASET INFORMATION

The `results.csv` file encompasses columns for:

- `date` - the match date
- `home_team` - the home team's name
- `away_team` - the away team's name
- `home_score` - home team's score at the end of the match, including extra time but excluding penalties
- `away_score` - away team's score at the end of the match, including extra time but excluding penalties
- `tournament` - tournament name
- `city` - the city or locality of the match
- `country` - the country hosting the match
- `neutral` - a TRUE/FALSE indicator of whether the venue was neutral

Assume that the result as shown in this dataset (win or tie) is the entire result – there are some cases of penalty shootouts and goals scored from penalties, but for complexity, we will ignore those for this exercise.

For clarity, current names are used for both teams and countries in historical matches. For example, an 1882 match featuring the team then known as Ireland against England is listed under Northern Ireland, reflecting the modern successor of the 1882 team. Country names are recorded as they were at the time of the match, but discrepancies between team and country names (e.g., Ghana vs. Gold Coast) are accounted for, with the `neutral` column marking such matches as non-neutral to clarify they were played at home.

Data: [results.csv](#)



PSUpload **Upload your csv or Excel file here:**

Page Break



ProbsGPTConvConf **Confirm you used ChatGPT to help complete your task**

- Yes I used ChatGPT (1)
- No I did not use ChatGPT (0)

Display This Question:

If Confirm you used ChatGPT to help complete your task = No I did not use ChatGPT

ProbsGPTConvWhy Explain why you did not use ChatGPT when you were instructed to. Make sure to use it for any remaining task where you are instructed to.

ProbsGPTConvLink **"Share" your ChatGPT conversations with us, so we can better understand how ChatGPT assisted you with the tasks. Only include conversations that you used for the task you just completed. If you used multiple ChatGPT conversations for the task, please share a link to each one.**

NOTE: Please do not delete your conversations for a week or so, so we can make sure to collect the data we need to from them. If you delete the conversation on your side, we will no longer be able to view your shared links.

The screenshot below shows how you can Share your conversation. Select the 3 dots on the Individual Conversation tab, and select "Share", then copy the link and paste it below.

.4.3 Post-Experiment Survey

(a) Post-survey

Now that you've completed the tasks, how would you rate your current level of focus and energy for completing this survey?

- Very high – I'm fully focused and ready
- Somewhat high – I feel quite prepared and alert
- Neutral – I'm neither tired nor particularly energized
- Somewhat low – I'm a bit tired or distracted
- Very low – I'm already feeling quite fatigued or unfocused

Please answer the below questions to the best of your knowledge

PLEASE DO NOT USE CHATGPT, OTHER LLM OR ANY OTHER SEARCH ENGINE (e.g., Google) TO ANSWER THESE QUESTIONS

1. Suppose we have a 'test_group' column in our dataframe (df) which has the values 'treatment' and 'control'. Which of the following code snippets will give us a dataframe filtered only to have the rows which correspond to 'treatment'? Select all that apply.
 - `df = df['treatment']`
 - `df = df[df['treatment']]`
 - `condition = df['test_group'] = 'treatment'`
 - `df = df[condition]`
 - `df = df['test_group'] = 'treatment'`
 - `df = df['test_group'] == 'treatment'`
 - `condition = df['test_group'] == 'treatment'`
 - `df = df[condition]`
 - `condition = df['test_group'] != 'control'`
 - `df = df[condition]`
2. If a coin is tossed 3 times, what is the probability of getting heads every time?
 - 1 out of 2
 - 1 out of 4
 - 1 out of 6
 - 1 out of 8
3. Distance-based algorithms are not affected by scaling
 - True
 - False
4. Which of these techniques can be used to handle missing data in categorical features? Select all that apply.
 - Removing rows having missing data
 - Replacing missing values with the most frequent category
 - Replacing missing values with the mean
 - Replacing missing values using predictive algorithms like classifiers
 - Replacing missing values using predictive algorithms like regressors

5. You are given a dataset of logos of famous companies , and you have to predict whether the review contains alphabets or not. Under which category does this problem fall? Select all that apply.
- Classification
 - Regression
 - Clustering
 - Natural language processing

1. This set of questions tests your ability to predict ("forecast") how well GPT-4 will perform at various types of questions. (In case you've been living under a rock these last few months, GPT-4 is a state-of-the-art "AI" language model that can solve all kinds of tasks.)

How likely is GPT-4 to solve this question correctly?

0 10 20 30 40 50 60 70 80 90 100

Develop an HTML page with JavaScript and canvas to draw a representation of the US flag that rotates 90 degrees clockwise each time it is clicked. ()



How likely is GPT-4 to solve this question correctly?

0 10 20 30 40 50 60 70 80 90 100

Here is some data about cities in Japan that I copied from Wikipedia. Based on this data, which cities have an even-numbered population?



City (Special Ward)	Prefecture	Population	Area (km ²)	Density (per km ²)	Founded
Special wards of Tokyo	Tokyo	9,375,104	621.81	13,890	
Yokohama	Kanagawa	3,732,616	437.38	8,500	1889-04-01
Osaka	Osaka	2,691,185	222.30	11,900	1889-04-01
Nagoya	Aichi	2,327,557	326.45	6,860	1889-10-01
Sapporo	Hokkaido	1,976,257	1,121.26	1,763	1922-08-01
Fukuoka	Fukuoka	1,588,924	340.96	4,515	1889-04-01
Kawasaki	Kanagawa	1,531,646	142.70	9,626	1924-07-01
Kobe	Hyōgo	1,524,601	552.23	2,772	1889-04-01
Kyoto	Kyoto	1,464,890	827.90	1,800	1889-04-01
Saitama	Saitama	1,324,854	217.49	5,483	2001-05-01
Hiroshima	Hiroshima	1,199,391	905.13	1,286	1889-04-01

How likely is GPT-4 to solve this question correctly?


0 10 20 30 40 50 60 70 80 90 100

I'm at a restaurant with a \$10 bill and want to use it exactly on some of the following items. Which ones should I buy: steak \$5.23 fries \$1.24 shake \$2.48 salad \$4.87 salmon \$4.13 cake \$1.00 ()




How likely is GPT-4 to solve this question correctly?

0 10 20 30 40 50 60 70 80 90 100

<p>Can you help me answer the following crossword clues. 1. "Lamented, in a way" (4 letters) 2. "Princess's irritant in a classic fairy tale" (3 letters) 3. "Bobbie Gentry's "___ to Billie Joe"" (3 letters) 4. "Leave no way out" (4 letters) 5. "Expression of false modesty from a texter" (4 letters) ()</p>	
--	--


How likely is GPT-4 to solve this question correctly?

0 10 20 30 40 50 60 70 80 90 100

<p>Who lost the Super Bowl two years after Pan-Am filed for bankruptcy? ()</p>	
--	--


How likely is GPT-4 to solve this question correctly?

0 10 20 30 40 50 60 70 80 90 100

<p>Write out the word "hello" as an ascii art drawing with # and _ ()</p>	
---	---

How likely is GPT-4 to solve this question correctly?

0 10 20 30 40 50 60 70 80 90 100

<p>What is the best next move for O in the following game of Tic Tac Toe?</p> <pre> - . O ----- . O X ----- X . X </pre>	
---	--

2. How did you find the use of Generative AI? Was it easy or difficult? Did it give you the answers you were looking for?

- The use of ChatGPT was easy and provided me with all the answers I was looking for
- The use of ChatGPT was easy and provided me with most the answers I was looking for
- The use of ChatGPT was easy, but did not provide me with most the answers I was looking for
- The use of ChatGPT was difficult, but provided me with all the answers I was looking for
- The use of ChatGPT was difficult, but provided me with most the answers I was looking for
- The use of ChatGPT was difficult and did not provide me with most the answers I was looking for

Next, please indicate the extent to which you agree or disagree with the following statements :

	Strongly Agree	Somewhat Agree	Neutral	Somewhat Disagree	Strongly Disagree
Generative AI helps me feel valuable in my role	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Generative AI elevates how important I feel my job is for society	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Generative AI elevates my professional status and level of influence within my organization	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Generative AI helps me feel more competent in my role	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Generative AI enables my ability to execute tasks and reach desired outcomes	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Generative AI enables my ability to execute data analytics tasks and reach desired outcomes	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Generative AI increases the value I place on my expertise and skill cultivation	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Generative AI increases my level of autonomy in making individual decisions in my role	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Generative AI helps me be more confident that I will meet my project managers expectations

Using Generative AI helps me stay aligned with my project managers expectations

Generative AI enables me to do what I really want to do in my role

I believe using Generative AI will contribute to the betterment of others in my work

I see Generative AI as my coworker

Generative AI will change the dynamic in my team

Generative AI improved how I perceive my role in the organization

I would recommend Generative AI to other consultants

I am proud of BCG's approach to Generative AI adoption within the firm

I believe BCG is at the leading edge of the Generative AI revolution

My managers and supervisors will expect more output from me because of Gen AI

Sustained use of ChatGPT for data science would have the potential to make me a better consultant in the 'Problem solving and insights' dimension

Sustained use of ChatGPT for data science would have the potential to make me a better consultant in the 'Communication and Presence' dimension

Sustained use of ChatGPT for data science would have the potential to make me a better consultant in the 'Practicality and Effectiveness' dimension

Rate how helpful you think Generative AI tools are for these use cases (Rating 1-7; with ability to say "I don't know")

Experience with GenAI



	1	2	3	4	5	6	7	I don't know
Brainstorm ideas	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Writing code for data analytics	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Writing code for data visualizations	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Learning how to use excel for data analysis and visualizations	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Identifying which machine learning models to use for a project	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Understanding the statistical significance of a result	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Writing code for data cleaning and preparation	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

3. GenAI benefits - In a few words, what do you think will be the biggest benefits of Generative AI for you?

4. GenAI risks - In a few words, what do you think will be the biggest risks of Generative AI for you?





5. On a scale of 1-7, where 1 = "Not at all" and 7 = "Extremely", please rate the following ...

	Not at all		Neither		Extremely		
0	1	2	3	4	5	6	7

How confident are you in your ability to contribute to data science projects?	
To what extent do you believe understanding data science concepts is important in the role of a BCG A/C?	

6. Given the capabilities of Generative AI, do you see the role of associates and consultants evolving in the next 5 years? If so, how?

7. Finally, answer the following questions on a scale of 0 to 10, where 0 is "Do not enjoy at all" and 10 is "enjoy to a great extent"

	Do not enjoy at all	Neutral	Enjoy to a great extent								
	0	1	2	3	4	5	6	7	8	9	10
How much do you think your coworkers enjoy their work?											
How much do you think your coworkers enjoy using ChatGPT for their work?											
How much would you enjoy doing more data analysis at work with the help fo ChatGPT?											
How much would you enjoy being tasked with data science tasks with the help of ChatGPT?											

Generally speaking, would you say that most people can be trusted, or that you need to be very careful in dealing with people?








- Most people can be trusted
- You need to be very careful in dealing with people
- Don't know

Please indicate the extent to which you agree or disagree with the following statements :

	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
ChatGPT can be trusted to give you correct information when researching a new topic	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
ChatGPT can be trusted to do quantitative analysis for you	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
ChatGPT can be trusted to clean data for you with minimal guidance	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
ChatGPT can be trusted to help you learn to do new things (e.g. use a new type of software)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

On a scale of 1-7, where 1 = "Not at all" and 7 = "Extremely", please rate the following ...

	Not at all		Neither			Extremely
	1	2	3	4	5	6 7

How confident are you in your ability to do data cleaning in python using ChatGPT as your guide?	
How confident are you in your ability to do quantitative analysis using ChatGPT as your guide?	
How confident are you in your ability to learn new skills with ChatGPT as your guide?	
How confident are you in identifying factual inaccuracies in ChatGPT's responses?	
How confident are you in judging the relevance of ChatGPT's responses to your questions?	
How confident are you in assessing the clarity and understandability of ChatGPT's output?	
How confident are you in evaluating the completeness of ChatGPT's answers to your queries?	

8. Investment Game

Next, you'll have an **exciting opportunity** to play a game with ChatGPT as a second player (**yes, ChatGPT can play games and make decisions!**) and **earn points that you will be able to redeem for exciting rewards!**

Note that **you will not see ChatGPT for this question**. Instead, we will email you with ChatGPT's response and with the reward options! The more points you have at the end of the game – the better your options will be!

Here's how the game works:

You (the Investor) and ChatGPT (the Responder) **will be given 100 tokens each**
As the Investor, you will have the opportunity to **pass some, all or none of your tokens to ChatGPT (the Responder), as you like**

Whatever amount you decide to give, **we will triple it and pass it to ChatGPT**

For example, if you decide to pass 0 tokens, we will give ChatGPT $3 \times 0 = 0$ and so it will have 100 tokens;

if you decide to pass 50 tokens, we will give ChatGPT $3 \times 50 = 150$ tokens so it will have a total of $150 + 100 = 250$ tokens;

if you decide to pass 100 tokens, we will give ChatGPT $3 \times 100 = 300$ tokens so it will have a total of $300 + 100 = 400$ tokens

After that, **ChatGPT (the Responder) will decide how many of the tokens it will give back to you**

In case this impacts your answer – here are the instructions we've given to ChatGPT:

"I would like to play the investment game with you. I'll be the investor and you will be the responder. As a starting point, you and I will each have 100 tokens. Whatever amount I decide to invest on you, there's a middle person who will triple that amount before passing it over to you. At that time, you will decide how much to pass back to me based on how much I invested and your total endowment. As a response, I need 4 numbers, the number I decided to pass over to you (label it "Investment amount"), the number that you will return back to me (label it "ChatGPT return"), total I have (label it "Total amount investor has") and total you've (label it "Total amount ChatGPT has")."

Please make sure that the total amount you and I have at the end of the game sums up to the amount I've left, the amount that was tripled by the middle person and the amount you had at the beginning of the game. Also ensure that every time we play, we start with a fresh endowment of 100 tokens each."

To make sure, you have a handle of the game, let's assume you decided to pass 20 tokens, how many total tokens does ChatGPT have after the researchers tripled the amount?

To make sure, you have a handle of the game, let's assume you decided to pass 50 tokens, how many total tokens does ChatGPT have after the researchers tripled the amount?

To make sure, you have a handle of the game, let's assume you decided to pass 80 tokens, how many tokens does ChatGPT have after the researchers tripled the amount?

Okay, let's play! **(Please do not use ChatGPT for this game, we will email you the results with the reward options)**

You now have 100 tokens and ChatGPT has 100 tokens. Don't forget, whatever number you decide to invest, we will triple it and pass it to ChatGPT. At that point, ChatGPT will decide how much to return back to you.

How many tokens would you like to pass to ChatGPT? (Please enter a number between 0 and 100)

- Next, could you guess how many tokens ChatGPT will give back to you. You will receive a bonus for a good guess (if your guess is within 5 tokens of the actual number), you will earn additional 10 tokens to redeem.

How much do you think ChatGPT has decided to give back to you? (The number should not be more than ChatGPT's endowment - $[3 * (\text{how much you decided to send})] + 100$)

End of survey

We thank you for your time spend participating in this study. Your response has been recorded. You might be selected to participate in a short follow-up interview.

.4.4 Grading Rubrics

Upskill Experiment - Coding Test

Assignment

Make sure to write and run the **Python code** needed to solve the question in this **Google Colab** notebook so that we may review your code.

Overview of Datasets

Dataset 1: Orders Data (`orders.csv`)

- **File Type:** CSV
- **Contents:**
 - `customer_id` : The unique identifier for each customer.
 - `order_info` : Information about each order in the format `order number ; date and time` . The order number in this dataset (once decoupled from the date and time) corresponds with that in the next.

Dataset 2: Products Data (`products.csv`)

- **File Type:** CSV
- **Contents:**
 - `customer_id` : The unique identifier for each customer associated with an order.
 - `order_id` : ID of each order in the format `order number` . The order number in this dataset corresponds with that in the previous once decoupled from the date and time.
 - `order_products` : Details about the products in each order in the following format: `{product_id: [product_cost, product_quantity], ...}` . Each product is sold either at its original price or a 20% discount.

Commonalities

- Both datasets share the `customer_id` field and `order_id` information with the order number.
- Each order ID is unique, whereas customer IDs may be repeated across multiple orders. However, each combination of order ID and customer ID is unique.
- Whenever you have information about either one of these, it is correct.

Data Quality and Cleaning Guidelines

- **Order and Customer IDs:** Entries are always correct when not NULL, and NULL values should be tried to be filled in wherever possible using data from elsewhere.
- **Date and Time Fields:** Entries with incorrect values should be removed.
- **Product Quantities and Product IDs:** Always correct unless marked as NULL, which indicates missing values.
- **Product Costs:** Each product ID is associated with a unique cost. Sometimes, the original unique cost is discounted at 20% so that is what is shown in a certain order. However, when not discounted,

sometimes there are junk or NULL values instead of the correct cost. Junk or NULL values in the product costs should be replaced with the original cost (the discounted cost should be left as is wherever it is shown but not added in elsewhere).

Tips for Handling Junk and NULL Values; duplicates

- Examine common values in each column to identify patterns and potential corrections.
- Attempt to fix junk or NULL values using information elsewhere in the data before considering row deletion.
- Date time fields can have incorrect fields that are not correctable, discard the affected rows and values to maintain data integrity.
- Check for duplicates at every stage

QUESTION:

Use the above datasets to answer the question: **Which 5 customer IDs had the highest average order by cost in May 2022?**

SOLUTION:

```
from google.colab import drive drive.mount('/content/drive')
```

Step 1: Start with loading and viewing the data

Imports and Reading in the File

1 point

```
In [1]: import pandas as pd
```

```
In [2]: orders = pd.read_csv('orders.csv')
```

```
In [3]: products = pd.read_csv('products.csv')
```

Studying the Files

[OPTIONAL]

```
In [4]: orders.head()
```

```
Out[4]:
```

	customer_id	order_info
0	107654909 15659	; 2023-08-04 01:33:45.202520256
1	251079598 14859	; 2023-11-23 14:53:10.639063904
2	344439380 17194	; 2023-03-23 04:04:23.546354640
3	752668623 15776	; 2020-10-09 19:55:36.453645364
4	705264936 17338	; 2021-10-18 08:03:18.739873984

```
In [5]: products.head()
```

```
Out[5]:
```

	index	customer_id	order_id	order_products
--	-------	-------------	----------	----------------

```

0    0  913891745.0    13702  {'129': [360, 178], '986': [391, 34], '317': [...
1    1  634096553.0    10795  {'317': [81, 175], '910': [0, 102], '129': [28...
2    2  189473854.0    12927                    {'910': [313, 103]}
3    3  774788031.0    13557  {'129': [NaT, 88], '722': [316, 17], '910': [3...
4    4  850303382.0    14520  {'129': [0, 125], '910': [313, 57], '986': [10...

```

Step 2: Split and clean the orders data

Remember that entries for customer IDs and order IDs are always correct when not NULL, and NULL values should be tried to be filled in wherever possible using data from elsewhere. Date and time fields entries with incorrect or junk values should be removed. And duplicates should be dropped.

Break up the order info

4 points

```
In [6]: orders['order_id'] = orders['order_info'].apply(lambda x: x.split(';')[0].strip())
```

```
In [7]: orders['order_date'] = orders['order_info'].apply(lambda x: x.split(';')[-1].strip())
```

Order date to datetime

2 points

```
In [8]: orders['order_date'] = pd.to_datetime(orders['order_date'], format='mixed')
```

Check order date for junk values

2 points for deleting and identifying junk values - THEY WILL NOT MAKE A DIFFERENCE TO THE CORRECTNESS

```
In [9]: orders.order_date.value_counts()
```

```
Out[9]: order_date
2099-12-31 23:59:59.000000000    558
2001-01-01 00:00:00.000000000    228
2022-05-03 07:25:26.192619264      5
2022-12-30 09:06:23.438343824      5
2021-05-29 23:50:38.343834384      5
...
2022-07-31 19:35:26.732673264      1
2021-01-28 22:44:14.905490548      1
2021-03-30 09:32:18.793879384      1
2022-01-17 14:16:27.218721872      1
2020-01-06 06:09:23.816381638      1
Name: count, Length: 10002, dtype: int64
```

Delete duplicates of Customer and Order ID together

3 points

```
In [10]: orders = orders[~orders[['customer_id', 'order_id']].apply(frozenset, axis=1).duplicated
```


Drop order_info column

[OPTIONAL]

```
In [11]: orders = orders[['customer_id', 'order_id', 'order_date']]
```

```
In [12]: orders.order_id.value_counts()
```

```
Out[12]: order_id
15659    1
18684    1
11383    1
19951    1
14804    1
..
11967    1
17991    1
11791    1
16555    1
13725    1
Name: count, Length: 10000, dtype: int64
```

Step 3: Clean the products data

Remember that entries for customer IDs and order IDs are always correct when not NULL, and NULL values should be tried to be filled in wherever possible using data from elsewhere.

Product Quantities and Product IDs: Always correct unless marked as NULL, which indicates missing values.

Product Costs: Each product ID is associated with a unique cost. Sometimes, the original unique cost is discounted at 20% so that is what is shown in a certain order. However, when not discounted, sometimes there are junk or NULL values instead of the correct cost. Junk or NULL values in the product costs should be replaced with the original cost (the discounted cost should be left as is wherever it is shown but not added in elsewhere).

Don't forget to drop duplicates!

Impute NULL values in products for customer_id in products dataframe

5 points

```
In [13]: products.describe()
```

```
Out[13]:
```

	index	customer_id	order_id
count	10000.00000	9.704000e+03	10000.00000
mean	4999.50000	5.465848e+08	15000.50000
std	2886.89568	2.632115e+08	2886.89568
min	0.00000	1.010432e+08	10001.00000
25%	2499.75000	3.076407e+08	12500.75000
50%	4999.50000	5.484692e+08	15000.50000
75%	7499.25000	7.769422e+08	17500.25000
max	9999.00000	9.987171e+08	20000.00000

```
In [14]: customers_dict = dict(zip(orders.order_id, orders.customer_id))
products['customer_id'] = products['order_id'].apply(lambda x: customers_dict[str(x)])
```

Delete junk values from order date in orders

[OPTIONAL]

```
In [15]: orders.order_date.value_counts()
```

```
Out[15]: order_date
2099-12-31 23:59:59.000000000    199
2001-01-01 00:00:00.000000000     74
2023-08-04 01:33:45.202520256      1
2021-03-20 07:44:18.145814584      1
2022-03-21 02:08:10.369036904      1
...
2022-03-24 17:44:41.908190816      1
2023-12-15 12:32:19.873987392      1
2020-05-08 11:49:42.178217822      1
2023-07-01 04:33:11.719171920      1
2020-01-06 06:09:23.816381638      1
Name: count, Length: 9665, dtype: int64
```

```
In [16]: orders = orders[~(orders.order_date.isin([pd.to_datetime('2099-12-31 23:59:59.000000000')
```

```
In [17]: orders.head()
```

```
Out[17]:
```

	customer_id	order_id	order_date
0	107654909	15659	2023-08-04 01:33:45.202520256
1	251079598	14859	2023-11-23 14:53:10.639063904
2	344439380	17194	2023-03-23 04:04:23.546354640
3	752668623	15776	2020-10-09 19:55:36.453645364
4	705264936	17338	2021-10-18 08:03:18.739873984

Read Products DF

[OPTIONAL]

```
In [18]: products.head()
```

```
Out[18]:
```

	index	customer_id	order_id	order_products
0	0	913891745	13702	{'129': [360, 178], '986': [391, 34], '317': [...
1	1	634096553	10795	{'317': [81, 175], '910': [0, 102], '129': [28...
2	2	189473854	12927	{'910': [313, 103]}
3	3	774788031	13557	{'129': [NaT, 88], '722': [316, 17], '910': [3...
4	4	850303382	14520	{'129': [0, 125], '910': [313, 57], '986': [10...

Replace all prices with correct price

10 points

```

In [19]: import json

In [20]: products['order_products'] = products['order_products'].apply(lambda x: json.loads(x.rep

In [21]: from collections import defaultdict
from collections import Counter

In [22]: def product_search(x):
    for key in list(x.keys()):
        product_dict[key].append(x[key][0])

In [23]: product_dict = defaultdict(list)
products['order_products'].apply(lambda x: product_search(x))

Out[23]: 0      None
1      None
2      None
3      None
4      None
...
9995   None
9996   None
9997   None
9998   None
9999   None
Name: order_products, Length: 10000, dtype: object

In [24]: from statistics import mode

In [25]: for key in product_dict.keys():
    print(key)
    print(Counter(product_dict[key]))

129
Counter({360: 4587, 288.0: 748, -100: 315, 0: 246, 100: 166})
986
Counter({391: 4542, 312.8: 727, -100: 284, 0: 227, 100: 190})
317
Counter({81: 4573, 64.8: 738, -100: 266, 0: 242, 100: 187})
722
Counter({316: 4617, 252.8: 694, -100: 310, 0: 251, 100: 197})
910
Counter({313: 4586, 250.4: 716, -100: 304, 0: 255, 100: 167})

In [26]: for key in product_dict.keys():
    product_dict[key] = mode(product_dict[key])

In [27]: def correct_cost(x):
    for key in x.keys():
        if x[key][0] == 100:
            x[key][0] = product_dict[key]
        elif x[key][0] == -100:
            x[key][0] = product_dict[key]
        elif x[key][0] == 0:
            x[key][0] = product_dict[key]
    return x

In [28]: products['order_products'] = products['order_products'].apply(lambda x: correct_cost(x))

```

Step 4: Merge the data sets to answer the question: Which 5 customer IDs had the highest average order by cost in May 2022?

Merge the data and get final answer

5 points

```
In [29]: products.order_id = products.order_id.apply(str)
```

```
In [30]: final = orders.merge(products, right_on=['customer_id', 'order_id'], left_on=['customer_
```

```
In [31]: final['order_month'] = final['order_date'].apply(lambda x: str(x.year) + ' ' + str(x.mon
```

```
In [32]: final['total_order_cost'] = final['order_products'].apply(lambda x: sum([y[0]*y[1] for y
```

```
In [33]: may_2022 = final[final['order_month'] == '2022 5']
```

```
In [34]: pd.DataFrame(may_2022.groupby('customer_id')['total_order_cost'].mean()).sort_values('to
```

```
Out[34]:
```

	total_order_cost
--	------------------

customer_id	
585775494	224539.0
613911991	204480.0
349369215	202248.0
272552610	199774.0
723415497	197863.0

INSTRUCTIONS

- Do not Google image search or send any images to GPT. Refrain from copying and pasting the exact question into Google or GPT unless completely stuck. Do not spend more than 1.5 hours on this task.

Question 1: The following is the first five rows of data containing financial and demographic information about domestic partners who have co-purchased a home in the last several years. Please note that the following table is illustrative and represents a snapshot sample of the data to solve this problem. All the information you need to solve the problem is contained within this snapshot.

Age 1	Age 2	Income 1	Income 2	Borough	ZIP Code	Date	Price	Mortgage
39	37	270000	180000	Manhattan	10076	1 January 2016	1,125,000	Yes
NULL	38	445000	670000	Manhattan	10025	1 January 2016	2,249,000	Yes
27	29	145000	225000	Queens	11106	2 January 2016	900,000	Yes
33	NULL	90000	76000	Brooklyn	11203	2 January 2016	415,000	Yes
68	55	78000	450000	Bronx	10474	2 January 2016	3,399,000	No

1. You have been tasked with predicting based on demographics and price whether a mortgage was taken out to by the house. You prompt ChatGPT for detailed instructions on how to do this, and ChatGPT recommend using a logistic regression model. It recommends the following steps.

1. Understand Your Dataset

- **Explore and Preprocess:** Start by exploring your dataset to understand the features available and their types (numerical, categorical). Clean the data by handling outliers and possibly irrelevant features. Preprocessing steps like encoding techniques (e.g., one-hot encoding) might be necessary for categorical data. Ensure that your dataset does not have missing values. You can either fill them in with a strategy (mean, median, mode) or remove the rows/columns with missing values, depending on the situation.

- a. Which of the following are among the steps you could take to address this point? Select all that apply.

- i. Plot the distribution of each of the numerical variables and remove rows with outliers from this dataset **+0.5 points**
- ii. One-hot encode the 'Borough' variable **+0.5 points**
- iii. Investigate relationships between variables **+0.5 points**
- iv. Convert date to a numerical variable **0 points**

- v. One-hot encode the ZIP code variable **-1 point only if vii. not selected**
- vi. One-hot encode the age variables **-0.5 points**
- vii. Bin the ZIP codes by neighborhoods and do not process further **-1 point**
- viii. Bin the ZIP codes by neighborhoods and one-hot encode **+1 point only if v. or vii. not selected**
- ix. Check columns with null values and remove those with >80% missing values **+0.5 points**
- x. Impute NULL values by using a summary statistic or by developing a simple model that predicts those values based on other features **+0.5 points**

-1 if more than one about ZIP code selected

Maximum: 3.5 points

Minimum: 0 points

Explanation of point assignment: There is only one reasonable handling of the ZIP code variable. There are several issues with one-hot encoding the raw ZIP code variable, including the curse of dimensionality and sparse resultant data. No reasonable data scientist would make that choice and therefore it is wrong. You also cannot bin without one-hot encoding because binned data is categorical. Choosing either one is a subtraction of a whole point but not an immediate 0. Given that this choice is more difficult there is more positive credit for getting this right than for getting other correct answers.

Date to numeric is contentious, therefore 0 penalty or reward (date is ordinal but sometimes represented as a number, although never treated like a numeric in the sense that you would never take a summary statistic such as mean or median of the date column (e.g. if you represent months or years numerically you would never take a mean of those), you would only take summary statistics such as mode because it is essentially categorical).

No one would ever one-hot encode age unless binned.

2. Split the Data

- **Train-Test Split:** Divide your dataset into a training set and a testing set (commonly a 70-30 or 80-20 split) to evaluate the model's performance on unseen data.

3. Train the Model

- **Training:** Use the training dataset to train your model, adjusting parameters as needed. For complex models, consider using cross-validation to fine-tune hyperparameters and prevent overfitting.

4. Evaluate the Model

- **Performance Metrics:** Evaluate your model on the test set using appropriate metrics such as accuracy, precision, recall, F1 score, and the ROC-AUC curve. These metrics will help you understand how well your model is performing in terms of both its ability to predict mortgages correctly and its robustness against false positives or negatives.

b. What issue necessitates using all these metrics? Which of the above steps is affected by this issue and how? (Answer in 100 words or less – bullet points ok)

(3 points – imbalanced data is the issue

OR

1 point for overfitting being recognized as the issue without reference to imbalance)

AND

2 points – affects step 2 (train-test split), as stratified sampling would be a fix (full credit for mention of stratified sampling even if step 2 not mentioned)

Total 5 points

c. Would you change the order of any of the above steps? Why or why not? (Answer in 100 words or less – bullet points ok)

2 points for identifying that steps 1 and 2 should be switched.

3 points for identifying that the issue is data leakage.

Total 5 points

2. You want to try a k-Nearest Neighbors model. Which of the following are not required (although recommended) for logistic regression, but absolutely necessary for k-Nearest Neighbors? Select all that apply.

a. Transform numerical variables (e.g. log) +2 points

b. Make sure there are only two classes to predict -2 points

c. Convert Mortgage column from string to binary -1 points

d. Standardize numerical variables +2 points

e. Impute the missing age with the other age in the same row -1 points

f. One-hot encode the appropriate variables -1 points

Maximum 4, minimum 0

Explanation of point assignment: The key here is what is absolutely necessary for kNN but not required for logistic regression, although recommended. Option (b) is not required at all for kNN so it has the biggest point subtraction. Option (c), (e), and (f) would need to be done for either model so they are not correct but do not warrant as much subtraction because they are not as egregious. The correct answers are (a) and (d) which are the numerical transformations, which kNN is particularly sensitive to.

3. You also try a decision tree model for the same classification problem, to compare performance. You realize your model is performing quite poorly on both training and validation sets. You double-check the code and there are no bugs. What could be causing this problem? Select all that apply.

a. Your model is underfit

b. Your model is overfit

c. The learning rate hyperparameter is too small

- d. The learning rate hyperparameter is too large
 - e. The decision tree is too shallow
 - f. The decision tree is too deep
 - g. None of the above
- h. Award the following points for the selection of the options, with a minimum of 0 and a maximum of 4:
- i. a = +2 points
 - j. b = -4 points
 - k. c = -2 points
 - l. d = -2 points
 - m. e = +2 points
 - n. f = -4 points
 - o. g = 0 points

Explanation of point assignment: Wrong answers are egregiously wrong (opposite of what is correct) and each result in an immediate 0. Two middle answers subtract only half of total points because they are not egregiously wrong but indicate a fundamental misunderstanding of decision tree model (which is one of the most simple models to understand). Therefore, if you pick the 2 correct answers and one of the hyperparameter answers, you get half of the total credit. Selecting one of the hyperparameter answers indicates that you might be guessing / selecting one of each and hoping for the best.

4. Next, you have been instructed to predict the price based on the other variables, and this time you have been instructed to use linear regression. Following instructions from ChatGPT, you perform a basic linear regression. You notice that your R^2 value is too low. You prompt ChatGPT for suggestions on how to diagnose the problem, and it is recommended that you check the residual plots. You notice that the residual plot does not appear random. What could this mean? Select all that apply.
- a. The observed values of your dependent variable are independent from each other -2 points
 - b. Your model is missing an important variable +1 points
 - c. There is some interaction between your variables +1 points
 - d. A higher order term might be required in your regression +1 points
 - e. Variance of the residual is the same for any value of X -2 points

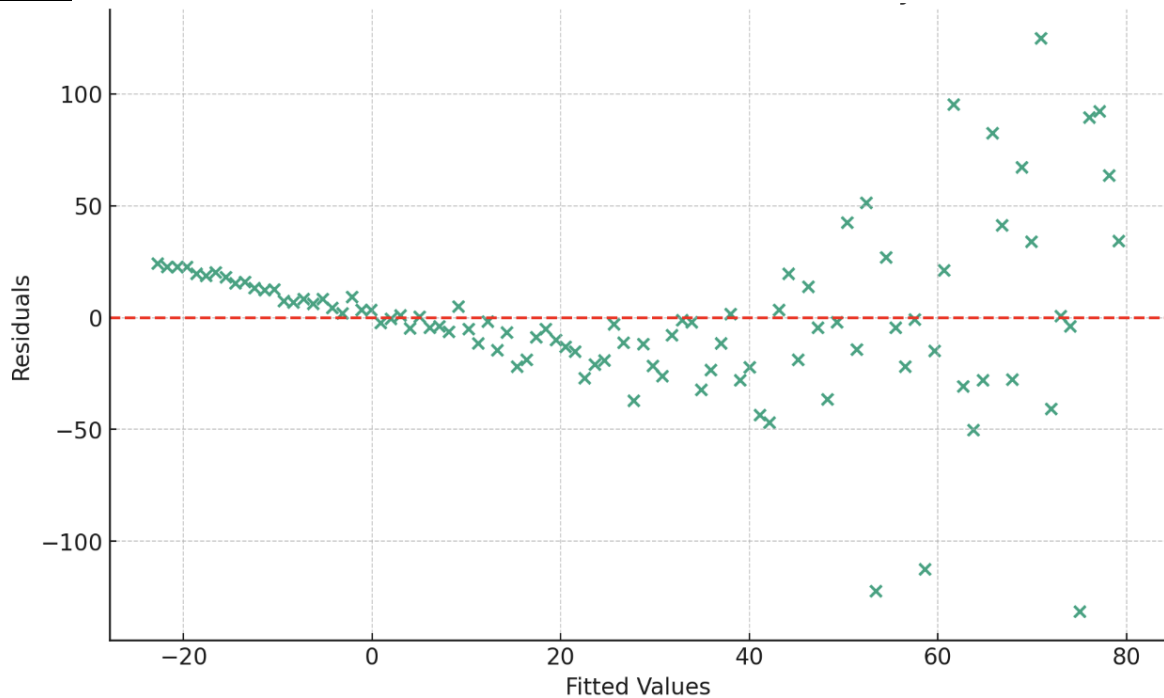
Explanation of point assignment: Wrong answers are unequivocally wrong as to things that would cause a low R^2 value (these are assumptions needed in order for regression to work, they would only be correct in the opposite version of those statements), they each subtract 2 but are not an immediate 0 unless both incorrect answers picked or 3 correct answers not picked

Max 3 min 0

5. For the following residual plots, what could be the characteristics of or issues with the data or model that are corresponding with these results (choose from the list provided for each image)? It is possible that more than one characteristic or issue applies to any given image, and it is possible that a characteristic or issue may apply to more than one image.

Explanation of point assignment: Offer no credit or no penalty for understandable mistakes, offer penalty for egregious mistakes, and offer reward for correctness, weighted by difficulty of getting the correct answers

Image A

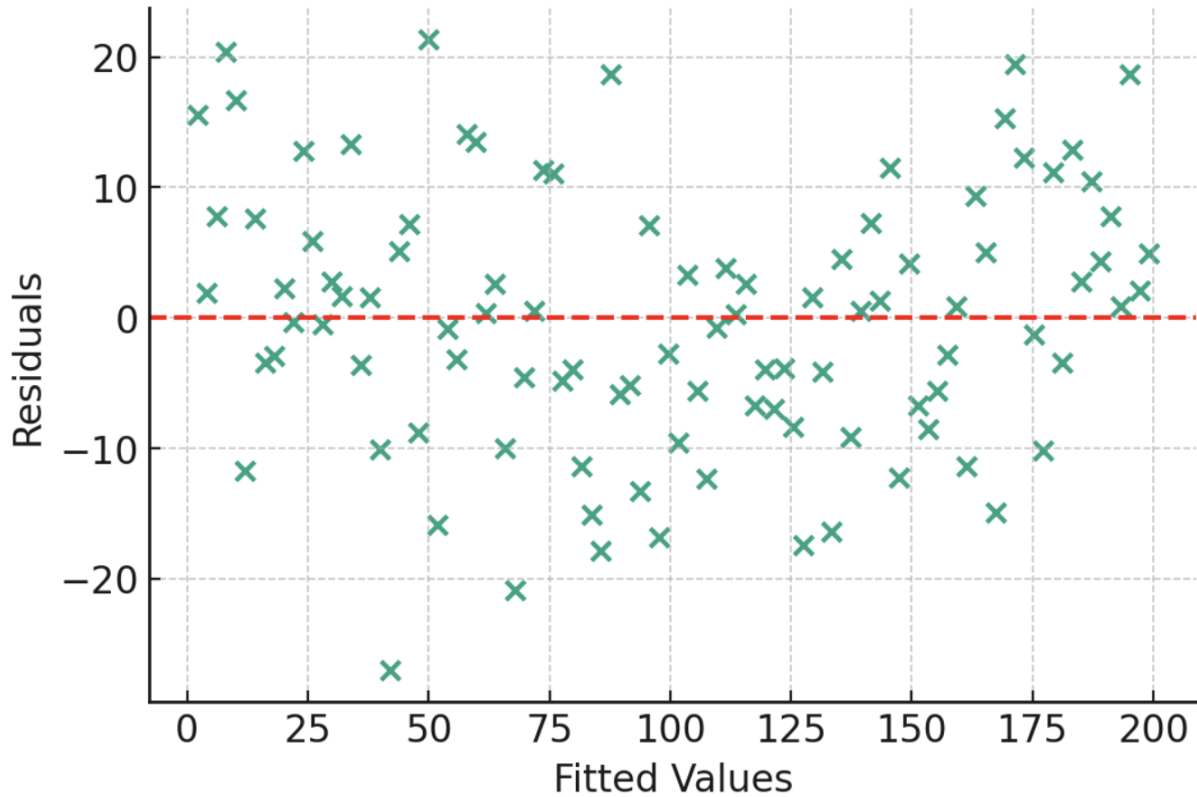


- a. No characteristic or issue is apparent **-2 points**
- b. Heteroscedastic data **1 points**
- c. Outliers **0 points**
- d. Response variable requires transformation **-1 points**
- e. A higher order variable might be required **2 points**

Max 3 min 0

Explanation of point assignment: Clearly there is a characteristic/issue apparent, outliers might be confusing, there's no apparent transformation for the response variable but there is evidence of heteroscedasticity (obvious) and non-linearity and higher order variables (not as obvious but still obvious)

Image B

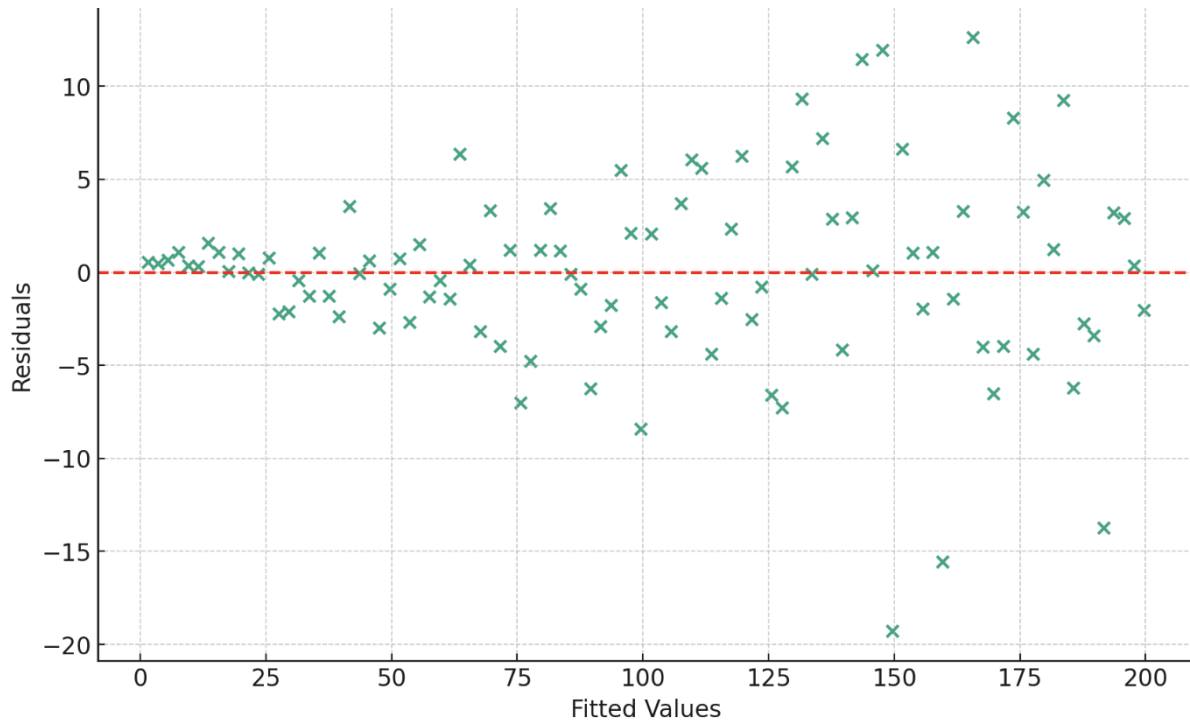


- a. No characteristic or issue is apparent **2 points**
- c. Heteroscedastic data **-2 points**
- d. Outliers **0 points**
- e. Response variable requires transformation **-2 points**
- f. A higher order variable might be required **-2 points**

Max 2 min 0

Explanation of point assignment: Clearly there is a characteristic/issue apparent, this should be identifiable as a normal residual plot, no penalty for outliers because inevitable confusion

Image C

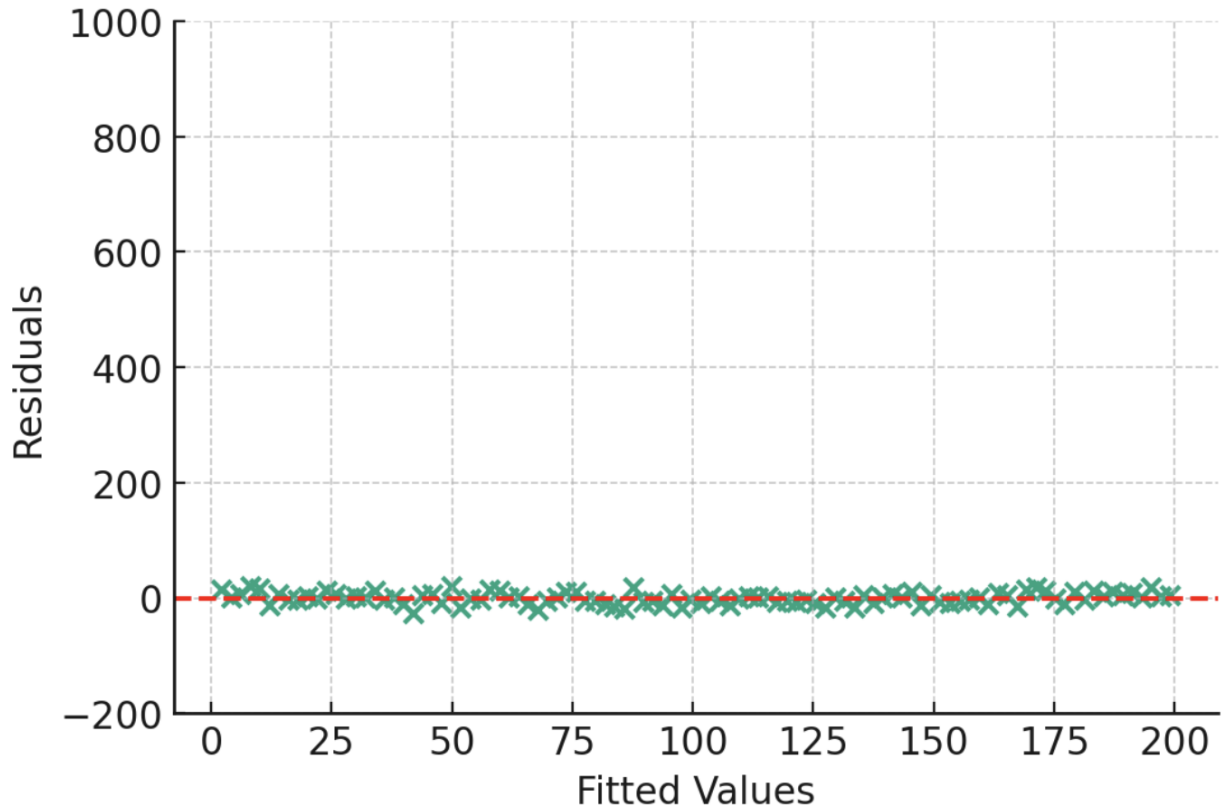


- a. No characteristic or issue is apparent **-2 points**
- d. Heteroscedastic data **2 points**
- e. Outliers **0 points**
- f. Response variable requires transformation **-2 points**
- g. A higher order variable might be required **-2 points**

Max 2 min 0

Explanation of point assignment: Clearly there is a characteristic/issue apparent, outliers might be confusing, there's no apparent transformation for the response variable, and there is a clear linear center making general linearity apparent

Image D

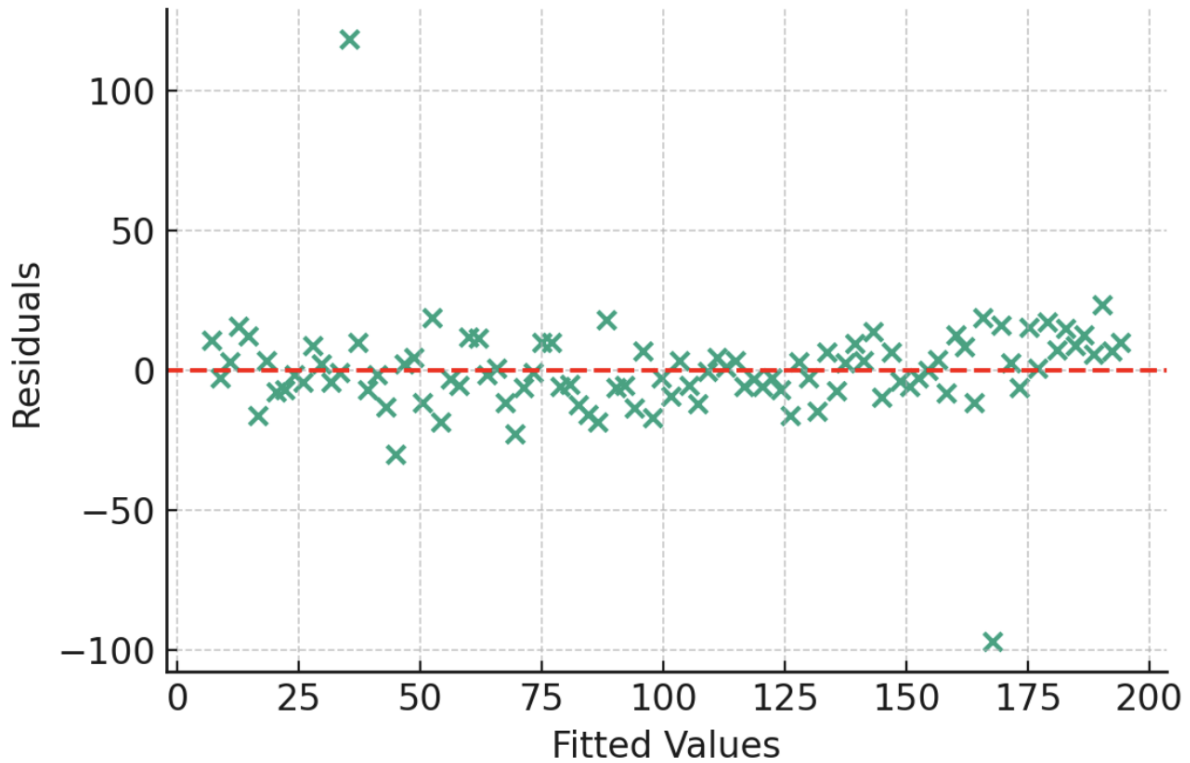


- a. No characteristic or issue is apparent **0 points**
- e. Heteroscedastic data **-2 points**
- f. Outliers **-2 points**
- g. Response variable requires transformation **2 points**
- h. A higher order variable might be required **-2 points**

Max 2 min 0

Explanation of point assignment: Some might think that this is an example of a normal residual plot zoomed out, since this is confusing there is no penalty. There is no visible heteroscedasticity or outlier, and there's no visible nonlinearity. It is quite apparent that the response variable requires transformation.

Image E

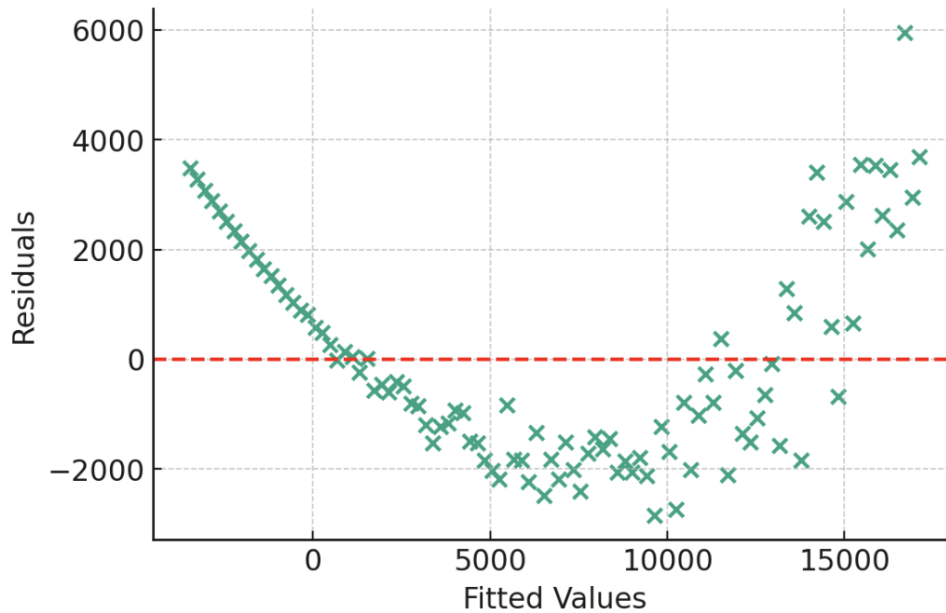


- a. No characteristic or issue is apparent **0 points**
 - f. Heteroscedastic data **-2 points**
 - g. Outliers **2 points**
 - h. Response variable requires transformation **0 points**
 - i. A higher order variable might be required **-2 points**

Max 2 min 0

Explanation of point assignment: Some might think that this is an example of a normal residual plot zoomed out, since this is confusing there is no penalty. There is no visible heteroscedasticity, nonlinearity, or requirement of transformation (although transformation requirement is confusing, so no penalty). Seemingly obvious for outliers.

Image F

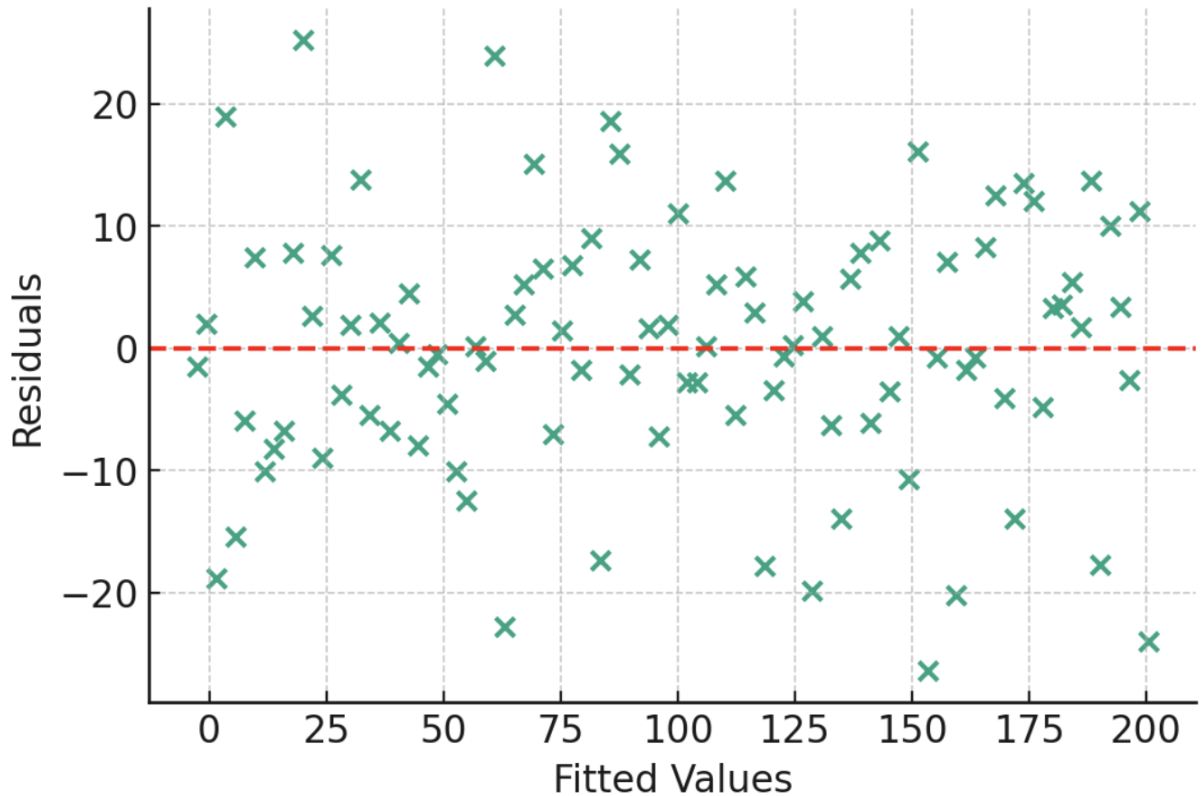


- a. No characteristic or issue is apparent **-2 points**
- g. Heteroscedastic data **2 points**
- h. Outliers **0 points**
- i. Response variable requires transformation **-2 points**
- j. A higher order variable might be required **1 points**

Max 3 min 0

Explanation of point assignment: Clearly there is a characteristic/issue apparent, outliers might be confusing, there's no apparent transformation for the response variable but there is clear evidence of non-linearity and higher order variables (obvious) and heteroscedasticity (not as obvious but still obvious)

Image G



- a. No characteristic or issue is apparent **2 points**
- h. Heteroscedastic data **-2 points**
- i. Outliers **0 points**
- j. Response variable requires transformation **-2 points**
- k. A higher order variable might be required **-2 point**

Max 2 min 0

Explanation of point assignment: Clearly there is a characteristic/issue apparent, this should be identifiable as a normal residual plot, no penalty for outliers because inevitable confusion

Characteristic or issues choices:

- a. No characteristic or issue is apparent
 - i. Heteroscedastic data
 - j. Outliers
 - k. Response variable requires transformation
 - l. A higher order variable might be required
6. You are asked to train a new model to predict price on the newest version of the dataset. In this version, there are several more fields collected with demographic information and financial information of the couples. However, this data is only from the last month. Which of the following steps recommended by ChatGPT could be beneficial to take to address some of the issues that are likely to arise because of this? Select all that apply.

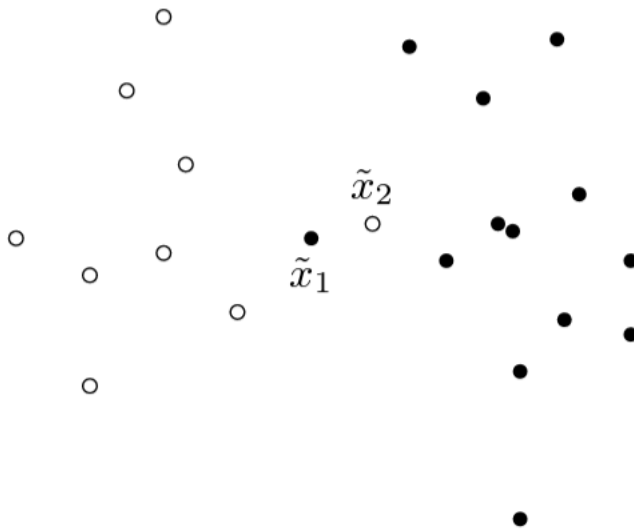
- a. Perform PCA +2
- b. Use a neural network instead of linear regression -4
- c. Use a regularized model instead of linear regression +2
- d. None of the above 0 points

Explanation of point assignment: Neural networks perform worse on less observations, this is an immediate 0

Max 4 min 0

Question 2:

1. You are asked to prepare a simple linear model to classify the following points into class 1 (black dots) and class 2 (white dots). What is the best empirical risk of this model that you can achieve with 0-1 loss? Justify your answer and show your working steps.



The formula for empirical risk is

$$\frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2$$

or Loss / N

Award 4 points for: The correct answer is 1/22, which is the minimal loss
OR

Award 3 points for: 21/22

OR

Award 1 point for: 2/22 or 20/22 [partial process without realizing that only one point will be misclassified, not 2 in the best case]

OR

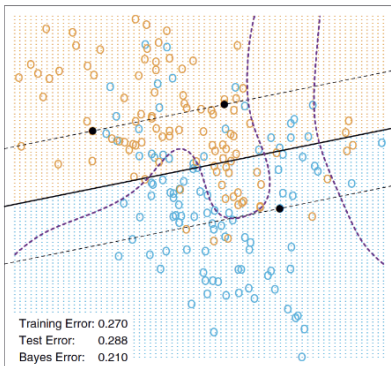
Award 1 point for: Identifying there will be 1 misclassification at best but not knowing what to do further

What you need to figure out to answer the question:

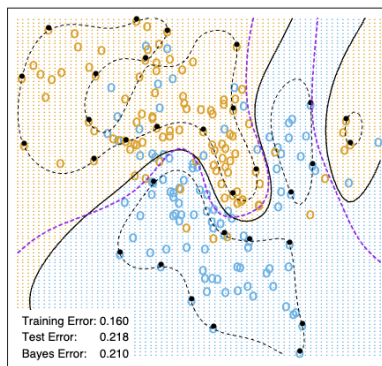
- With 0-1 Loss, this turns into # of misclassifications / # of observations
- With a linear classifier, you would at best misclassify at least 1 observation

Explanation of point assignment: Want to award partial credit for frequent errors where some of the process is correct

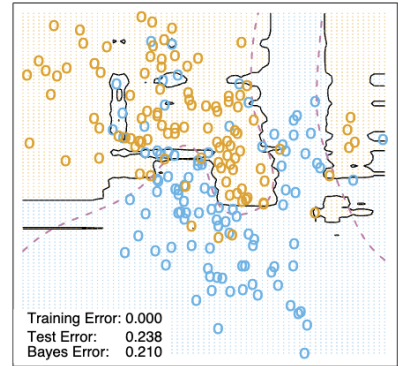
2. ChatGPT has run 3 classifiers on your data and provided a visual output, but not specified which models yielded which output. For each of the three images, name a classifier that could create the boundary represented by the solid black line, and one that could not (class 1 is the orange dots, and class 2 is the blue dots). You can ignore the dashed line, you can use the metrics on the bottom-left but you do not need them. Justify your answer.



(a)



(b)



(c)

(a)

2 points, can create this boundary (any): Logistic regression, linear SVM (support vector machine), Naïve Bayes or other linear classification model

2 points, cannot create this boundary: Any of the others

Max 2 points

(b)

2 points, can create this boundary (any): Multi-layer perceptron, poly kernel SVM, sigmoid kernel SVM, kNN (k nearest neighbors), GAUSSIAN Naïve Bayes, other valid (regular Naïve Bayes not acceptable)

2 points, cannot create this boundary: Any of the linear models, ideally

Max 2 points

(c)

2 points, can create this boundary (any): Decision tree or random forest, RBF kernel SVM, kNN

2 points, cannot create this boundary: Any of the linear models, ideally, or any of the other kernels on SVM

Max 2 points

Explanation of point assignment: People (in pilots and experiment) tend not to answer the flip side (which model is unable to create this decision boundary), so there is no additional credit for it but there is credit for getting it right if you get the other wrong

Question 3: Imagine you're a logistics manager and one of your delivery trucks has gone missing. You believe it lost its signal while on either Route A or Route B, with a 65% and 35% chance of being on each route respectively. Based on the coverage area of these routes, if the truck is on Route A and you search for an entire day, there's a 45% chance you'll find it. However, if it's on Route B and you search for a day, the probability of locating is 75%.

1. If you only had one day to search for the truck, on which route would you focus your search efforts in order to maximize your chances of finding it? Please explain your choice and breakdown your calculations.

The first step is to calculate the probability of finding the truck in each Route, given that the truck is lost with 65% chance in Route A and 35% in Route B.

Probability of finding the truck in Route A = $P(\text{Find in A} \mid \text{Truck in Route A}) \cdot P(\text{Truck in Route A}) = 0.65 \cdot 0.45 = 0.2925$ (2 points)

Probability of finding the truck in Route B = $P(\text{Find in B} \mid \text{Truck in Route B}) \cdot P(\text{Truck in Route B}) = 0.35 \cdot 0.75 = 0.2625$ (2 points)

You should search in Route A. (1 point)

SUMMARY:

Award 1 point for Route A without explanation

Award 5 points for Route A with accompanying process steps

2. Assume that you made the rational decision on the first day, but didn't manage to locate the truck. The truck remains at the position that it was originally lost at and has not been moved. You have another day committed for search - has your initial idea of which route the truck is on changed? Where should you search now? Please explain your choice and breakdown your calculations.

The rational decision was to search in Route A. Since you made that decision and did not find your truck, here is the new calculation. Now we have more information about the probability that the truck is in Route A past the priori probability. (3 points for this realization, if unaccompanied by correct calculations)

$P(\text{Posterior Truck in A} \mid \text{Truck not found on Day 1 in A}) = \frac{P(\text{Truck not found on Day 1 in A} \mid \text{Truck in A}) \cdot P(\text{Prior Truck in A})}{P(\text{Truck not found on Day 1 in A})}$ (2 points)

$P(\text{Truck not found on Day 1 in A} \mid \text{Truck in A}) = 0.55$

$P(\text{Prior truck in A}) = 0.65$

$P(\text{Truck not found on Day 1 in A}) = P(\text{Truck not found on Day 1 in A} \mid \text{Route A}) + P(\text{Truck not found in Day 1 in A} \mid \text{Route B}) = 0.55 + 1 = 1.55$

Therefore $P(\text{Posterior Truck in A} \mid \text{Truck not found on Day 1 in A}) = 0.55 \cdot 0.65 / 1.55 = 0.231$ (3 points)

Probability of finding the truck in Route A = $P(\text{Find in A} \mid \text{In Route A}) \cdot P(\text{Posterior Truck in A}) = 0.23 \cdot 0.45 = 0.1035$ (2 points)

Probability of finding the truck in Route B = $P(\text{Find in B} \mid \text{Route B}) \cdot P(\text{Posterior Truck in Route B}) = 0.35 \cdot (1 - 0.23) = 0.269$ (2 points)

Therefore, you should switch to Route B. (1 point)

SUMMARY:

Award 1 point for Route B without explanation.

Award 3 points for realization or intuition that probability numbers have changed (prior probabilities are no longer valid).

Award 10 points for Route B with accompanying process steps.

Bibliography

Abebe, Girum, A Stefano Caria, and Esteban Ortiz-Ospina, “The selection of talent: experimental and structural evidence from Ethiopia,” *American Economic Review*, 2021, *111* (6), 1757–1806.

Acemoglu, Daron and Jorn-Steffen Pischke, “Why Do Firms Train? Theory and Evidence,” *The Quarterly Journal of Economics*, 1998, *113* (1), 79–119.

— **and Pascual Restrepo**, “The race between man and machine: Implications of technology for growth, factor shares, and employment,” *American economic review*, 2018, *108* (6), 1488–1542.

Agarwal, Nikhil, Alex Moehring, Pranav Rajpurkar, and Tobias Salz, “Combining human expertise with artificial intelligence: Experimental evidence from radiology,” Technical Report, National Bureau of Economic Research 2023.

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

Agrawal, Ajay, John Horton, Nicola Lacetera, and Elizabeth Lyons, “Digitization and the contract labor market,” *Economic Analysis of the Digital Economy*, 2015, 219.

— , **Joshua Gans, and Avi Goldfarb**, “Prediction, judgment, and complexity: a theory of decision-making and Artificial Intelligence,” in “The economics of Artificial Intelligence: an agenda,” University of Chicago Press, 2018, pp. 89–110.

—, —, and —, *Prediction machines: the simple economics of Artificial Intelligence*, Harvard Business Press, 2018.

—, **Nicola Lacetera, and Elizabeth Lyons**, “Does standardized information in online markets disproportionately benefit job applicants from less developed countries?,” *Journal of international Economics*, 2016, *103*, 1–12.

Alvero, AJ, Sonia Giebel, Ben Gebre-Medhin, Anthony Lising Antonio, Mitchell L Stevens, and Benjamin W Domingue, “Essay content and style are strongly related to household income and SAT scores: Evidence from 60,000 undergraduate applications,” *Science advances*, 2021, *7* (42), eabi9031.

Barron, John M. and John Bishop, “Extensive Search, Intensive Search, and Hiring Costs: New Evidence on Employer Hiring Activity,” *Economic Inquiry*, 1985, *23* (3), 363–382.

Belot, Michèle, Philipp Kircher, and Paul Muller, “Providing advice to jobseekers at low cost: an experimental study on online advice,” *The Review of Economic Studies*, 2018, *86* (4), 1411–1447.

Bertrand, Marianne and Sendhil Mullainathan, “Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination,” *American Economic Review*, 2004, *94* (4), 991–1013.

Blatter, Marc, Samuel Muehlemann, and Samuel Schenker, “The costs of hiring skilled workers,” *European Economic Review*, 2012, *56* (1), 20–35.

Bolton, Gary, Ben Greiner, and Axel Ockenfels, “Engineering trust: reciprocity in the production of reputation information,” *Management Science*, 2013, *59* (2), 265–285.

Briscese, Guglielmo, Giulio Zanella, and Veronica Quinn, “Providing government assistance online: a field experiment with the unemployed,” *Journal of Policy Analysis and Management*, 2022, *41* (2), 579–602.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell et al., “Language models are few-shot learners,” *Advances in Neural Information Processing Systems*, 2020, 33, 1877–1901.

Brynjolfsson, Erik, Danielle Li, and Lindsey R Raymond, “Generative AI at work,” Technical Report, National Bureau of Economic Research 2023.

– , – , **and** – , “Generative AI at work,” Technical Report, National Bureau of Economic Research 2023.

Cai, Hongbin, Ginger Zhe Jin, Chong Liu, and Li an Zhou, “Seller reputation: from word-of-mouth to centralized feedback,” *International Journal of Industrial Organization*, 2014, 34, 51–65.

Card, David, Jochen Kluge, and Andrea Weber, “Active labour market policy evaluations: A meta-analysis,” *The economic journal*, 2010, 120 (548), F452–F477.

Chan, Jason and Jing Wang, “Hiring preferences in online labor markets: Evidence of a female hiring bias,” *Management Science*, 2018, 64 (7), 2973–2994.

Crépon, Bruno, Esther Duflo, Marc Gurgand, Roland Rathelot, and Philippe Zamora, “Do labor market policies have displacement effects? Evidence from a clustered randomized experiment,” *The Quarterly Journal of Economics*, 2013, 128 (2), 531–580.

Dell’Acqua, F., E. McFowland, E. R. Mollick, H. Lifshitz-Assaf, K. Kellogg, S. Rajendran, and K. R. Lakhani, “Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality,” Technical Report, Harvard Business School Technology & Operations Mgt. Unit Working Paper 2023.

Dell’Acqua, Fabrizio, “Falling asleep at the wheel: Human/AI Collaboration in a Field Experiment on HR Recruiters,” Technical Report, Working paper 2022.

- Deming, David J**, “The growing importance of social skills in the labor market,” *The Quarterly Journal of Economics*, 2017, 132 (4), 1593–1640.
- **and Kadeem Noray**, “Earnings dynamics, changing job skills, and STEM careers,” *The Quarterly Journal of Economics*, 2020, 135 (4), 1965–2005.
- den Broek, Elmira Van, Anastasia Sergeeva, and Marleen Huysman**, “When the Machine Meets the Expert: An Ethnography of Developing AI for Hiring.,” *MIS quarterly*, 2021, 45 (3).
- Djankov, Simeon and Federica Saliola**, “The changing nature of work,” *Journal of International Affairs*, 2018, 72 (1), 57–74.
- Eloundou, Tyna, Sam Manning, Pamela Mishkin, and Daniel Rock**, “GPTs Are GPTs: An early look at the labor market impact potential of large language models,” *arXiv preprint arXiv:2303.10130*, 2023.
- Farber, Henry S, Dan Silverman, and Till Von Wachter**, “Determinants of callbacks to job applications: An audit study,” *American Economic Review*, 2016, 106 (5), 314–18.
- Felten, Edward W, Manav Raj, and Robert Seamans**, “Occupational heterogeneity in exposure to generative AI,” *SSRN*, 2023.
- Filippas, Apostolos, Andrey Fradkin, and John J Horton**, “Subsidizing Job Search Considered Harmful: Evidence from a Field Experiment (Preliminary),” 2023.
- , **John Joseph Horton, and Joseph Golden**, “Reputation inflation,” *Marketing Science*, 2022.
- Flesch, Rudolph**, “A new readability yardstick.,” *Journal of applied psychology*, 1948, 32 (3), 221.
- Fradkin, Andrey, Elena Grewal, and David Holtz**, “Reciprocity and unveiling in two-sided reputation systems: evidence from an experiment on airbnb,” *Marketing Science*, 2021, 40 (6), 1013–1029.

- Ghose, Anindya and Panagiotis G Ipeirotis**, “Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics,” *IEEE transactions on knowledge and data engineering*, 2010, 23 (10), 1498–1512.
- Goldfarb, Avi and Catherine Tucker**, “Digital economics,” *Journal of Economic Literature*, 2019, 57 (1), 3–43.
- Heckman, James J, Lance Lochner, and Christopher Taber**, “Explaining rising wage inequality: Explorations with a dynamic general equilibrium model of labor earnings with heterogeneous agents,” *Review of economic dynamics*, 1998, 1 (1), 1–58.
- HF canonical model maintainers**, “DistilBERT-base-uncased-finetuned-sst-2-English (Revision bfdd146),” 2022.
- Hong, Yili, Jing Peng, Gordon Burtch, and Ni Huang**, “Just DM me (politely): direct messaging, politeness, and hiring outcomes in online labor markets,” *Information Systems Research*, 2021, 32 (3), 786–800.
- Horton, John J.**, “Online labor markets,” *Internet and Network Economics: 6th International Workshop, WINE 2010, Stanford, CA, USA, December 13-17, 2010. Proceedings*, 2010.
- , “The effects of algorithmic labor market recommendations: evidence from a field experiment,” *Journal of Labor Economics*, 2017, 35 (2), 345–385.
- Jha, S. and S. S. Bhattacharyya**, “Learning orientation and performance orientation: Scale development and its relationship with performance,” *Global Business Review*, 2013, 14 (1), 43–54.
- Kang, Sonia K, Katherine A DeCelles, András Tilcsik, and Sora Jun**, “Whitened résumés: race and self-presentation in the labor market,” *Administrative Science Quarterly*, 2016, 61 (3), 469–502.
- Kincaid, J Peter, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom**, “Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel,” 1975.

- Kokkodis, Marios and Sam Ransbotham**, “Learning to successfully hire in online labor markets,” *Management Science*, 2022.
- Lee, David S**, “Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects,” *Review of Economic Studies*, 2009, 76 (3), 1071–1102.
- Luca, Michael and Oren Reshef**, “The effect of price on firm reputation,” *Management Science*, 2021, 67 (7), 4408–4419.
- Mankiw, N Gregory and Michael D Whinston**, “Free entry and social inefficiency,” *The RAND Journal of Economics*, 1986, pp. 48–58.
- Marinescu, Ioana**, “The general equilibrium impacts of unemployment insurance: Evidence from a large online job board,” *Journal of Public Economics*, 2017, 150, 14–29.
- **and Ronald Wolthoff**, “Opening the black box of the matching function: the power of words,” *Journal of Labor Economics*, 2020, 38 (2), 535–568.
- Martin-Lacroux, Christelle and Alain Lacroux**, “Do employers forgive applicants’ bad spelling in résumés?,” *Business and Professional Communication Quarterly*, sep 2017, 80 (3), 321–335.
- Miron, E., M. Erez, and E. Naveh**, “Do personal characteristics and cultural values that promote innovation, quality, and efficiency compete or complement each other?,” *Journal of Organizational Behavior*, 2004, 25 (2), 175–199.
- Mok, Aaron**, “Don’t expect ChatGPT to help you land your next job,” Oct 2023. Accessed: 2024-02-22.
- Mollick, Ethan R and Lilach Mollick**, “New Modes of Learning Enabled by AI Chatbots: Three Methods and Assignments,” 12 2022. Available at SSRN: <https://ssrn.com/abstract=4300783> or <http://dx.doi.org/10.2139/ssrn.4300783>.
- Moss-Racusin, Corinne A., John F. Dovidio, Victoria L. Brescoll, Mark J. Graham, and Jo Handelsman**, “Science faculty’s subtle gender biases favor male students,” *Proceedings of the National Academy of Sciences*, 2012, 109 (41), 16474–16479.

- Noy, Shakked and Whitney Zhang**, “Experimental evidence on the productivity effects of generative Artificial Intelligence,” *SSRN*, 2023.
- Oestreicher-Singer, Gal and Arun Sundararajan**, “The visible hand? demand effects of recommendation networks in electronic markets,” *Management Science*, 2012, 58 (11), 1963–1981.
- Oreopoulos, Philip**, “Why do skilled immigrants struggle in the labor market? A field experiment with thirteen thousand resumes,” *American Economic Journal: Economic Policy*, 2011, 3 (4), 148–71.
- Pallais, Amanda**, “Inefficient hiring in entry-level labor markets,” *American Economic Review*, nov 2014, 104 (11), 3565–3599.
- Peng, Sida, Eirini Kalliamvakou, Peter Cihon, and Mert Demirer**, “The impact of ai on developer productivity: Evidence from github copilot,” *arXiv preprint arXiv:2302.06590*, 2023.
- Rogerson, Richard, Robert Shimer, and Randall Wright**, “Search-theoretic models of the labor market: A survey,” *Journal of economic literature*, 2005, 43 (4), 959–988.
- Sajjadiani, Sima, Aaron J Sojourner, John D Kammeyer-Mueller, and Elton Mykorezi**, “Using machine learning to translate applicant work history into predictors of performance and turnover,” *Journal of Applied Psychology*, 2019, 104 (10), 1207.
- Singh, Jyoti Prakash, Seda Irani, Nripendra P Rana, Yogesh K Dwivedi, Sunil Saumya, and Pradeep Kumar Roy**, “Predicting the “helpfulness” of online consumer reviews,” *Journal of Business Research*, 2017, 70, 346–355.
- Smith, Morgan**, “ChatGPT can help you write a standout CV in seconds, job experts say: It’s ‘the ultimate resume-writing cheat code’,” 2023. Accessed: 2024-02-22.
- Stanton, Christopher T and Catherine Thomas**, “Landing the first job: The value of intermediaries in online hiring,” *The Review of Economic Studies*, 2016, 83 (2), 810–854.

Sterkens, Philippe, Ralf Caers, Marijke De Couck, Michael Geamanu, Victor Van Driessche, and Stijn Baert, “Costly mistakes: why and when spelling errors in resumes jeopardise interview chances,” *Working paper*, 2021.

Tafti, Elena Ashtari, *Technology, skills, and performance: the case of robots in surgery*, Institute for Fiscal Studies, 2022.

Tambe, Prasanna, Peter Cappelli, and Valery Yakubovich, “Artificial intelligence in human resources management: Challenges and a path forward,” *California Management Review*, 2019, 61 (4), 15–42.

Weiss, Daphne, Sunny X. Liu, Hannah Mieczkowski, and Jeffrey T. Hancock, “Effects of using Artificial Intelligence on interpersonal perceptions of job applicants,” *Cyberpsychology, Behavior, and Social Networking*, 2022, 25 (3), 163–168.

Wiles, Emma and John Horton, “More, but Worse: The Impact of AI Writing Assistance on the Supply and Quality of Job Posts,” March 2024.

—, **Zanele Munyikwa, and John Horton**, “Algorithmic Writing Assistance on Jobseekers’ Resumes Increases Hires,” January 2023, (30886).

Zhao, Yingyan, “Your (country’s) reputation precedes you: Information asymmetry, externalities and the quality of exports,” *Unpublished Paper, George Washington University*, 2018.