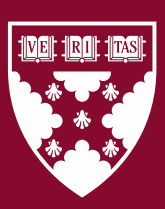
Working Paper 24-013

Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality

Fabrizio Dell'Acqua
Edward McFowland III
Ethan Mollick
Hila Lifshitz-Assaf
Katherine C. Kellogg
Saran Rajendran
Lisa Krayer
François Candelon
Karim R. Lakhani



Harvard Business School

Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality

Fabrizio Dell'Acqua
Harvard Business School

Edward McFowland III
Harvard Business School

Ethan Mollick
The Wharton School

Hila Lifshitz-Assaf
Warwick Business School

Katherine C. Kellogg
MIT Sloan School of Management

Saran Rajendran
Boston Consulting Group

Lisa Krayer
Boston Consulting Group

François Candelon
Boston Consulting Group

Karim R. Lakhani
Harvard Business School

Working Paper 24-013

We thank Michael Bervell, John Cheng, Pallavi Deshpande, Maxim Ledovskiy, John Kalil, Kelly Kung, Rick Lacerda, MarcAntonio Awada, Paula Marin Sariego, Rafael Noriega, Alejandro Ortega, Rahul Phanse, Quoc-Anh Nguyen, Nitya Rajgopal, Ogbemi Rewane, Kyle Schirmann, Andrew Seo, Tanay Tiwari, Elliot Tobin, Lebo Nthoiwa, Patrick Healy, Saud Almutairi, Steven Randazzo, Anahita Sahu, Aaron Zheng, and Yogesh Kumaar for helpful research assistance. For helpful feedback, we thank Maxime Courtaux, Clement Dumas, Gaurav Jha, Jesse Li, Max Männig, Michael Menietti, Rachel Mural, Zahra Rasouli, Esther Yoon, Leonid Zhukov, and David Zuluaga Martínez. Lakhani would like to thank Martha Wells, Anne Leckie, Iain Banks, and Alastair Reynolds for inspiring AI futures. We used Poe, Claude, and ChatGPT for light copyediting and graphics creations. Lakhani is an advisor to Boston Consulting Group on AI Strategy and learning engagement.

Funding for this research was provided in part by Harvard Business School.

Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality*

Fabrizio Dell'Acqua[1], Edward McFowland III[1], Ethan Mollick[2], Hila Lifshitz-Assaf[1,3], Katherine C. Kellogg[4], Saran Rajendran[5], Lisa Krayer[5], François Candelon[5], and Karim R. Lakhani[1]

[1]Digital Data Design Institute, Harvard Business School; [2]The Wharton School, University of Pennsylvania; [3]Warwick Business School, Artificial Intelligence Innovation Network; [4]MIT Sloan School of Management; [5]Boston Consulting Group, Henderson Institute

September 15, 2023

*Fabrizio Dell'Acqua (fdellacqua@hbs.edu), Edward McFowland III (emcfowland@hbs.edu), Ethan Mollick (emollick@wharton.upenn.edu), Hila Lifshitz-Assaf (hila.lifshitz-assaf@wbs.ac.uk), Katherine C. Kellogg (kkellogg@mit.edu), Saran Rajendran (rajendran.saran@bcg.com), Lisa Krayer (krayer.lisa@bcg.com), François Candelon (candelon.francois@bcg.com), Karim R. Lakhani (klakhani@hbs.edu). We thank Michael Bervell, John Cheng, Pallavi Deshpande, Maxim Ledovskiy, John Kalil, Kelly Kung, Rick Lacerda, MarcAntonio Awada, Paula Marin Sariego, Rafael Noriega, Alejandro Ortega, Rahul Phanse, Quoc-Anh Nguyen, Nitya Rajgopal, Ogbemi Rewane, Kyle Schirmann, Andrew Seo, Tanay Tiwari, Elliot Tobin, Lebo Nthoiwa, Patrick Healy, Saud Almutairi, Steven Randazzo, Anahita Sahu, Aaron Zheng, and Yogesh Kumaar for helpful research assistance. For helpful feedback, we thank Maxime Courtaux, Clement Dumas, Gaurav Jha, Jesse Li, Max Männig, Michael Menietti, Rachel Mural, Zahra Rasouli, Esther Yoon, Leonid Zhukov, and David Zuluaga Martínez. Lakhani would like to thank Martha Wells, Anne Leckie, Iain Banks, and Alastair Reynolds for inspiring AI futures. We used Poe, Claude, and ChatGPT for light copyediting and graphics creations. Lakhani is an advisor to Boston Consulting Group on AI Strategy and learning engagement. All errors are our own.

Abstract

The public release of Large Language Models (LLMs) has sparked tremendous interest in how humans will use Artificial Intelligence (AI) to accomplish a variety of tasks. In our study conducted with Boston Consulting Group, a global management consulting firm, we examine the performance implications of AI on realistic, complex, and knowledge-intensive tasks. The pre-registered experiment involved 758 consultants comprising about 7% of the individual contributor-level consultants at the company. After establishing a performance baseline on a similar task, subjects were randomly assigned to one of three conditions: no AI access, GPT-4 AI access, or GPT-4 AI access with a prompt engineering overview. We suggest that the capabilities of AI create a "jagged technological frontier" where some tasks are easily done by AI, while others, though seemingly similar in difficulty level, are outside the current capability of AI. For each one of a set of 18 realistic consulting tasks within the frontier of AI capabilities, consultants using AI were significantly more productive (they completed 12.2% more tasks on average, and completed tasks 25.1% more quickly), and produced significantly higher quality results (more than 40% higher quality compared to a control group). Consultants across the skills distribution benefited significantly from having AI augmentation, with those below the average performance threshold increasing by 43% and those above increasing by 17% compared to their own scores. For a task selected to be outside the frontier, however, consultants using AI were 19 percentage points less likely to produce correct solutions compared to those without AI. Further, our analysis shows the emergence of two distinctive patterns of successful AI use by humans along a spectrum of human-AI integration. One set of consultants acted as "Centaurs," like the mythical half-horse/half-human creature, dividing and delegating their solution-creation activities to the AI or to themselves. Another set of consultants acted more like "Cyborgs," completely integrating their task flow with the AI and continually interacting with the technology.

1 Introduction

The capabilities of Artificial Intelligence to produce human-like work have improved rapidly, especially since the release of OpenAI's ChatGPT, one of several Large Language Models (LLMs) that are widely available for public use. As AI capabilities overlap more with those of humans, the integration of human work with AI poses new fundamental challenges and opportunities, in particular in knowledge-intensive domains. In this paper, we examine this issue using randomized controlled field experiments with highly skilled professional workers. Our results demonstrate that AI capabilities cover an expanding, but uneven, set of knowledge work we call a "jagged technological frontier." Within this growing frontier, AI can complement or even displace human work; outside of the frontier, AI output is inaccurate, less useful, and degrades human performance. However, because the capabilities of AI are rapidly evolving and poorly understood, it can be hard for professionals to grasp exactly what the boundary of this frontier might be. We find that professionals who skillfully navigate this frontier gain large productivity benefits when working with the AI, while AI can actually decrease performance when used for work outside of the frontier.

Though LLMs are new, the impact of other, earlier forms of AI have been the subject of considerable scholarly discussion (e.g., Brynjolfsson et al. (2018); Furman and Seamans (2019); Puranam (2021)). Because of the limitations of these earlier forms of AI, nonroutine tasks that were difficult to codify seemed protected from automation (Autor et al., 2003; Acemoglu and Restrepo, 2019), especially as previous waves of technology had mostly automated lower-skilled occupations (Goldin and Katz, 1998). The release of ChatGPT in November, 2022 changed both the nature and urgency of the discussion. LLMs proved unexpectedly capable at creative, analytical, and writing tasks, including scoring at top levels at graduate and professional examinations (Girotra et al., 2023; Geerling et al., 2023; Kung et al., 2023; Boussioux et al., 2023). This represented an entirely new category of automation, one whose abilities overlapped with the most creative, most educated, and most highly paid workers (Eloundou et al., 2023).

Studies of previous generations of AI (Brynjolfsson et al., 2023) and controlled

experiments on the impact of recently released LLMs (e.g., Noy and Zhang (2023); Choi and Schwarcz (2023)) suggest that these systems can have a large impact on worker performance. In our study, we focus on complex tasks, selected by industry experts to replicate real-world workflows as experienced by knowledge workers. Most knowledge work includes this sort of flow, a set of interdependent tasks, some of which may be good fit for current AI, while some are not. We examine both kinds of tasks, and build on recent studies to suggest ways of understanding the rapidly evolving impact of AI on knowledge workers, under which circumstances organizations may benefit, and how this might change as the technology advances.

This is important because understanding the implications of LLMs for the work of organizations and individuals has taken on urgency among scholars, workers, companies, and even governments (Agrawal et al., 2018; Iansiti and Lakhani, 2020; Berg et al., 2023). Previous forms of AI led to considerable debate in the literature around how and whether professionals should adopt AI for knowledge work (Anthony et al., 2023) and the potential impact this might have on organizations (Raisch and Krakowski, 2021; Glaeser et al., 2021; Brynjolfsson et al., 2021). Some scholars focused on the potential for AI to help professionals improve their effectiveness and efficiency (DeStefano et al., 2022; Balakrishnan et al., 2022; Valentine and Hinds, 2023). Others demonstrated that, for critical tasks, it can be risky for professionals to use AI (Lebovitz et al., 2021), especially black-boxed (e.g., Lebovitz et al. (2022); Waardenburg et al. (2022)), and showed how professionals are struggling to use it effectively (Pachidi et al., 2021; Van den Broek et al., 2021). Finally, another group of researchers argued that the "algorithmic management" afforded by AI can create negative personal impacts for professionals (Kellogg et al., 2020; Möhlmann et al., 2021; Tong et al., 2021) and raise accountability and ethical questions (Choudhury et al., 2020; Cowgill et al., 2020; Rahman et al., 2024). Yet, most of the studies predate ChatGPT, and investigate forms of AI designed to produce discrete predictions based on past data. These systems are quite different from LLMs.

Specifically, outside of their technical differences from previous forms of machine learning, there are three aspects of LLMs that suggest they will have a much more rapid, and widespread, impact on work. The first is that LLMs have surprising

capabilities that they were not specifically created to have, and ones that are growing rapidly over time as model size and quality improve. Trained as general models, LLMs nonetheless demonstrate specialist knowledge and abilities as part of their training process and during normal use (Singhal et al., 2022; Boiko et al., 2023). While considerable debate remains on the concept of emergent capabilities from a technological perspective (Schaeffer et al., 2023), the effective capabilities of AIs are novel and unexpected, widely applicable, and are increasing greatly in short time spans. Recent work has shown that AI performs at a high level in professional contexts ranging from medicine to law (Ali et al., 2023; Lee et al., 2023), and beats humans on many measures of innovation (Boussioux et al., 2023; Girotra et al., 2023). And, while score performance on various standardized academic tests is an imperfect measure of LLM capabilities, it has been increasing substantially with each generation of AI models (OpenAI, 2023).

The general ability of LLMs to solve domain-specific problems leads to the second differentiating factor of LLMs compared to previous approaches to AI: their ability to directly increase the performance of workers who use these systems, without the need for substantial organizational or technological investment. Early studies of the new generation of LLMs suggest direct performance increases from using AI, especially for writing tasks (Noy and Zhang, 2023) and programming (Peng et al., 2023), as well as for ideation and creative work (Boussioux et al., 2023; Girotra et al., 2023). As a result, the effects of AI are expected to be higher on the most creative, highly paid, and highly educated workers (Eloundou et al., 2023; Felten et al., 2023)

The final relevant characteristic of LLMs is their relative opacity. This extends to the failure points of AI models, which include a tendency to produce incorrect, but plausible, results (hallucinations or confabulations), and to make other types of errors, including in math and when providing citations. The advantages of AI, while substantial, are similarly unclear to users. It performs well at some jobs, and fails in other circumstances in ways that are difficult to predict in advance. Contributing further to the opacity is that the best ways to use these AI systems are not provided by their developers and appear to be best learned via ongoing user trial-and-error and the sharing of experiences and heuristics via various online forums like user groups, hackathons, Twitter feeds and YouTube channels.

Taken together, these three factors – the surprising abilities of LLMs, their ability to do real work with virtually no technical skill required of the user, and their opacity and unclear failure points – suggest that the value and downsides of AI may be difficult for workers and organizations to grasp. Some unexpected tasks (like idea generation) are easy for AIs, while other tasks that seem to be easy for machines to do (like basic math) are challenges for some LLMs. This creates a "jagged Frontier," where tasks that appear to be of similar difficulty may either be performed better or worse by humans using AI. Due to the "jagged" nature of the frontier, the same knowledge workflow of tasks can have tasks on both sides of the frontier, see Figure 1. The future of understanding how AI impacts work involves understanding how human interaction with AI changes depending on where tasks are placed on this frontier, and how the frontier will change over time. Investigating how humans navigate this jagged frontier, and the subsequent performance implications, is the focus of our work.

We collaborated with a global management consulting firm (Boston Consulting Group - BCG) and advised them on designing, developing, and executing two pre-registered randomized experiments to assess the impact of AI on high human capital professionals. Subsequently, the author team received the data that the company collected for the purpose of this experiment and conducted the analysis presented in this paper. The study was structured in three phases: an initial demographic and psychological profiling, a main experimental phase involving multiple task completions, and a concluding interview segment. We tested two distinct tasks: one situated outside the frontier of AI capabilities and the other within its bounds. The experiment aimed to understand how AI integration might reshape the traditional workflows of these high human capital professionals.

Our results show that this generation of LLMs are highly capable of causing significant increases in quality and productivity, or even completely automating some tasks, but the actual tasks that AI can do are surprising and not immediately obvious to individuals or even to producers of LLMs themselves. Because this frontier is expanding and changing, the overall results suggest that AI will have a large impact on work, one which will increase with LLM capabilities, but where the impacts occur will be uneven.

2 Methods

We collected data from two randomized experiments to assess the causal impact of AI, specifically GPT-4 — the most capable of the AI models at the time of the experiments (Spring 2023) – on high human capital professionals working traditionally without AI.[1] We pre-registered our study, detailing the design structure, the experimental conditions, the dependent variables, and our main analytical approaches.[2] Our aim was to determine how introducing this AI into the tasks of highly-skilled knowledge workers might augment, disrupt, or influence their traditional workflows.

BCG individual contributor consultants around the world were offered the opportunity to spend 5 hours working on this experiment to evaluate the impact of AI on their activities. Approximately 7% of BCG's global individual contributor consultants' cohort engaged in and completed the experiment.

The experiment was structured into three distinct phases. Initially, consultants entered the study by completing a survey that captured their demographic and psychological profiles, as well as details about their role within the company. A few weeks after enrolling, participants received a link to complete the main experimental phase. This phase commenced with a pre-task survey, followed by the tasks detailed subsequently, and concluded with a post-task survey. In the final phase, participants were interviewed to share their experiences and perspectives on the role of AI in their profession.

In the first phase, we administered an enrollment survey to gather information about potential participants.[3] This survey captured details such as office location, internal affiliation, and tenure at BCG. Participants also completed psychological assessments, specifically providing insights into their Big 5 personality traits (Soto and John, 2017), innovativeness (Agarwal and Prasad, 1998), self-perceived creativity (Miron-Spektor et al., 2004), and paradox mindset (Miron-Spektor et al., 2018). Furthermore, the survey included a short section on their reading habits (including their familiarity with

[1] The project has received IRB approval, IRB23-0392.

[2] Pre-registration completed on Open Science Foundation, osf.io/ytaev. The pre-registration is available from the authors upon request and will be made publicly available after article acceptance or after the OSF embargo period has passed, whichever comes first

[3] Out of the 852 consultants who responded to the survey, 758 - about 89% - completed the experiment.

AI characters in fiction), and demographic details like gender, native language, and educational background. We utilized these data for stratified random assignment and as controls in our regression models, as described below.

The study encompassed 758 strategy consultants, each of whom completed the initial survey and experimental tasks. Each participant was assigned to one of two distinct experiments. Stratification of participants was based on multiple criteria, both between experiments and across experimental treatments. These criteria included gender, location, tenure at BCG, individual openness to innovation, and native English-speaking status. This information was collected with the survey administered during phase one, a few weeks before the main experiment.

In order to ensure genuine engagement and effort from participants, we incentivized their performance in the experiment. Participants who diligently participated in all aspects of the experiment were honored with an "office contribution" recognition, carrying financial implications related to their annual bonuses. Furthermore, to recognize and reward excellence, the top 20% of participants received additional recognition, and the top 5% was also awarded with a small gift. Executives at BCG reported that the recognition received by top participants was important because it was shared with the committee that oversees their career development and performance assessments.

Subjects were allocated to one of two distinct experiments, each involving a unique type of task, with no overlap between the groups. Both tasks were designed in collaboration with multiple people at BCG to represent some of the typical job activities encountered by individual contributor consultants. Approximately half of the participants (385 consultants) tackled a series of tasks where they were prompted to conceptualize and develop new product ideas, focusing on aspects such as creativity, analytical skills, persuasiveness and writing skills. The other half (373 consultants) engaged in business problem-solving tasks using quantitative data, customer and company interviews, and including a persuasive writing component. Both sets of tasks were developed to be realistic, and were designed with the input of professionals in the respective sectors. A senior level executive at the company commented on these tasks being “very much in line with part of the daily activities” of the subjects involved.

Notably, some forms of these tasks are also used by the company to screen job applicants, typically from elite academic backgrounds (including Ph.D.s), for their highly-coveted positions.

Both experiments followed a consistent structure. Initially, participants undertook a task without the aid of AI, establishing a baseline for performance and enabling within-subject analyses. Following this, participants were randomly assigned to one of three conditions to assess the influence of AI on their tasks, with these conditions being consistent across both experiments. The first group (a control condition) proceeded without AI support; the second ("GPT Only") had the assistance of an AI tool based on GPT-4; and the third ("GPT + Overview") not only utilized the same AI tool but also benefited from supplementary prompt engineering overview, which increased their familiarity with AI. These materials included instructional videos and documents that outlined and illustrated effective usage strategies.

Rather than relying on self-reported metrics or indirect indicators, we directly assessed participants' skills through a task that closely mirrored the main experiment. In both experiments, we employ an assessment task that, while different from the experimental task, is highly comparable, ensuring a precise evaluation of skills for this specific task type.[4] Our findings indicate that performance in the assessment task is a predictor of performance in the experimental task, allowing us to study the differential effects of introducing AI to participants of different skill levels.

Each task assigned to participants came with a specific time allocation. In the experiment using a task inside the frontier, the assessment task duration was set for 30 minutes, while the subsequent one was allotted 90 minutes. Conversely, in the outside-the-frontier experiment, both the first and second tasks were designated 60 minutes each, though participants could complete them earlier if they finished ahead of time. It is important to note that for the task inside the frontier, participants were required to remain on the task's page for the entire duration of the task, and could not complete the exercise earlier. This approach ensured that our analysis for the inside-the-frontier tasks focused

[4] Dell'Acqua et al. (2023) adopts a comparable experimental framework to evaluate subjects' competencies.

chiefly on the qualitative differences, rather than any timing improvements brought about by using AI. These timeframes were automatically enforced, with the experimental system advancing to the next question once the stipulated time for a task elapsed.

In every experimental task, subjects assigned to the AI conditions had access to a company platform. This platform, developed using the OpenAI API, facilitated an interactive experience with OpenAI's GPT-4, mirroring the dynamics of ChatGPT. It enabled the collection of all participants' prompts and AI's corresponding responses, providing a comprehensive view into the collaborative behaviors between subjects and AI. All subjects used the same version of the tool, accessing GPT-4 as available at the end of April, 2023, and using default system prompts and temperature.

Aside from the thematic differences, the tasks differed in another key way. While both were designed to be comparably complex and realistic, the first task was selected to be within the potential technological frontier of GPT-4. The second experiment was designed so that GPT-4 would make an error when conducting the analysis, ensuring the work fell just outside the frontier.

3 Results

3.1 Quality and Productivity Booster - Inside the Frontier

The inside-the-frontier experiment focused on creative product innovation and development. The initial assessment task asked participants to brainstorm innovative beverage concepts. From their set of ideas, they identified the most viable option and devised a comprehensive plan for its market debut. After this task, subjects moved to the main experimental phase and the context transitioned to the main experimental task.

In this experimental task, participants were tasked with conceptualizing a footwear idea for niche markets and delineating every step involved, from prototype description to market segmentation to entering the market. An executive from a leading global footwear company verified that the task design covered the entire process their company

typically goes through, from ideation to product launch.[5] Participants responded to a total of 18 tasks (or as many as they could within the given time frame). These tasks spanned various domains. Specifically, they can be categorized into four types: creativity (e.g., "Propose at least 10 ideas for a new shoe targeting an underserved market or sport."), analytical thinking (e.g., "Segment the footwear industry market based on users."), writing proficiency (e.g., "Draft a press release marketing copy for your product."), and persuasiveness (e.g., "Pen an inspirational memo to employees detailing why your product would outshine competitors."). This allowed us to collect comprehensive assessments of quality. All tasks and details are reported in Appendix A.

In the experiment, the primary outcome variable is the quality of the subjects' responses. To quantify this quality, we employed a set of human graders to evaluate each question that participants didn't leave unanswered.[6] Each response was evaluated by two human graders. We then calculated the mean grade assigned by humans to each question. This gave us 18 dependent variables (one per each question). We subsequently averaged these scores across all questions to derive a composite "Quality" score, which we use in our main analyses. As an additional assessment, we also utilized GPT-4, to independently score the subjects' responses. Similarly to the human grades, we produced a score for each one of the 18 questions, and then a composite "Quality (GPT)" score.

Figure 2 uses the composite human grader score and visually represents the performance distribution across the three experimental groups, with the average score plotted on the y-axis. A comparison of the dashed lines and the overall distributions of the experimental conditions clearly illustrates the significant performance enhancements associated with the use of GPT-4. Both AI conditions show clear superior performance to the control group not using GPT-4.

Table 1 presents the results of the analyses using response quality as the dependent variable and highlights the performance implications of using AI. Columns 1, 2, and 3 utilize human-generated grades as the dependent variable, while Column 4 uses the

[5]The executive noted the only step missing from this exercise was an evaluation of how the new product integrates with the company's existing product lines. As our experiment used a fictional company, we did not require participants to present their product suggestions in relation to existing ones.

[6]Graders were from BCG, or MBA students at a top program.

composite grades generated by GPT-4. Across all specifications, both treatments — GPT + Overview and GPT Only — demonstrate positive effects. In Column 1, GPT + Overview leads to a 1.75 increase in scores over the control mean of 4.1, which is a 42.5% increase; GPT Only led to a 1.56, or 38% increase. Notably, Columns 2, 3, and 4 incorporate performance metrics from the assessment task and the treatment coefficients they report remain very consistent. Column 4 uses GPT scores as the dependent variable, and shows coefficients of 1.34 for the GPT + Overview treatment and 1.21 for the GPT Only treatment over the control group, which are equal respectively to 18.6% and 16.8% increases in performance.[7]

The beneficial impacts of using AI remain consistent across all our specifications. We merged our AI treatments and used all our pre-registered quality variables as dependent variables. This included individual grades for each question as evaluated by humans, as well as grades evaluated by GPT-4, based on the three specifications outlined in Columns 1-3 of Table 1. This resulted in a comprehensive set of 108 regressions. All of these regressions showed a significant effect of introducing AI on consultants' performance. Figures 3 and 4 show 54 of these regressions each. Additionally, three dashed lines report the average effects of each regression. The mean effect size when comparing subjects using AI with a control with no GPT-4 access is 1.69 (a 46.6% increase over the control mean) when using human evaluations and 1.36 (20.2%) when using GPT-4 evaluations.

Another key observation from the table is the differential impact of the two AI treatments. Specifically, the GPT + Overview treatment consistently exhibits a more pronounced positive effect compared to the GPT Only treatment. The bottom of the table displays a p-value that tests whether the effects of receiving GPT + Overview were equivalent to those of being assigned to GPT Only, showing this value to be below or around the conventional 5% threshold in all specifications. This underscores the importance of the added overview in enhancing the efficacy of AI assistance. However, we should note that the overview increased "retainment" (i.e., copying and pasting the GPT-4 output), and retainment itself was associated with better performance.[8] The table

[7]These percentage improvements are relatively lower also because GPT-4 tends to be a more lenient grader and scores our control baseline higher.

[8]Appendix C provides further details.

also highlights various other factors, such as gender, native English proficiency, tenure, location, and tech openness, and their influence on the outcomes.[9]

Table 2 presents the results related to the percentage of task completion by subjects, which is the dependent variable in this analysis. Across Columns 1, 2, and 3, both treatments — GPT + Overview and GPT Only — demonstrate a positive effect on task completion. On average, these coefficients indicate a 12.2% increase in completion rates.[10] The control group completed on average 82% of their tasks, while the GPT + Overview condition completed about 93% and GPT Only about 91%. Column 2 incorporates the performance metric from the assessment, and Column 3 further extends the analysis by including the same set of controls as in Table 1. The coefficients suggest that the integration of AI tools enhances the rate of task completion very significantly, at the same time as it increases quality.

Figure 5 presents an important trend: the most significant beneficiaries of using AI are the bottom-half-skill subjects, consistent with findings from Noy and Zhang (2023) and Choi and Schwarcz (2023).[11] By segmenting subjects exposed to one of the two AI conditions into two distinct categories — top-half-skill performers (those ranking in the top 50% on the assessment task) and bottom-half-skill performers (those in the bottom 50%) — we observed performance enhancements in the experimental task for both groups when leveraging GPT-4. When comparing the two groups, though, we see the bottom-half-skill performers exhibited the most substantial surge in performance, 43%, compared to the top-half-skill subjects, 17%. Note that the top-half-skill performers also receive a significant boost, although not as much as the bottom-half-skill performers.

For the task inside the frontier, we did not allow any subjects to complete the task before the allotted time was over. Instead, their final question was an especially long

[9]We employ binary variables for all these factors. "Female" is set to 1 if a subject identifies as female and 0 otherwise. "English Native" is 1 if a subject claims native proficiency in English and 0 otherwise (nearly every subject indicates either Native or Advanced proficiency in English). "Low Tenure" is 1 if a subject has been with BCG for a year or less, and 0 otherwise. "Location" is 1 if a subject's office is located in Europe or the Middle East, and 0 otherwise. Lastly, "Tech Openness" is 1 if the subject expressed a higher receptivity to technology in their enrollment survey, and 0 otherwise.

[10]When directly comparing the two AI treatments at the bottom of the table, the difference in their impacts is not statistically significant.

[11]It is important to note that "higher-skill" and "lower-skill" here are relative. All these consultants would appear to be extremely high-skill in most other real-world contexts.

one asking them to "synthesize the insights you have gained from the previous questions and create an outline for a Harvard Business Review-style article of approximately 2,500 words." However, while participants were required to take the full time allotted to this task, we nevertheless tracked the amount of time that they took to reach this last question, having completed the first 17 questions. Table 3 uses this Timing variable as the dependent variable. The GPT + Overview treatment makes subjects faster by 1129 seconds (18.8 minutes or 22.5% faster than the control), while the GPT Only treatment reduces time spent on the first 17 questions by 1388 seconds (23.13 minutes or 27.63% faster than the control).

Exploring the variation in content generated by subjects, our focus was on understanding the diversity of their responses in relation to others. By employing Google's Universal Sentence Encoder (USE), we semantically analyzed the ideas presented by subjects in the first question of the inside-the-frontier experiment ("Generate ideas for a new shoe aimed at a specific market or sport that is underserved. Be creative, and give at least 10 ideas"). Our findings indicate that while subjects using AI produce ideas of higher quality, as discussed in the previous sections, there is a marked reduction in the variability of these ideas compared to those not using AI. This suggests that while GPT-4 aids in generating superior content, it might lead to more homogenized outputs. Figure 6 shows the distribution of similarity across experimental groups. For a detailed analysis, refer to Appendix D.

Our results reveal significant effects, underscoring the prowess of AI even in tasks traditionally executed by highly skilled and well-compensated professionals. Not only did the use of AI lead to an increase in the number of subtasks completed by an average of 12.5%, but it also enhanced the quality of the responses by an average of more than 40%. These effects support the view that for tasks that are clearly within its frontier of capabilities, even those that historically demanded intensive human interaction, AI support provides huge performance benefits.

3.2 Quality Disruptor - Outside the frontier

In refining the task within the frontier and recognizing the substantial quality and productivity gains from AI integration, we sought a task that AI couldn't easily complete through simple copying and pasting of our instructions as a prompt. We designed the task outside the frontier using as a starting point the type of business cases that BCG uses for its highly competitive job interviews. Our goal was to design a task where consultants would excel, but AI would struggle without extensive guidance. After several iterations, we settled on a task based on an existing business case that used data on a spreadsheet, as well as a file presenting interviews with company insiders, which were adjusted and adapted to the goals of this experiment. To be able to solve the task correctly, participants would have to look at the quantitative data using subtle but clear insights from the interviews. While the spreadsheet data alone was designed to seem to be comprehensive, a careful review of the interview notes revealed crucial details. When considered in totality, this information led to a contrasting conclusion to what would have been provided by AI when prompted with the exercise instructions, the given data, and the accompanying interviews.

In this second experiment, the primary objective was for subjects to offer actionable strategic recommendations to a hypothetical company. First, participants worked on the assessment task, where they had to analyze the company's channel performance. Using insights from interviews and financial data, participants were asked to provide information and informed advice to the CEO. Their recommendations were to pinpoint which channel held the most potential for growth.

As subjects completed their assessment task, they moved to the main experimental task. The focus transitioned from the examination of the company's distribution channels to brand analysis, as subjects had to analyze the company's brand performance. Similarly to the assessment task, participants used insights from interviews and financial data to provide recommendations for the CEO. Their recommendations were to pinpoint which brand held the most potential for growth. Additionally, participants were also expected to suggest actions to improve the chosen brand, regardless of the exact brand they had

chosen. Details of these tasks are reported in Appendix A.

For this task outside the frontier, our primary metric of evaluation is 'correctness.' This is represented as a binary variable, where a value of '1' indicates that subjects provided the accurate recommendation, and '0' signifies otherwise. Figure 7 visually presents the correctness percentages across different groups, highlighting a noticeable dip in performance among the AI treatment groups when juxtaposed with the control group. Subjects in the control group were correct about this exercise about 84.5% of the time, while the AI conditions scored at 60% and 70% (for an average decrease of 19 percentage points when combining the AI treatment conditions and comparing them to the control condition).

Table 4 delves into the impact of the AI treatments on the correctness of tasks in the outside-the-frontier experiment using linear regressions with correctness as a binary dependent variable. Both AI treatments— GPT + Overview and GPT Only — show a significant negative impact, with the GPT + Overview group experiencing a more pronounced decrease (24 percentage points versus 13 percentage points). Column 2 introduces the performance metric from the assessment, while Column 3 further refines the analysis by incorporating the same set of controls as in Tables 1 and 2. When directly comparing the two treatments, the difference in their impacts is statistically significant at the 10% threshold across specifications.

Table 5 examines the influence of the AI treatments on the time taken by participants to complete tasks in the outside-the-frontier experiment. The dependent variable here is "Timing," representing the duration subjects spent on the task calculated in seconds. Column 2 further refines the analysis by incorporating the same set of controls as in Table 1. Both AI treatments — GPT + Overview and GPT Only — indicate a reduction in the time spent, more than 11 minutes for GPT + Overview (a 30% decrease in timing over the control mean), and more than 6 minutes for GPT Only (a decrease of 18%) when compared to the control. When we compare the two coefficients, we find that GPT + Overview shows a more substantial decrease in time, and one that is statistically significant, compared to the GPT Only group.

Table 6 further examines the quality of recommendations provided by subjects in

the outside-the-frontier experiment. The dependent variable, 'Recommendation Quality,' measures the quality of the recommendations given. These were scored (on a 1-10 scale) by two different sets of human graders (one of BCG consultants not involved in the experiment, and the other of business school students with grading experience) following a simple rubric, developed by the authors relying on their experience in consulting (see Appendix B for the full rubric). Looking at Column 1, the treatment GPT + Overview leads to a 1.47 greater score (25.1% increase over the control mean), while GPT Only increases the score by 1.05 (17.9% over the control mean). Across all specifications, subjects using AI (both GPT + Overview and GPT Only) consistently outperformed those not using AI in terms of recommendation quality, regardless of the correctness of their answer. When we control for various factors (as in Table 1) in columns 2, 3, and 4, the positive impact of AI remains robust. This result holds true regardless of the correctness of the subjects' answers, as evidenced in columns 3 and 4. Column 3 reports the same regression in column 2 only for the subset of participants whose answers were incorrect, while column 4 does the same for the subject of participants whose answers were correct. In both instances, the effects of using AI are positive.

Figure 8 shows the quality of recommendations for people who were wrong in their answers and for those who were right (Figure 8). Both distributions show a clear shift for the experimental groups using AI. That is, the quality of the recommendation about a business problem-solving case increased, regardless of whether the underlying recommendation was correct or not. This finding underscores the multifaceted ways in which AI can influence performance in a workflow among highly skilled professionals.

3.3 Navigating the frontier

The experiments show that the shape and position of the frontier are vital to understanding the impact of AI on work. On tasks within the frontier, AI significantly improved human performance. Outside of it, humans relied too much on the AI and were more likely to make mistakes. Not all users navigated the jagged frontier with equal adeptness. While some completed their task incorrectly, others showcased a remarkable ability to harness the power of AI effectively. We conducted further analyses of the

strategies adopted by those who managed to perform correctly in the outside-the-frontier experimental task while using LLMs. Understanding the characteristics and behaviors of these participants may prove important as organizations think about ways to identify and develop talent for effective collaboration with AI tools.

We identified two predominant models that encapsulate their approach.

The first is Centaur behavior. Named after the mythical creature that is half-human and half-horse, this approach involves a similar strategic division of labor between humans and machines closely fused together.[12] Users with this strategy switch between AI and human tasks, allocating responsibilities based on the strengths and capabilities of each entity. They discern which tasks are best suited for human intervention and which can be efficiently managed by AI.

The second model we observed is Cyborg behavior.[13] Named after hybrid human-machine beings as envisioned in science fiction literature, this approach is about intricate integration. Cyborg users don't just delegate tasks; they intertwine their efforts with AI at the very frontier of capabilities. This strategy might manifest as alternating responsibilities at the subtask level, such as initiating a sentence for the AI to complete or working in tandem with the AI.

We are continuing to examine these behaviors using the experimental task's GPT logs and rich qualitative data gathered from the experiments in order to create an emerging understanding of these behaviors. We are also investigating the impact of these behaviors on performance. Appendix E further explores these behaviors.

4 Discussion

In this paper, we study the integration of AI with a jagged capability frontier into contemporary human real-world, high-end knowledge work tasks. We use a randomized field experiment to illuminate the dual role of AI as both a booster, enhancing efficiency and productivity, and a disruptor, negatively impacting performance in tasks outside its

[12]This term was prominently used by chess champion Garry Kasparov to describe how humans and AI might collaborate in chess.

[13]Clynes and Kline (1960) first proposed this term.

frontier. Our findings underscore the transformative potential of AI and offer insights into harnessing its capabilities for optimal outcomes. A crucial feature of our experiment was the availability of our experimental subjects. Specifically, we tapped into a high human capital population, with participants who were not only highly skilled but also engaged in tasks that closely mirrored part of their professional activities.

We found that the utility of AI can fluctuate over the course of a professional's workflow, with some tasks falling inside while others fall outside of the frontier. For tasks inside the frontier, these findings carry large performance implications. Across 18 realistic business tasks, AI significantly increased performance and quality for every model specification, increasing speed by more than 25%, performance as rated by humans by more than 40%, and task completion by more than 12%. Further, it operated in a way that benefitted bottom-half performers the most, though all users benefitted from AI. Thus, AI seems to both level performance differences across levels of ability and raise the quality of work for inside-the-frontier tasks. These findings suggest a need to understand how work can be organized to better integrate AI. Tasks outside the AI frontier also present opportunities for individuals to operate as either cyborgs or centaurs. Cyborgs integrate AI and human capabilities at a granular, sub-task level, while centaurs strategically delegate between human and AI sub-tasks. In our sample, we observed these two behaviors (see Appendix E). It is clear that the best approaches to using AI are not fully understood and need to be deeply examined by scholars and practitioners.

It was only when we identified tasks outside the frontier that we saw performance decreases as a result of AI. On those tasks, this study highlights the importance of validating and interrogating AI (Lebovitz et al., 2022) and of continuing to exert cognitive effort and experts' judgment when working with AI (Dell'Acqua, 2022). Professionals who had a negative performance when using AI tended to blindly adopt its output and interrogate it less ("unengaged interaction with AI" Lebovitz et al. (2022)). Appendix B presents a "retainment" score measuring how common blind adoption was among our subjects. These findings raise questions regarding when and how to know whether to trust LLMs (Glikson and Woolley, 2020) in particular as AI tools impact employees' performance and their subsequent evaluation (Teodorescu et al., 2021; Gkeredakis, 2022).

For AI designers and companies, these findings offer an important path for designing AI tools and for building navigation capabilities of the users (Lebovitz et al., 2021; Valentine and Hinds, 2023).

More broadly, these findings raise questions regarding using AI for high-risk tasks and responsible AI, a topic that is highly debated by AI policymakers and academics (Waardenburg et al., 2022; Rahman et al., 2024; *AI Act*, 2023). An immediate danger emerging from these findings, for instance, is that people will stop delegating work inside the frontier to junior workers, creating long-term training deficits (Beane, 2019).[14] Navigating the frontier requires expertise, which will need to be built through formal education (Osterman, 2011; Myers and Kellogg, 2022), on-the-job training (Kellogg et al., 2021), and employee-driven upskilling (Leonardi and Neeley, 2022; Iansiti and Nadella, 2022).

In this study, we focus on the individual level of analysis but, as we had such a large number of professionals from within a single organization (BCG), this study clearly offers organization-level implications as well. First, our findings bring in a nuanced response to the ongoing debate on whether organizations should adopt AI for high-end knowledge work. The results suggest that this debate should move beyond the dichotomous decision of adopting or not AI and instead focus on the knowledge workflow and the tasks within it, and in each of them, evaluate the value of using different configurations and combinations of humans and AI. This will require rethinking collaboration between humans and AI (Anthony et al., 2023; Faraj and Leonardi, 2022; Feuerriegel et al., 2022), how new roles will emerge and be created (Barrett et al., 2012; Allen and Choudhury, 2022; Kellogg, 2022; Sergeeva et al., 2020), how adoption will be shaped by stakeholders in the organization and beyond (Sendak et al., 2020; Hillebrand et al., 2023), new capabilities and strategies (Iansiti and Lakhani, 2020) and new forms of organizing (Bailey et al., 2022; Beane and Leonardi, 2022). Moreover, the significant impact on creativity suggests AI should play a role in the future of organizing for innovation (Kittur et al., 2019; Amabile, 2020; Raj et al., 2023; Lifshitz-Assaf and Lazar, 2023; Boussioux et al., 2023).

The potential for diminished diversity of ideas stemming from AI usage could

[14] AI could also play a role in supporting human training (Gaessler and Piezunka, 2023).

pose challenges for organizations (e.g., Page (2019) (see appendix D)). As companies increasingly integrate AI into their operations, they must consider whether employing a variety of AI models, possibly multiple LLMs, or increased human-only involvement, could counteract this homogenization (with the associated tradeoffs). This underscores the significance of maintaining a diverse AI ecosystem. Moreover, the optimal AI strategy might vary based on a company's production function. While some organizations might prioritize consistently high average outputs, others might value maximum exploration and innovation. Finally, there may be some general equilibrium considerations to incorporate. Outputs like grants or articles might indeed be of superior quality when AI-assisted. However, in a competitive landscape where many are leveraging AI, outputs generated without AI assistance might stand out and achieve notable success due to their distinctiveness. Clearly, the interplay between the quality and variability of ideas is intricate, and further investigation is essential to unpack the nuances and implications of these results.

Finally, we note that our findings offer multiple avenues for interpretation when considering the future implications of human/AI collaboration. Firstly, our results lend support to the optimism about AI capabilities for important high-end knowledge work tasks such as fast idea generation, writing, persuasion, strategic analysis, and creative product innovation. In our study, since AI proved surprisingly capable, it was difficult to design a task in this experiment outside the AI's frontier where humans with high human capital doing their job would consistently outperform AI. However, navigating AI's jagged capabilities frontier remains challenging. Even for experienced professionals engaged in tasks akin to some of their daily responsibilities, this demarcation is not always evident. As the boundaries of AI capabilities continue to expand, often exponentially, it becomes incumbent upon human professionals to recalibrate their understanding of the frontier and for organizations to prepare for a new world of work combining humans and AI. Overall, AI seems poised to significantly impact human cognition and problem-solving ability. Similarly to how the internet and web browsers dramatically reduced the marginal cost of information sharing, AI may also be lowering the costs associated with human thinking and reasoning, with potentially

broad and transformative effects.

References

Acemoglu, Daron and Pascual Restrepo, "Automation and new tasks: How technology displaces and reinstates labor," *Journal of Economic Perspectives*, 2019, *33* (2), 3–30.

Agarwal, Ritu and Jayesh 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, Joshua Gans, and Avi Goldfarb, *Prediction machines: the simple economics of artificial intelligence*, Harvard Business Press, 2018.
AI Act

AI Act, *Technical Report PE 698.792, European Parliament, Brussels June 2023.*

Ali, R., O. Y. Tang, I. D. Connolly, J. S. Fridley, J. H. Shin, P. L. Z. Sullivan, ..., and W. F. Asaad, *"Performance of ChatGPT, GPT-4, and Google bard on a neurosurgery oral boards preparation question bank," Neurosurgery, 2023, pp. 10–1227.*

Allen, Ryan and Prithwiraj Choudhury, *"Algorithm-augmented work and domain experience: The countervailing forces of ability and aversion," Organization Science, 2022, 33 (1), 149–169.*

Amabile, Teresa M., *"Creativity, Artificial Intelligence, and a World of Surprises," Academy of Management Discoveries, September 2020, 6 (3), 351–354.*

Anthony, C., B. A. Bechky, and A. L. Fayard, *""Collaborating" with AI: Taking a system view to explore the future of work," Organization Science, 2023.*

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

Bailey, D. E., S. Faraj, P. J. Hinds, P. M. Leonardi, and G. von Krogh, *"We are all theorists of technology now: A relational perspective on emerging technology and organizing," Organization Science, 2022, 33 (1), 1–18.*

Balakrishnan, M., K. Ferreira, and J. Tong, *"Improving Human-Algorithm Collaboration: Causes and Mitigation of Over- and Under-Adherence," Working Paper December 2022.*

Barrett, M., E. Oborn, W. J. Orlikowski, and J. Yates, *"Reconfiguring boundary relations: Robotic innovations in pharmacy work," Organization Science, 2012, 23 (5), 1448–1466.*

Beane, M. I. and P. M. Leonardi, *"Pace layering as a metaphor for organizing in the age of intelligent technologies: Considering the future of work by theorizing the future of organizing," Journal of Management Studies, 2022.*

Beane, Matthew, *"Shadow learning: Building robotic surgical skill when approved means fail," Administrative Science Quarterly, 2019, 64 (1), 87–123.*

Berg, Justin, Manav Raj, and Robert Seamans, *"Capturing Value from Artificial Intelligence," Academy of Management Discoveries ja, 2023.*

Boiko, D. A., R. MacKnight, and G. Gomes, *"Emergent autonomous scientific research capabilities of large language models," arXiv preprint arXiv:2304.05332, 2023.*

Boussioux, Leonard, Jacqueline N Lane, Miaomiao Zhang, Vladimir Jacimovic, and Karim R. Lakhani, *"The Crowdless Future? How Generative AI Is Shaping the Future of Human Crowdsourcing," The Crowdless Future, 2023.*

Brynjolfsson, Erik, Danielle Li, and Lindsey R. Raymond, *"Generative AI at work," Working Paper w31161, National Bureau of Economic Research 2023.*

_ , Tom Mitchell, and Daniel Rock, *"What can machines learn and what does it mean for occupations and the economy?," in "AEA papers and proceedings," Vol. 108 American Economic Association 2018, pp. 43–47.*

_ , Wang Jin, and Kristina McElheran, *"The power of prediction: predictive analytics, workplace complements, and business performance," Business Economics, 2021, 56, 217–239.*

Choi, Jonathan H. and Daniel Schwarcz, *"AI Assistance in Legal Analysis: An Empirical Study," 2023. Available at SSRN: https://ssrn.com/abstract=4539836.*

Choudhury, Prithwiraj, Evan Starr, and Rajshree Agarwal, *"Machine learning and human capital complementarities: Experimental evidence on bias mitigation," Strategic Management Journal, 2020, 41 (8), 1381–1411.*

Clynes, Manfred E. and Nathan S. Kline, *"Cyborgs and space," Astronautics, 1960, 14 (9), 26–27.*

Cowgill, Bo, Fabrizio Dell'Acqua, Samuel Deng, Daniel Hsu, Nakul Verma, and Augustin Chaintreau, *"Biased Programmers? Or Biased Data? A Field Experiment in Operationalizing AI Ethics," Proceedings of the 21st ACM Conference on Economics and Computation, 2020, pp. 679–681.*

Dell'Acqua, Fabrizio, Bruce Kogut, and Patryk Perkowski, *"Super Mario Meets AI: The Effects of Automation on Team Performance and Coordination in a Videogame Experiment," The Review of Economics and Statistics, 2023.*

Dell'Acqua, Fabrizio, *"Falling asleep at the wheel: Human/AI Collaboration in a Field Experiment on HR Recruiters," 2022.*

DeStefano, Timothy, Katherine Kellogg, Michael Menietti, and Luca Vendraminelli, *"Why Providing Humans with Interpretable Algorithms May, Counterintuitively, Lead to Lower Decision-making Performance," MIT Sloan Research Paper No. 6797, 2022. Available at SSRN: https://ssrn.com/abstract=4246077 or http://dx.doi.org/10.2139/ssrn.4246077.*

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.*

Faraj, Samer and Paul M. Leonardi, *"Strategic organization in the digital age: Rethinking the concept of technology," Strategic Organization, 2022, 20 (4), 771–785.*

Felten, Edward W., Manav Raj, and Robert Seamans, *"Occupational heterogeneity in exposure to generative ai," 2023. Available at SSRN: https://ssrn.com/abstract=4414065.*

Feuerriegel, S., Y. R. Shrestha, G. von Krogh, and C. Zhang, *"Bringing artificial intelligence to business management," Nature Machine Intelligence, 2022, 4 (7), 611–613.*

Furman, Jason and Robert Seamans, *"AI and the Economy," Innovation policy and the economy, 2019, 19 (1), 161–191.*

Gaessler, Fabian and Henning Piezunka, *"Training with AI: Evidence from chess computers," Strategic Management Journal, 2023.*

Geerling, Wayne, G. Dirk Mateer, Jadrian Wooten, and Nikhil Damodaran, *"ChatGPT has aced the test of understanding in college economics: Now what?," The American Economist, 2023, p. 05694345231169654.*

Girotra, K., L. Meincke, C. Terwiesch, and K. T. Ulrich, *"Ideas are dimes a dozen: Large language models for idea generation in innovation," 2023. Available at SSRN: https://ssrn.com/abstract=4526071.*

Gkeredakis, Manos, *"Fair Algorithms in Organizations: A Performative-Sensemaking Model," in "ICIS 2022 Proceedings" 2022, p. 1.*

Glaeser, Edward, Andrew Hillis, Hyunjin Kim, Scott Duke Kominers, and Michael Luca, *"Decision Authority and the Returns to Algorithms," 2021.*

Glikson, Ella and Anita Williams Woolley, *"Human trust in artificial intelligence: Review of empirical research," Academy of Management Annals, 2020, 14 (2), 627–660.*

Goldin, Claudia and Lawrence F. Katz, *"The origins of technology-skill complementarity," Quarterly Journal of Economics, 1998, 3 (4), 693–732.*

Hillebrand, L, S Raisch, and J Schad, *"Artificial Intelligence in Management: An Integrative Review and Research Agenda," Working Paper 2023.*

Iansiti, Marco and Karim R. Lakhani, *Competing in the age of AI: Strategy and leadership when algorithms and networks run the world, Harvard Business Press, 2020.*

_ and Satya Nadella, *"Democratizing Transformation," Harvard Business Review, 2022, 100 (5-6), 42–49.*

Kellogg, K. C., *"Local adaptation without work intensification: experimentalist governance of digital technology for mutually beneficial role reconfiguration in organizations," Organization Science, 2022, 33 (2), 571–599.*

_ , J. E. Myers, L. Gainer, and S. J. Singer, *"Moving violations: Pairing an illegitimate learning hierarchy with trainee status mobility for acquiring new skills when traditional expertise erodes," Organization Science, 2021, 32 (1), 181–209.*

_ , M. A. Valentine, and A. Christin, *"Algorithms at work: The new contested terrain of control," Academy of Management Annals, 2020, 14 (1), 366–410.*

Kittur, A., L. Yu, T. Hope, J. Chan, H. Lifshitz-Assaf, K. Gilon, F. Ng, R.E. Kraut, and D. Shahaf, *"Scaling Up Analogical Innovation with Crowds and AI," Proceedings of the National Academy of Sciences, 2019, 116 (6), 1870–1877.*

Kung, Tiffany H., Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga et al., *"Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models," PLoS digital health, 2023, 2 (2), e0000198.*

Lebovitz, Sarah, Hila Lifshitz-Assaf, and Natalia Levina, *"To engage or not to engage with AI for critical judgments: How professionals deal with opacity when using AI for medical diagnosis," Organization Science, 2022, 33 (1), 126–148.*

_ , Natalia Levina, and Hila Lifshitz-Assaf, *"Is AI ground truth really true? The dangers of training and evaluating AI tools based on experts' know-what," MIS quarterly, 2021, 45 (3).*

Lee, P., S. Bubeck, and J. Petro, *"Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine," New England Journal of Medicine, 2023, 388 (13), 1233–1239.*

Leonardi, P. M. and T. Neeley, *The digital mindset: What it really takes to thrive in the age of data, algorithms, and AI 2022.*

Lifshitz-Assaf, Hila and Moran Lazar, *"Would Archimedes Shout "Eureka" If He Had Google? Innovating with Search Algorithms," September 2023. Available online 10 September 2023.*

Miron-Spektor, Ella, Amy Ingram, Joshua Keller, Wendy K. Smith, and Marianne W. Lewis, *"Microfoundations of organizational paradox: The problem is how we think about the problem," Academy of management journal, 2018, 61 (1), 26–45.*

_ , Miriam Erez, and Eitan 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.*

Myers, J. E. and K. C. Kellogg, *"State actor orchestration for achieving workforce development at scale: Evidence from four US states," ILR Review, 2022, 75 (1), 28–55.*

Möhlmann, M., L. Zalmanson, O. Henfridsson, and R. W. Gregory, *"Algorithmic Management of Work on Online Labor Platforms: When Matching Meets Control," MIS quarterly, 2021, 45 (4).*

Noy, Shakked and Whitney Zhang, *"Experimental evidence on the productivity effects of generative artificial intelligence," 2023. Available at SSRN: https://ssrn.com/abstract=4375283.*

OpenAI, *"GPT-4 Technical Report," ArXiv, 2023. Published 15 March 2023, Computer Science.*

Osterman, P., *The promise, performance, and policies of community colleges. Reinventing higher education: The promise of innovation: 129–158., Cambridge, MA: Harvard Education Press, 2011.*

Pachidi, S., H. Berends, S. Faraj, and M. Huysman, *"Make way for the algorithms: Symbolic actions and change in a regime of knowing," Organization Science, 2021, 32 (1), 18–41.*

Page, Scott E., *The diversity bonus: How great teams pay off in the knowledge economy, Princeton University Press, 2019.*

Peng, S., E. Kalliamvakou, P. Cihon, and M. Demirer, *"The impact of ai on developer productivity: Evidence from github copilot," arXiv preprint arXiv:2302.06590, 2023.*

Puranam, Phanish, *"Human–AI collaborative decision-making as an organization design problem," Journal of Organization Design, 2021, pp. 1–6.*

Rahman, H., A. Karunakaran, and L. Cameron, *"Taming Platform Power: Taking Accountability into Account in the Management of Platforms," Academy of Management Annals, 2024. Forthcoming.*

Raisch, Sebastian and Sebastian Krakowski, *"Artificial intelligence and management: The automation–augmentation paradox," Academy of Management Review, 2021, 46 (1), 192–210.*

Raj, Manav, Justin Berg, and Rob Seamans, *"Art-ificial Intelligence: The Effect of AI Disclosure on Evaluations of Creative Content," arXiv preprint arXiv:2303.06217, 2023.*

Schaeffer, R., B. Miranda, and S. Koyejo, *"Are emergent abilities of Large Language Models a mirage?," arXiv preprint arXiv:2304.15004, 2023.*

Sendak, M., M. C. Elish, M. Gao, J. Futoma, W. Ratliff, M. Nichols, A. Bedoya, S. Balu, and C. O'Brien, *""The human body is a black box" supporting clinical decision-making with deep learning," in "Proceedings of the 2020 conference on fairness, accountability, and transparency" 2020, pp. 99–109.*

Sergeeva, A. V., S. Faraj, and M. Huysman, *"Losing touch: An embodiment perspective on coordination in robotic surgery," Organization Science, 2020, 31 (5), 1248–1271.*

Singhal, K., S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, ..., and V. Natarajan, *"Large language models encode clinical knowledge," arXiv preprint arXiv:2212.13138, 2022.*

Soto, Christopher J. and Oliver P. John, *"Short and extra-short forms of the Big Five Inventory–2: The BFI-2-S and BFI-2-XS," Journal of Research in Personality, 2017, 68, 69–81.*

Teodorescu, Mike H. M., Lily Morse, Yazeed Awwad, and Gerald C. Kane, *"Failures of Fairness in Automation Require a Deeper Understanding of Human-ML Augmentation," MIS Quarterly, 2021, 45 (3), 1483–1500.*

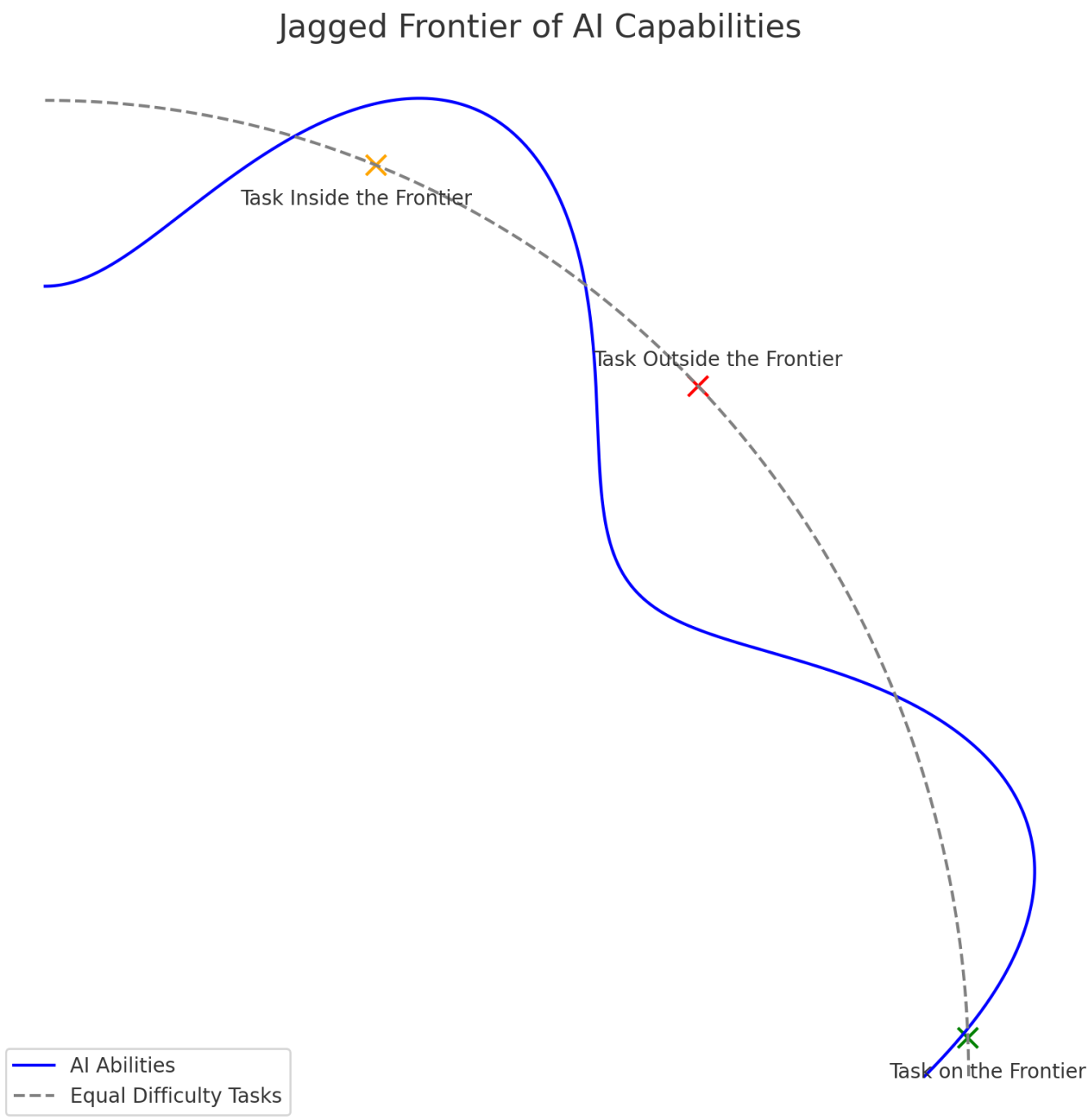
Tong, Siliang, Nan Jia, Xueming Luo, and Zheng Fang, *"The Janus face of artificial intelligence feedback: Deployment versus disclosure effects on employee performance," Strategic Management Journal, 2021.*

Valentine, M.A. and R. Hinds, *"From Resistance to Reskilling: How Experts Develop Valued New Skills Through Algorithmic Capability Building," Working Paper, Stanford University 2023.*

Van den Broek, E., A. Sergeeva, and M. Huysman, *"When the Machine Meets the Expert: An Ethnography of Developing AI for Hiring," MIS Quarterly, 2021, 45 (3).*

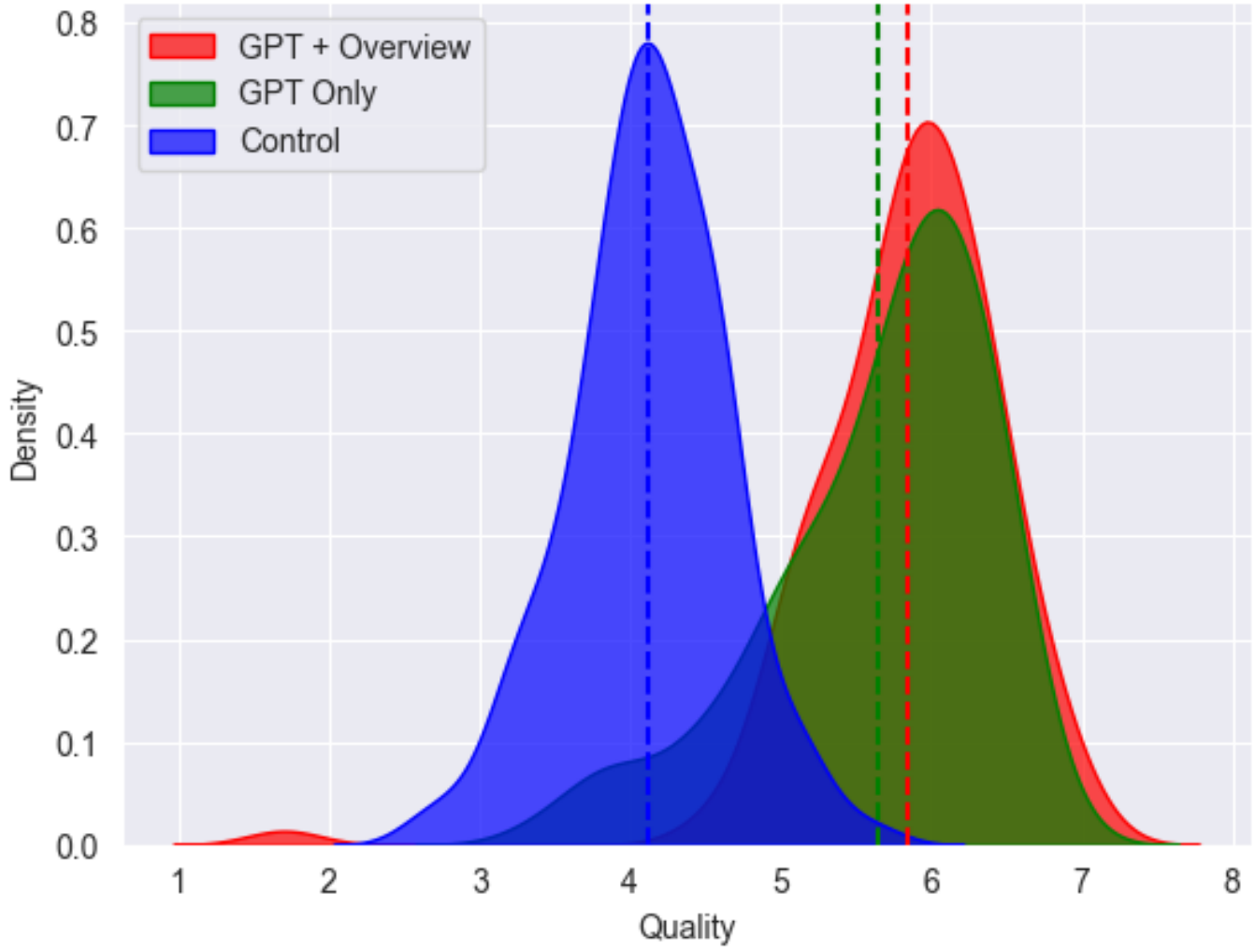
Waardenburg, L., M. Huysman, and A. V. Sergeeva, *"In the land of the blind, the one-eyed man is king: Knowledge brokerage in the age of learning algorithms," Organization Science, 2022, 33 (1), 59–82.*

Figure 1: **The Jagged AI Frontier**



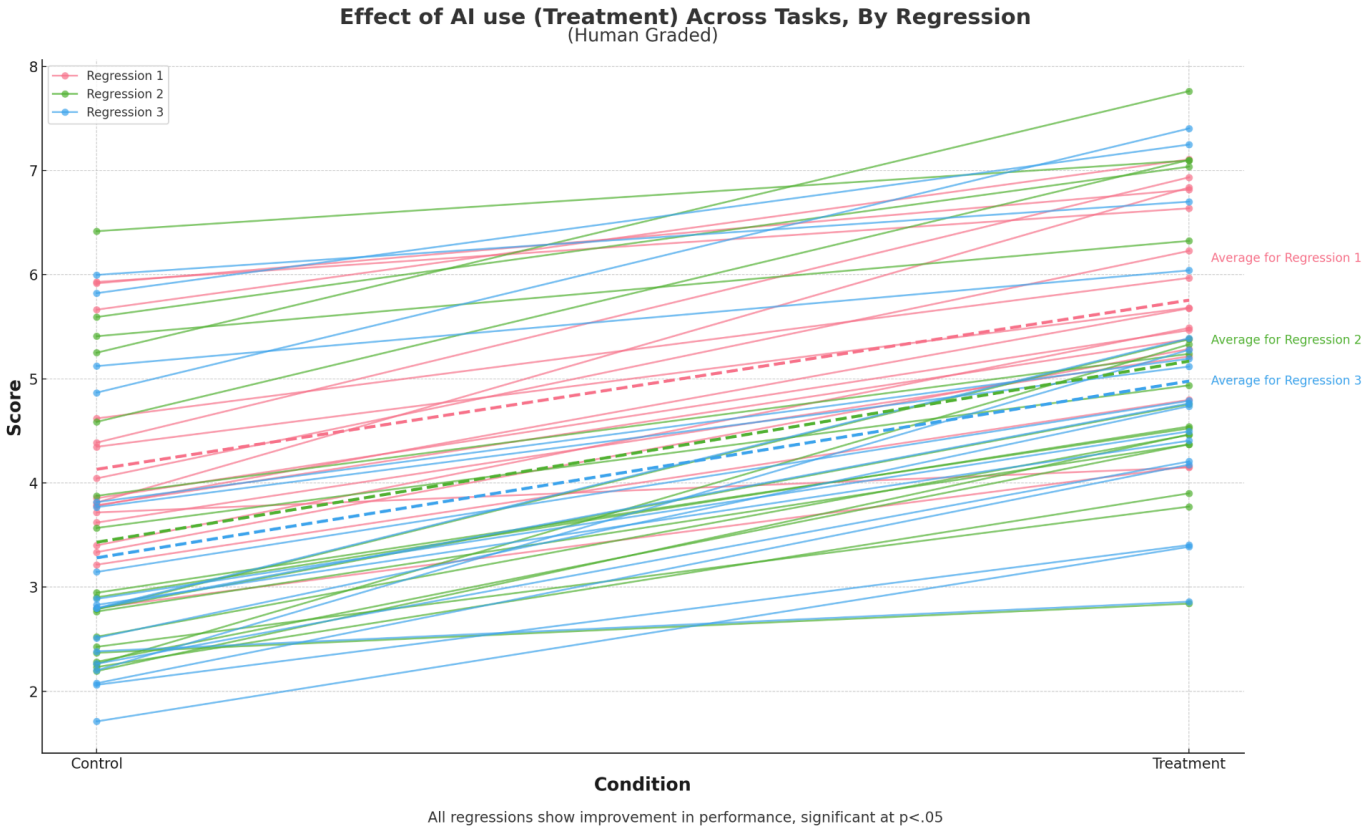
Notes: This figure displays the AI frontier as jagged. Tasks with the same perceived difficulty may be on one side or the other of the frontier. ChatGPT produced this image starting from the authors' prompts.

Figure 2: **Performance Distribution - Inside the Frontier**



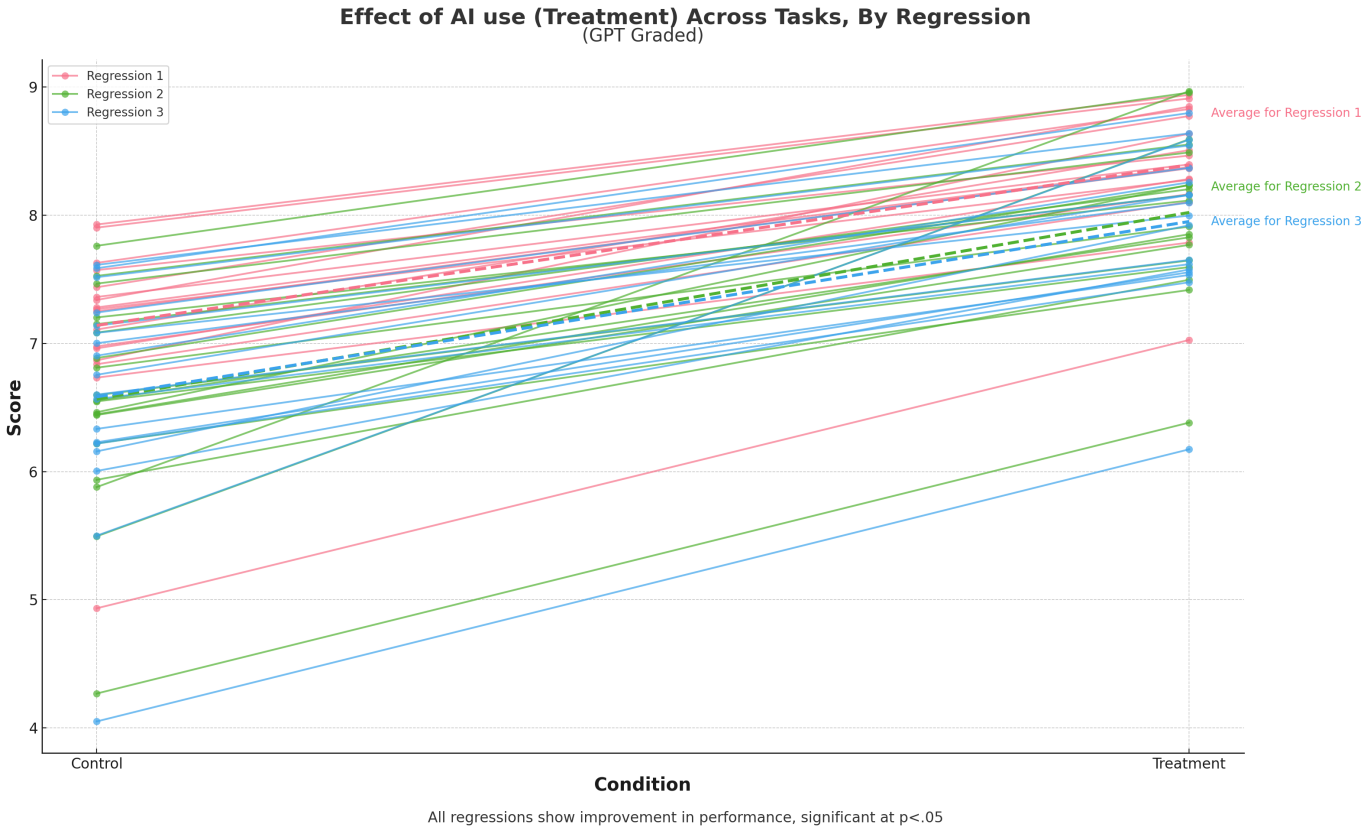
Notes: This figure displays the full distribution of performance in the experimental task inside the frontier for subjects in the three experimental groups (red for subjects in the GPT+Overview condition; green for subjects in the GPT Only condition; blue for subjects in the control condition).

Figure 3: **Performance - Inside the Frontier - Human Grades**



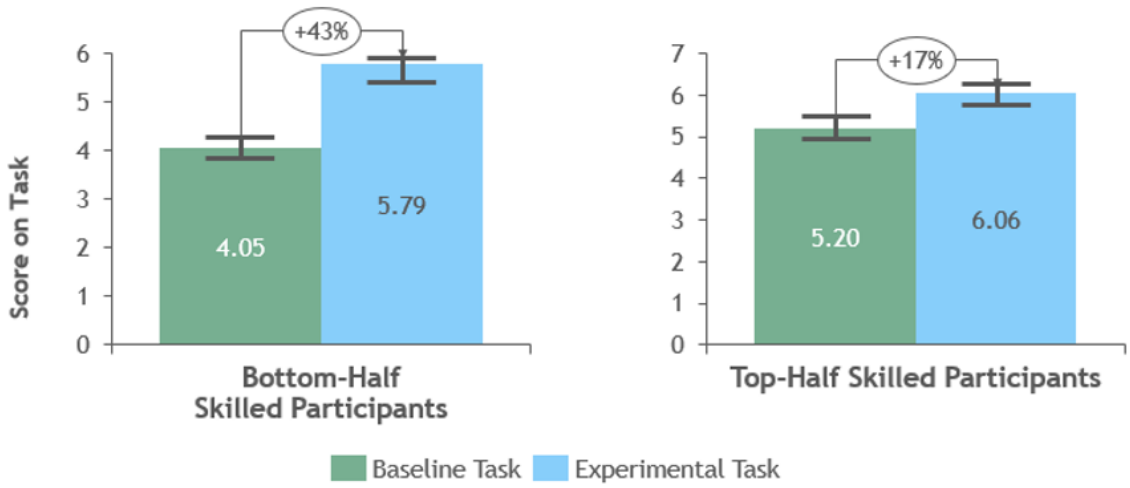
Notes: The figure displays a series of horizontal lines, each representing the estimated treatment effect for a specific question as evaluated by human graders and linear regression model. In total, there are 54 lines: 18 questions and 3 regression models for each question. These are the three regressions reported in Columns 1-3 in Table 1 Additionally, three dashed lines report the average effects in each regression. The y-axis of the figure is labeled with the outcome variable's scale. ChatGPT produced this image using data provided by the authors.

Figure 4: **Performance - Inside the Frontier - GPT Grades**



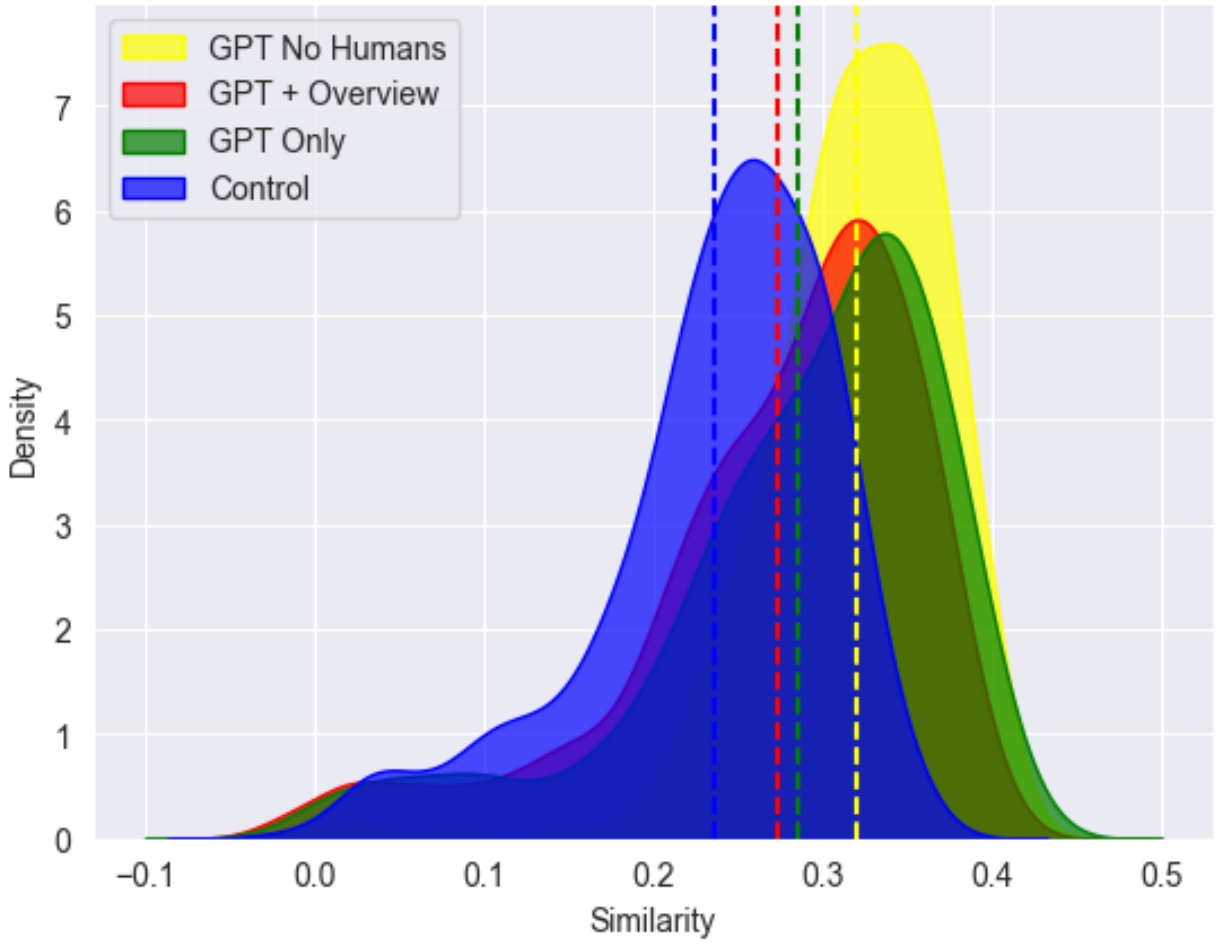
Notes: The figure displays a series of horizontal lines, each representing the estimated treatment effect for a specific question as evaluated by GPT and linear regression model. In total, there are 54 lines: 18 questions and 3 regression models for each question. These are the three regressions reported in Columns 1-3 in Table 1 Additionally, three dashed lines report the average effects in each regression. The y-axis of the figure is labeled with the outcome variable's scale. ChatGPT produced this image using data provided by the authors.

Figure 5: **Bottom-Half Skills and Top-Half Skills - Inside the Frontier**



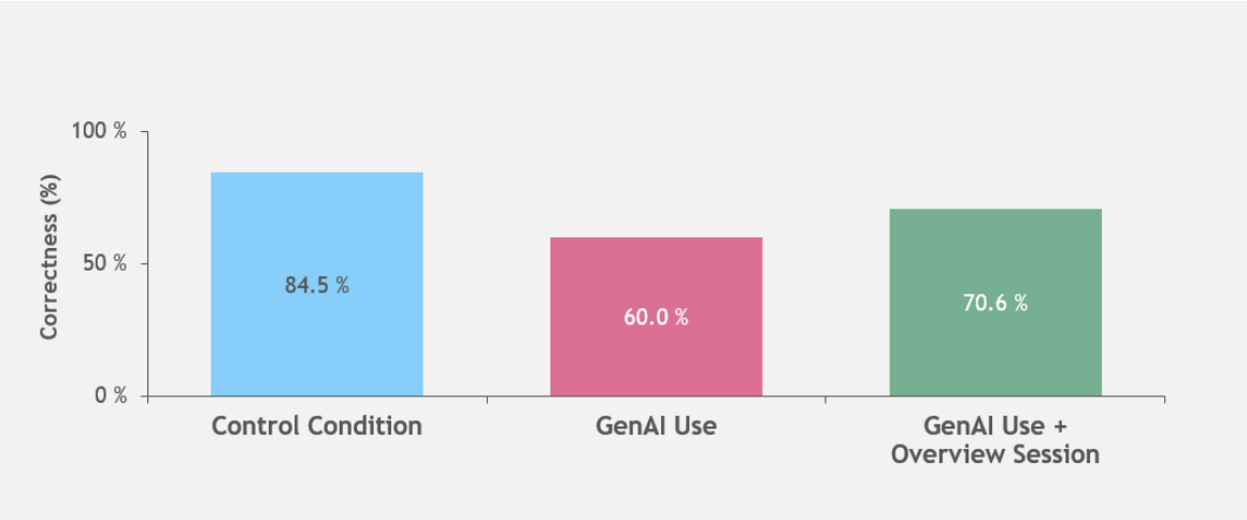
Notes: This figure displays the average performance of subjects in the bottom-half performance distribution in the assessment task (on the left), and those in the top-half performance distribution in the assessment task (on the right). The bars in green report their performance in the assessment task, while the bars in blue report their performance in the experimental task. The y-axis is labeled with the average scores (on a 1-10 scale).

Figure 6: **Similarity across Participants**



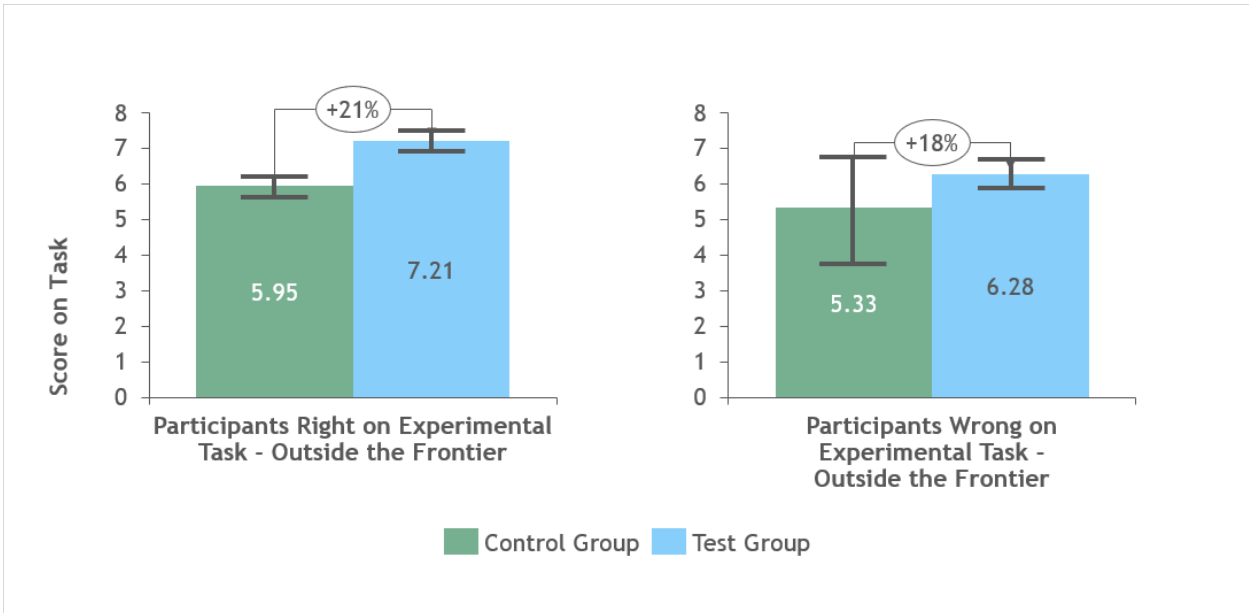
Notes: This figure displays the distribution of Average Within-Subject Semantic Similarity by experimental condition. Red for subjects in the GPT+Overview condition; green for subjects in the GPT Only condition; blue for subjects in the control condition; yellow for the additional "GPT No Human" condition produced with a simulation in 100 independent ChatGPT sessions responding to the experimental task's instructions as a prompt.

Figure 7: **Performance - Outside the Frontier**



Notes: This figure displays average performance for the task outside the frontier. It reports the percentage of subjects in each experimental group providing a correct response in the experimental task.

Figure 8: **Recommendation Quality**



Notes: This figure displays the average performance of subjects who were correct in the experimental task outside the frontier (on the left), and those who were incorrect on that task (on the right). The green bars represent the recommendation quality of the control group, while the blue bars indicate the average recommendation quality of the treatment groups. The y-axis denotes the average recommendation scores, ranging from 1 to 10.

—

Table 1: **Inside the Frontier - Quality**

	(1) Quality	(2) Quality	(3) Quality	(4) Quality (GPT)
GPT + Overview	1.746***	1.752***	1.769***	1.349***
	(0.074)	(0.070)	(0.075)	(0.058)
GPT Only	1.556***	1.585***	1.592***	1.216***
	(0.080)	(0.077)	(0.078)	(0.059)
Assessment		0.161***	0.158***	
		(0.048)	(0.051)	
Assessment (GPT)				0.167**
				(0.070)
Female		-0.182***	-0.183***	-0.042
		(0.070)	(0.070)	(0.049)
English Native		0.088	0.088	0.097*
		(0.072)	(0.076)	(0.055)
Low Tenure		0.061	0.062	0.018
		(0.064)	(0.065)	(0.048)
Location		-0.047	-0.048	0.049
		(0.080)	(0.082)	(0.061)
Tech Openness		0.086	0.075	0.065
		(0.068)	(0.076)	(0.058)
Controls			X	X
R2	0.598	0.631	0.634	0.638
GPT = GPT + Overview	0.029	0.053	0.047	0.049
Control Mean	4.099	4.099	4.099	7.207
Observations	385	385	385	385

Notes: This table examines the effects of introducing GPT-4 on the quality of the responses for the experimental task inside the frontier. Each column displays the results of a distinct linear regression model. Columns 1-3 have the average response quality, graded by two human evaluators, as their dependent variable. In contrast, Column 4 uses the average response quality in the experimental task as determined by GPT. Columns 2-3 incorporate the average response quality from the assessment task, as graded by human evaluators, while Column 4 utilizes the GPT-evaluated metric. Columns 3 and 4 include additional controls for measures of familiarity with AI, ChatGPT use, and perceptions about AI's automation abilities. The bottom of the table displays p-values from an F-test comparing the effects of receiving the GPT + Overview treatment versus the GPT Only treatment. All regressions include robust standard errors. *** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

Table 2: **Inside the Frontier - Completion**

	(1) Percent Compl.	(2) Percent Compl.	(3) Percent Compl.
GPT + Overview	0.111***	0.109***	0.105***
	(0.020)	(0.020)	(0.020)
GPT Only	0.090***	0.088***	0.082***
	(0.021)	(0.021)	(0.018)
Assessment		-0.007	0.006
		(0.013)	(0.011)
Female		-0.001	0.014
		(0.018)	(0.016)
English Native		0.014	-0.003
		(0.020)	(0.019)
Low Tenure		0.025	0.025*
		(0.017)	(0.015)
Location		-0.021	-0.004
		(0.020)	(0.018)
Tech Openness		0.017	0.009
		(0.017)	(0.017)
Percent Compl. (Assess)			0.367***
			(0.052)
Controls			X
R2	0.083	0.100	0.303
GPT = GPT + Overview	0.282	0.265	0.216
Control Mean	0.824	0.824	0.824
Observations	385	385	385

Notes: This table examines the effects of introducing GPT-4 on the subject's task completion for the experimental task inside the frontier. Each column displays the results of a distinct linear regression model. The dependent variable across all columns is the percentage of total questions that subjects successfully completed. Columns 2-3 use the average response quality in the assessment task as evaluated by two human graders as a control. Column 3 additionally includes the percentage of completed questions in the assessment task and additional controls for measures of familiarity with AI, ChatGPT use, and perceptions about AI's automation abilities. The bottom of the table displays p-values from an F-test comparing the effects of receiving the GPT + Overview treatment versus the GPT Only treatment. All regressions include robust standard errors. *** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

Table 3: **Inside the Frontier - Timing**

	(1) Timing	(2) Timing	(3) Timing
GPT + Overview	-1129.143***	-1105.622***	-1094.640***
	(135.181)	(136.614)	(129.681)
GPT Only	-1388.415***	-1356.364***	-1351.998***
	(150.204)	(152.252)	(138.056)
Assessment		159.955*	-0.376
		(88.270)	(81.973)
Female		53.466	-4.081
		(141.422)	(134.884)
English Native		-70.594	144.690
		(147.939)	(143.001)
Low Tenure		-89.041	-18.741
		(128.835)	(117.147)
Location		45.125	12.766
		(151.060)	(144.315)
Tech Openness		-45.955	-96.996
		(132.155)	(132.153)
Timing (Assessment)			1.474***
			(0.183)
Controls			X
R2	0.196	0.206	0.345
GPT = GPT + Overview	0.137	0.155	0.124
Control Mean 1	5023	5023	5023
Observations	385	385	385

Notes: This table examines the effects of introducing GPT-4 on timing the experimental task inside the frontier. The dependent variable represents the total seconds taken to reach the final question. Each column displays the results of a distinct linear regression model. Columns 2-3 use the average response quality in the assessment task as evaluated by two human graders as a control. Column 3 additionally includes the timing necessary to reach the last question in the assessment task, as well as additional controls for measures of familiarity with AI, ChatGPT use, and perceptions about AI's automation abilities. The bottom of the table displays p-values from an F-test comparing the effects of receiving the GPT + Overview treatment versus the GPT Only treatment. All regressions include robust standard errors. *** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$ *** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

Table 4: **Outside the Frontier - Correctness**

	(1) Correctness	(2) Correctness	(3) Correctness
GPT + Overview	-0.245***	-0.245***	-0.248***
	(0.054)	(0.054)	(0.054)
GPT Only	-0.139***	-0.145***	-0.145***
	(0.053)	(0.052)	(0.053)
Assessment		0.109**	0.117**
		(0.046)	(0.047)
Female		-0.063	-0.073
		(0.050)	(0.053)
English Native		-0.096**	-0.097**
		(0.046)	(0.048)
Low Tenure		-0.118***	-0.118**
		(0.045)	(0.046)
Location		-0.098*	-0.104*
		(0.052)	(0.053)
Tech Openness		0.012	0.016
		(0.050)	(0.052)
Controls			X
R2	0.051	0.099	0.109
GPT = GPT + Overview	0.082	0.095	0.088
Control Mean	0.844	0.844	0.844
Observations	373	373	373

Notes: This table examines the effects of introducing GPT-4 on the correctness of the responses for the experimental task outside the frontier. Each column displays the results of a distinct linear regression model. All columns use a binary correctness variable, as evaluated by human graders, as their dependent variable. Column 3 additionally includes a binary correctness metric from the assessment task, as graded by human evaluators, and introduces further controls for measures of familiarity with AI, ChatGPT use, and perceptions about AI's automation abilities. The bottom of the table displays p-values from an F-test comparing the effects of receiving the GPT + Overview treatment versus the GPT Only treatment. All regressions include robust standard errors.*** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

Table 5: **Outside the Frontier - Timing**

	(1) Timing	(2) Timing	(3) Timing
GPT + Overview	-689.191***	-671.526***	-677.139***
	(115.266)	(94.987)	(96.131)
GPT Only	-407.329***	-279.837***	-287.775***
	(121.833)	(95.751)	(97.945)
Assessment Timing		0.681***	0.681***
		(0.046)	(0.046)
Female		18.777	5.163
		(87.671)	(90.485)
English Native		-114.277	-118.673
		(85.932)	(90.779)
Low Tenure		82.151	86.135
		(81.736)	(83.910)
Location		57.024	48.935
		(95.524)	(97.050)
Tech Openness		34.603	65.572
		(85.090)	(89.744)
Controls			X
R2	0.085	0.407	0.414
GPT = GPT + Overview	0.022	0.000	0.000
Control Mean	2260	2260	2260
Observations	373	373	373

Notes: This table examines the effects of introducing GPT-4 on timing the experimental task outside the frontier. The dependent variable represents the total seconds taken to complete the exercise. Each column displays the results of a distinct linear regression model. Columns 2-3 include the timing of the assessment task, as well as additional controls for measures of familiarity with AI, ChatGPT use, and perceptions about AI's automation abilities. The bottom of the table displays p-values from an F-test comparing the effects of receiving the GPT + Overview treatment versus the GPT Only treatment. All regressions include robust standard errors. *** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

Table 6: **Outside the Frontier - Recommendation Quality**

	(1) Rec. Quality	(2) Rec. Quality	(3) Rec. Quality	(4) Rec. Quality
GPT + Overview	1.475***	1.331***	1.343*	1.544***
	(0.242)	(0.233)	(0.725)	(0.253)
GPT Only	1.046***	0.941***	1.570*	0.841***
	(0.289)	(0.265)	(0.807)	(0.285)
Assessment - Rec.		0.303***	0.277***	0.324***
		(0.044)	(0.102)	(0.048)
Female		-0.423*	-0.876	-0.186
		(0.243)	(0.582)	(0.259)
English Native		-0.105	-0.642	0.202
		(0.224)	(0.558)	(0.241)
Low Tenure		-0.057	-0.639	0.230
		(0.214)	(0.554)	(0.221)
Location		0.178	-0.148	0.324
		(0.236)	(0.543)	(0.247)
Tech Openness		0.197	-0.122	0.323
		(0.220)	(0.514)	(0.227)
Controls		X	X	X
R2	0.085	0.235	0.212	0.313
GPT = GPT + Overview	0.098	0.112	0.664	0.011
Control Mean	5.856	5.856	5.325	5.954
Observations	372	372	105	267

Notes: This table examines the effects of introducing GPT-4 on the quality of the recommendations provided in the experimental task outside the frontier. Each column displays the results of a distinct linear regression model. Columns 1-2 have the quality of recommendations provided in the experimental task, graded by two human evaluators, as their dependent variable. Column 2 includes the quality of recommendations provided in the assessment task, as well as additional controls for measures of familiarity with AI, ChatGPT use, and perceptions about AI's automation abilities. Columns 3 and 4 run the same regression as Column 2, using different samples. Column 3 takes into account only subjects who provided an incorrect response to the experimental task. Column 4 takes into account only those who provided a correct response. The bottom of the table displays p-values from an F-test comparing the effects of receiving the GPT + Overview treatment versus the GPT Only treatment. All regressions include robust standard errors. *** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

Appendix

A Tasks

In the following Appendix, we detail the specific tasks undertaken by subjects during the experiments. Both experiments started with an assessment task, serving as the initial phase where all subjects were required to complete it without any assistance from Generative AI. This initial task is crucial as it establishes a baseline, capturing the abilities and skills of the subjects in the absence of AI support. The subsequent task represents the core experimental phase. Here, subjects have the opportunity to leverage Generative AI, with the level of access determined by their respective treatment assignments. We report these tasks below for both experiments.

Outside the Frontier

Assessment Task

The CEO, Harold Van Muylders, would like to understand which of the three distribution channels that the company uses (fully owned stores, franchisee stores, or online) to focus his efforts. Please find attached interviews from company insiders on this issue. In addition, the attached Excel sheet provides financial data broken down by distribution channels.

Aim: Please prepare a 500-750 word note to the CEO. The note should focus on the following:

- If the CEO must pick one distribution channel to focus on to drive profit growth in the company, what channel should that be? What is the rationale for this choice? Please support your views with data and/or interview quotations as appropriate.
- Please also suggest innovative and tactical actions the CEO can take to boost profit growth in your chosen distribution channel. Please be creative, and feel free to rely on your own business judgement on what is appropriate for Kleding.

Experimental Task

The CEO, Harold Van Muylders, would like to understand Kleding's performance by the company's three brands (Kleding Man, Kleding Woman, and Kleding Kids) to uncover deeper issues. Please find attached interviews from company insiders on this issue. In addition, the attached excel sheet provides financial data broken down by brands.

Aim: Please prepare a 500-750 word note to the CEO. The note should focus on the following:

- If the CEO must pick one brand to focus on and invest to drive revenue growth in the company, what brand should that be? What is the rationale for this choice? Please support your views with data and/or interview quotations.

- Please also suggest innovative and tactical actions the CEO can take to improve this chosen brand. Please be creative, and feel free to rely on your own business judgement on what is appropriate for Kleding.

Inside the Frontier

Assessment Task

You are working for a beverage company in the unit developing new products. Your boss asked you to present an idea for a new product at the next manager meeting. Please, respond to the questions below.

1. Generate ideas for a new drink in markets that are underserved. Be creative, and give at least 10 ideas.
2. Pick the best idea, and explain why, so that your boss and other managers can understand your thinking.
3. Describe a potential prototype drink in vivid detail in one paragraph (3-4 sentences).
4. Come up with a list of steps needed to launch the product. Be concise but comprehensive.
5. Come up with a name for the product: consider at least 4 names, write them down, and explain the one you picked.

Experimental Task

You are working for a footwear company in the unit developing new products. Your boss asked you to present an idea for a new product at the next manager's meetings. Please, respond to the questions below.

1. Generate ideas for a new shoe aimed at a specific market or sport that is underserved. Be creative, and give at least 10 ideas.
2. Pick the best idea, and explain why, so that your boss and other managers can understand your thinking.
3. Describe a potential prototype shoe in vivid detail in one paragraph (3-4 sentences).
4. Come up with a list of steps needed to launch the product. Be concise but comprehensive.
5. Come up with a name for the product: consider at least 4 names, write them down, and explain the one you picked.
6. Use your best knowledge to segment the footwear industry market by users. Keep it general, and do not focus yet on your specific target and customer groups.

7. List the initial segments might you consider (do not consider more than 3).
8. List the presumed needs of each of these segment. Explain your assessment.
9. Decide which segment is most important. Explain your assessment.
10. Come up with a marketing slogan for each of the segments you are targeting.
11. Suggest three ways of testing whether your marketing slogan works well with the customers you have identified.
12. Write a 500-word memo to your boss explaining your findings.
13. Your boss would like to test the idea with a focus group. Please, describe who you would bring into this focus group.
14. Suggest 5 questions you would ask the people in the focus group.

    Now, imagine your new product entering the market.

15. List (potential) competitor shoe companies in this space.
16. Explain the reasons your product would win this competition in an inspirational memo to employees.
17. Write marketing copy for a press release.
18. Please, synthesize the insights you have gained from the previous questions and create an outline for a Harvard Business Review-style article of approximately 2,500 words. In this article, your goal should be to describe your process end-to-end so that it serves as a guide for practitioners in the footwear industry looking to develop a new shoe. Specifically, in this article, please describe your process for developing the new product, from initial brainstorming to final selection, prototyping, market segmentation, and marketing strategies. Please also include headings, subheadings, and a clear structure for your article, which will guide the reader through your product development journey and emphasize the key takeaways from your experience. Please also share lessons learned and best practices for product development in the footwear industry so that your article serves as a valuable resource for professionals in this field.

B Evaluation Rubric - Recommendation Quality

Score	Description	Example
1	• Participant does not identify tactical actions for client to boost profits	*"From a perspective target of driivng profit growth, the recommendation would be to further develop franchisee which is the channel with the strongest financials: very steady margins: ~2% CAGR for the full channel over the last 4 years, in line with revenue growth for the channel. The business model is very straightforward for the client with limited financial risks.* *In addition, there is very good prospects for the channel, as per News Article."(**1**)*
2-4	• Participant alludes to recommendations on how to boost profit but does not explicitly call out tactical actions. • Description of actions lacks specificity. Little to no description of business reasoning and impact on profit.	*"3. Gives time for Kleding to improve its fully owned and online channels.* *By shifting focus to a profitable channel, it will give Kleding time to fix its fully owned and online channels, which require more work. For fully owned stores, we need to evaluate the entire store portfolio and potentially close those that are unprofitable.* *We believe it is important to keep some so that we don't become overly reliant on franchisees and lose leverage in profit sharing negotiations. Meanwhile, we may need to hire people who are more knowledgeable about e-commerce to improve both the top-line and bottom-line."(**3**)*
5	• Participant identifies tactical actions for client to boost profits. • Tactical actions are aligned with overall channel strategy • Participant does not describe how to implement strategy. Explanation lacks specificity. • Business reasoning unclear.	*"Going forward Kleding should deploy 3 tactics to increase profitability though franchisee stores –* *1. Negotiate real time access to Kleding sales data at Franchise stores and the ability to marketing and collection team to make adjustments to visual merchandising and discounts based on the data* *2. A clear communication route between the Kleding team and Franchise store managers so changes in strategy can be applied immediately* *3. Training to Franchise store staff on customer engagement, product knowledge and tactics for competitive selling"(5)*
6-8	• Participant identifies tactical actions for client to boost profit • Tactical actions are aligned with overall channel strategy • Participant describes tactical actions in detail and outlines how to implement • Business reasoning lacks clarity. Participant does not fully connect explanation to client concerns. Impact on profit unclear.	*"Key takeaway 3: Consider closing or shrinking underperforming stores* *The immediate action plan is to bring back the fully owned stores' strategy to the smaller/more intimate model. It's important to consider that some of the underperforming stores will be under long-term contracts, which can lead to fines and other immediate operational costs. Therefore, for some of the stores we need to consider whether would cost more to close the stores or to keep them and have a negative contribution margin*

		for the coming years. Trying to shrink the stores, instead of fully closing them can also lead to a more friendly contract renegotiation."(7)
9-10	• Participant identifies tactical actions for client to boost profit • Tactical actions are aligned with overall channel strategy • Participant includes elaborate description of how to implement tactical actions • Actions backed up by sound business reasoning. Participant draws on historical pain points and belief audits. • Participant outlines impact of action on profit. • Points deducted if explanation is incomplete	*"Tactical actions:* 1. *Close large unprofitable stores: as we learnt from Head of Finance, bigger stores mean higher rents, higher staff costs, high amortization costs. Since none of the new stores opened since 2015 has been profitable, we should close those stores. Despite near term cost, it would help to stop losing money earlier than later.* 2. *Focus on quality over quantity: instead of showcasing all items in the collection, the stores should focus on selling the top items that have potential to be bestsellers. At the same time, it should try to create a more intimate shopping experience in influencing customers' decisions.* 3. *Invest in employee training programs: as Head of Owned Stores mentioned, the rapid expansion has lead to lack of employee trainings, therefore less knowledgeable staff and suboptimal sales skills. By investing in employee training, they can understand the products deeper and target potential customers better, as well as improving their sales skills and customer service skills. With happier customers, there will be higher sales from returning customers, and overall brand image improvement, driving long term sales. There are also generative AI created sales simulations sellers can utilize to practice those skills.* 4. *Invest in advanced supply chain system: delays and stockouts can all lower customer satisfaction and lead to lower sales. With GenAI, there are more tools that can efficiently manage operations and provide customized suggestions."*

C Retainment

In this Appendix, we will focus on the concept of retainment, which captures the degree to which subjects with access to GenerativeAI directly *retain* the content produced by the AI in their submitted answers. To investigate retainment behavior, we will consider the creative problem-solving experiment, where subjects conceptualize new product ideas through a series of questions.

Measuring Retainment

We operationalize the measurement of retainment by first considering how to compute the similarity between two pieces of text. We utilize Restricted Damerau-Levenshtein distance (RDL) which essentially measures the smallest amount of character edits (deletions, insertions, substitutions, or adjacent transpositions) required to change one piece of text into another. Next, we note that we observe the entire session log of interactions a subject has with ChatGPT, i.e, each textual prompt the subject inputs during the session and each corresponding textual output the subject receives in response. Given the varying nature and sequence of how subjects prompt during their sessions, it is not generally possible to systematically map a single prompt/response to each question they are asked to answer. However, for a given question asked of the subject, we can measure how similar (in RDL distance) their provided answer is to each response they received from GPT during their session. We consider the smallest distance to indicate the subject's retainment for the given question, and use this distance to produce a normalized measure of retainment between the answer (a) and the corresponding most similar response (r^*): $retainment(a) = 1 - \frac{RDL(a, r^*)}{max\{length(a), length(r^*)\}}$. This measure of retainment is between 0 and 1; a 0 indicates that the subject completed changed every character between their submitted answer and the response they received, while a 1 indicates that the subject reported the character-for-character response as their answer. Our retainment measure would not fully capture when a subject's answer is a combination of multiple responses from ChatGPT, so ours is a conservative retainment measure.

Results

We compute each subject's average retainment—their retainment averaged across all answers they

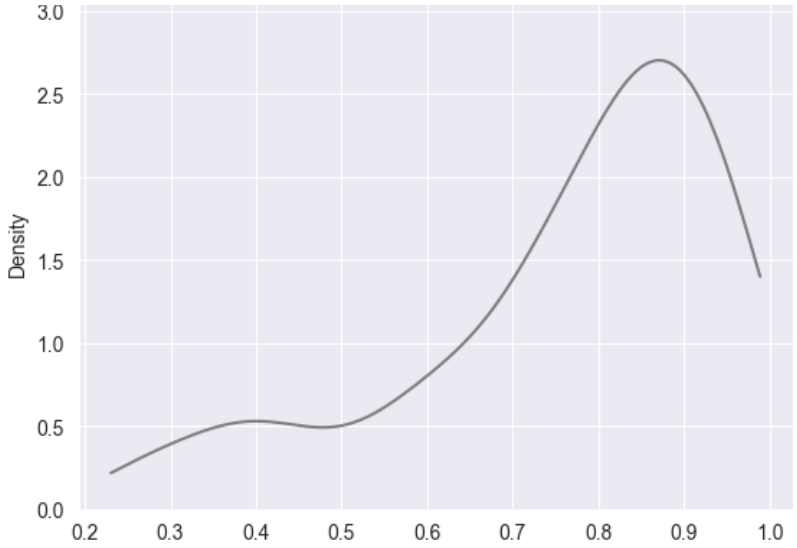


Figure 1: Distribution of Average Subject Answer Retainment

submitted—and display its distribution in Figure 1. From the distribution, we can see that a majority of subjects with access to ChatGPT retain a very high amount of its response in their submitted answers. The mode of average retainment is approximately 0.87, for context such a retainment value for a singular answer is quite high and can be obtained by changing a few words. Therefore, this result seems to clearly indicate that subjects are essentially "copying and pasting" ChatGPT responses as their submitted answers. It is natural to assume a high level of retainment indicates an abdication of judgment by the subjects, and while this is one way to interpret Figure 1, we caution against simply drawing this conclusion from the current results. Alternatively, a subject could be engaging in "high-quality" prompting behavior, e.g., helping ChatGPT to iteratively refine its responses until it is perfected, and only then retaining a high amount of the response in their answers. Currently, our analysis cannot distinguish between the two, and in general, it is not straightforward how to assess prompting quality.

We further decompose Figure 1, and consider the distribution of average retainment for subjects who did and did not receive training (in addition to ChatGPT access) in Figure 2 and observe that while this pattern of high retainment is present in both groups, it is more extreme in the group that received

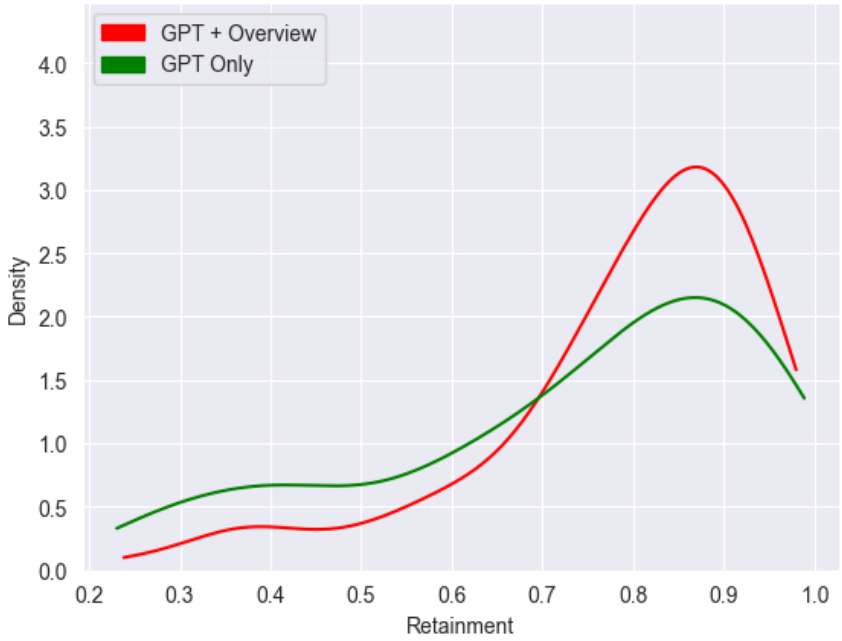


Figure 2: Distribution of Average Subject Retainment by Training Condition

training. This result appears to indicate that those subjects who receive training are more prone to retaining large amounts of ChatGPT responses in their answers. This would be consistent with the hypothesis that high retainment can come from engaging in high-quality prompting, which in turn can be improved by the training. However, it could also indicate that the training increased trust in ChatGPT's ability to effectively answer the questions and further increased subjects' willingness to abdicate judgment. Given the ongoing conversation on the spectrum of A.I.'s propensity to augment vs. replace human decision-making, a deep investigation into what is driving the retainment results of Figures 1 and 2 presents a fruitful future direction of research. There is considerable value in understanding if and which

subjects are prone to abdicating their judgment to A.I. and if (even light) training interventions can ameliorate or exacerbate this behavior.

Finally, while we do not know the mechanism by which subjects are selecting to retain high amounts of ChatGPT responses, we can objectively measure if there is a relationship between the level of retainment and answer quality. We obtained evaluations, ranging from 1 to 10, for each answer submitted by each subject, in one of four categories the given question was constructed to capture: creativity, persuasiveness, analyses, or writing quality. Figure 3 shows a scatter plot of the subject's score and their

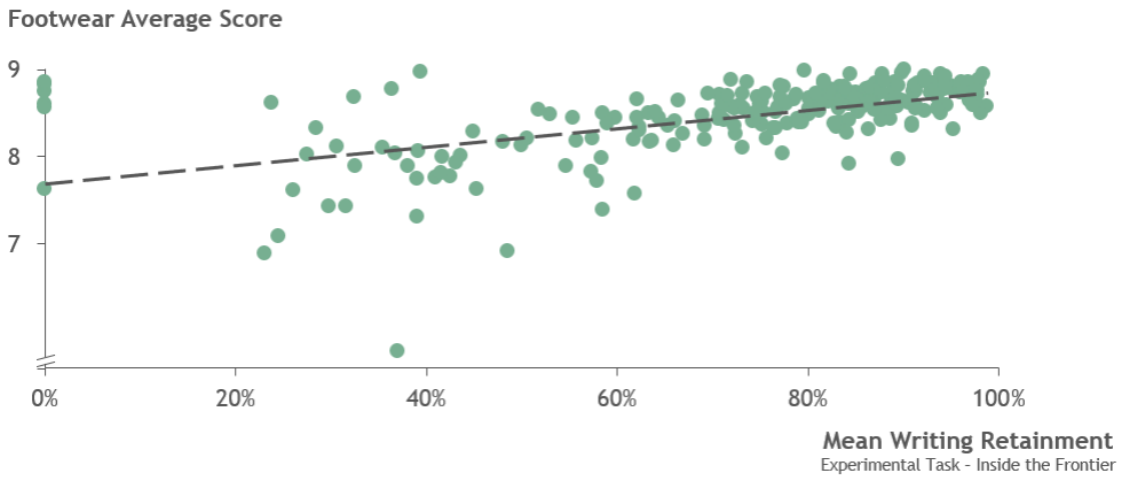


Figure 3: Relationship between Average Answer Quality and Average Retention

retention (both averaged over all questions), as well as the regression line capturing the linear relationship. Essentially all subjects do a fair job (the regression intercept is approximately 7.55 and highly significnat) which is not surprising given that the subjects represent the upper-end of the skills distribution and this experimental task is very representative of their work tasks. However, even in this highly-skilled subject pool, increased retainment of the response from ChatGPT is related to increased performance (the coefficient is approximately 1.21 and highly significant). A cursory interpretation of the result would be that for these types of creativity and writing tasks, results are generally better by minimizing human intervention. While potentially accurate this question merits deeper examination in future work, to layer in prompting behaviors or other dimensions that may explain the subject's *choice* to retain more (or less) of ChatGPT's responses. Either way, what is clear is that those who choose to retain to a relatively high degree produced significantly better answers on average, and that this certainly can have a profound impact on how organizations consider the use of GenerativeAI technologies by their employees.

D Collective Variation

In this Appendix, we will focus on the concept of *variation* in the content created by subjects. Specifically, we are interested in the diversity of the answers an individual subject produces with respect to the answers of other subjects. Importantly, we consider how the properties of these distributions are impacted by access to GenerativeAI. This question is motivated by various debates on the impact of GenerativeAI on firms, including how the ability for GenerativeAI to increase the productivity of individual employees translates to overall improvements for the firm. To investigate the variation in the content creation, we will consider the creative problem-solving experiment, and focus specifically on the first question where subjects are asked to generate at least 10 ideas for a new shoe aimed at a specific market or sport that is underserved.

Measuring Variation

We operationalize the measurement of variation by first considering how to compute the similarity between the ideas described by two different pieces of text. While simple character or word similarity is a possibility, this may fail to appropriately capture similarity between two pieces of text that express the same idea, but do not use words in common. This highlights the importance of choosing the appropriate *representation* for the pieces of text, and led us to use Goolge's Universal Sentence Encoder (USE) to build a *representation* that encodes the underlying meaning of the idea expressed in the text. More specifically, USE is designed to encode a piece of text into a 512-dimensional embedding vector that captures its underlying semantic meaning, in order to optimize transfer learning to downstream natural language processing tasks. Given these *semantic vector representations* of two ideas described in text, we can compute their vector inner-product to measure their *semantic* similarity. A subject produces approximately 10 ideas in their answer, so we separate and encode each of these ideas. We then compute the average inner-products between the idea of the given subject with all of the ideas produced by other subjects, within the same experimental condition. For each subject's idea, we then have a between-subject idea similarity and can average it across all the subject's 10 ideas to produce a subject level *between semantic similarity*. This measure of semantic similarity allows us to capture the variation in ideas of the collective experimental condition, where higher amounts of semantic similarity indicate less variation.

Results

In addition to the existing three experimental conditions, we artificially create a fourth condition (GPT Only), where we directly provide our question the initial prompt to ChatGPT and treat its response as an answer. We carry out this simulation in 100 independent ChatGPT sessions, each meant to represent a separate *GPT Only subject* and compute the same between semantic similarity analyses within this GPT Only condition.

We consider the between semantic similarity measured for each subject, and plot its distribution and mean (with a vertical line) separately for each experimental condition in Figure 1. We observe that subjects without access to ChatGPT tend to produce ideas with less semantic similarity (more conceptual variation) than those without access, implying that usage of ChatGPT reduces the range of ideas the subjects generate on average. We also observe that the GPT Only group has the highest degree of between semantic similarity, measured across each of the simulated subjects. These two results taken together point toward an interesting conclusion: the variation across responses produced by ChatGPT is smaller than what human subjects would produce on their own, and as a result when human subjects

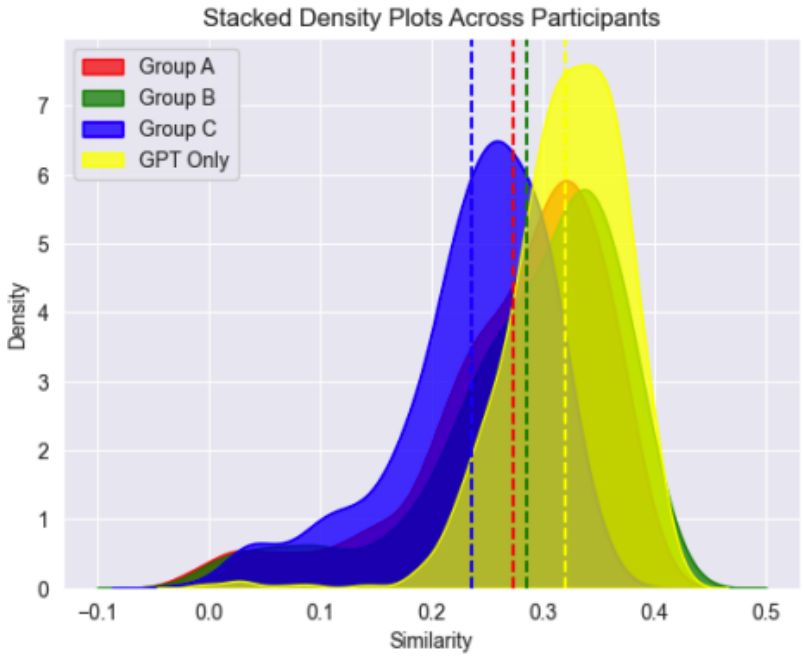


Figure 1: Distribution of Average Within Subject Semantic Similarity by experimental condition: Group A (Access to ChatGPT), Group B (Access to ChatGPT + Training), Group C (No access to ChatGPT), and GPT Only (Simulated ChatGPT Sessions).

use ChatGPT there is a reduction in the variation in the eventual ideas they produce. This result is perhaps surprising one would assume that ChatGPT, with its expansive knowledge base, would instead be able to produce many very distinct ideas, compared to human subjects alone. Moreover, the assumption is that when a human subject is also paired with ChatGPT the diversity of their ideas would increase.

While Figure 1 indicates access to ChatGPT reduces variation in the human-generated ideas, it provides no commentary on the underlying quality of the submitted ideas. We obtained evaluations of each subject's idea list along the dimension of creativity, ranging from 1 to 10, and present these results in Table 1. The idea lists provided by subjects with access to ChatGPT are evaluated as having significantly higher quality than those subjects without ChatGPT. Taken in conjunction with the between semantic similarity results, it appears that access to ChatGPT helps each individual construct higher quality ideas lists on average; however, these ideas are less variable and therefore are at risk of being more redundant.

While these collective results may appear to imply important limitations of ChatGPT (i.e., the convergence of ideas) from a firm's perspective, there are questions that still must be answered before arriving at such an important conclusion. Simply because the *full set* of 10 ideas for subjects with access to ChatGPT is generally of higher quality and less variable does not imply the *best* idea(s) for each subject follow this same pattern. Moreover, we cannot treat all subjects as homogenous, as some may use ChatGPT to improve the variability of their ideas while maintaining (or increasing) their quality; therefore, it is important to understand which types of subjects are potentially pushed toward more redundant ideas These various hypotheses warrant a deeper investigation into the mechanisms behind what we observe, and how to intervene to achieve the desired ends.

Appendix E

Centaur and Cyborg Practices

In many industries and for many types of analytic tasks, the discussion is no longer about *whether to adopt* AI but rather about *how to use* AI. This field experiment advances our understanding about how and when humans deeply engage with AI in knowledge work. By studying the knowledge work of 244 professional consultants as they used AI to complete a real-world, analytic task, we found that new human-AI collaboration practices and reconfigurations are emerging as humans attempt to navigate the jagged frontier. Here, we detail a typology of practices we observed, which we conceptualize as *Centaur and Cyborg practices.*

Centaur behavior. Named after the mythical creature that is half-human and half-horse, this approach involves a similar strategic division of labor between humans and machines closely fused together. Users with this strategy switch between AI and human tasks, allocating responsibilities based on the strengths and capabilities of each entity. They discern which tasks are best suited for human intervention and which can be efficiently managed by AI. From a frontier perspective, they are highly attuned to the jaggedness of the frontier and not conducting full sub-tasks with genAI but rather dividing the tasks into sub-tasks where the core of the task is done by them or genAI. Still, they use genAI to improve the output of many sub-tasks, even those led by them.

For example, Figure 1 depicts how user BA1 applied the centaur practice of drawing on AI for a particular writing task (a relative strength of AI) while drawing on human knowledge for the tasks of data analysis and generating recommendations (relative strengths of humans). Specifically, BA1 used human knowledge to complete the task of generating a recommendation and related information, and then switched to AI to draft a memo. BA1 asked AI:

You are writing a memo to a CEO of a company to inform him of where his company should focus as they attempt to drive revenue growth for one of their three brands, Kleding man. You want to provide him with the following facts for why he should focus on driving revenue growth for Kleding man…

Our early analysis of Centaur practices suggests that particular practices tend to be used at particular times in the analytic process; for example, users engaged in the practice of drawing on AI for its strength in refining user text at the beginning or end of the analytic processes, to either help set up the analytic process, accessing general information or methods, or to structure and edit the final output.

Cyborg behavior. Named after hybrid beings, as envisioned in science fiction literature, that seamlessly blend machine components with human biology, this approach is about intricate integration. Users do not just have a clear division of labor here between genAI and themselves; they intertwine their efforts with AI at the very frontier of capabilities. This manifests at the subtask level, when for an external observer it might even be hard to demarcate whether the output was produced by the human or the AI as they worked tightly on each of the activities related to the sub task.
For example, **Figure 3** depicts a different user, BA3, assigning his own professional persona (of a consultant) to the genAI to provide the input. Since LLMs are trained on a large breadth of

data, assigning a persona can be used to guide AI to a particular set of data. In this example, BA3 was responding to the previous output from AI, first asking AI to make editorial changes to the outputs, and then assigning a persona, instructing AI to "act as a consultant in the answering of the question."

We detail in Table 1 below examples of the centaur and cyborg practices we identify. Centaur practices entail using individual's knowledge of the current strengths of generative AI relative to theirs to switch between human and AI for each of the sub-task of the tasks accordingly throughout the workflow (example in **Figure 1**: **Centaur Practices – Using AI to Refine Input from User's Strength)**. Cyborg practices entail tightly integrating AI and human outputs in the tasks (example in **Figure 2**: **Cyborg Practices – Using AI to Validate Work and Requesting Editorial Changes)**. **Table 1** details the practices used by centaurs versus those used by cyborgs.

Our analysis was conducted on a sub-task level and it could be that the same individual would be applying centaur behavior for one type of tasks and cyborg for another. Our early analysis suggests that several factors may explain . In particular, it seems that humans with different levels of skill in the task domain, with different degrees of hands-on practice with generative AI, and with different perceptions of the relationship between humans and AI, may use different sets of practices. We are also investigating this is detail and the impact of these behaviors on performance. We assume some of these practices enable navigating the jagged frontier in a superior way to others.

Table 2: Centaur and Cyborg Practices

	Practices and Descriptions
Centaur Practices	**Use individual's knowledge of the current strengths of generative AI relative to theirs to switch between human and AI for each of the modules/sub-task of the tasks accordingly throughout the workflow.** **Examples of Behavior/practices (given the state of AI at the time of the experiment):** • **Mapping Problem domain**: Asking AI for general information related to the problem's domain for the human to use for their sub-task. • **Gathering methods information:** Asking AI for specific information on methods that the human is employing to solve their sub-task • **Refining human generated content**: Users providing their own output and using AI to refine its presentation.
Cyborg Practices	**Use AI for each of the sub-tasks throughout the whole workflow. Apply principles based on current knowledge about how to best elicit useful outputs from AI and/or continually question AI and experiment to reach a better output.** **For example:** • **Assigning a persona**: Instructing AI to simulate a specific type of personality or character • **Requesting editorial changes to AI output**: Asking AI to make editorial changes to the outputs AI has produced

	• **Teaching through examples**: Giving example of correct answer before asking AI a question • **Modularizing tasks**: Breaking down tasks into multiple sub-steps for AI to execute • **Validating:** Asking AI to check its inputs, analysis, and outputs • **Demanding logic explanation:** Asking AI to explain a confusing output; or why a particular recommendation was made • **Exposing Contradictions** Pointing out logical or factual inconsistencies • **Elaborating:** Asking AI to bring more breadth of details and nuance on an interesting or unexpected point • **Directing a Deep dive:** Directing AI to focus on a particular data point, content or task • **Adding user's own data:** Adding data after an output is generated to re-do the analysis in iterative cycles • **Pushing back:** Disagreeing with the output and ask AI to reconsider

Figure 1: Centaur Practices – Using AI to Refine Input from User's Strength

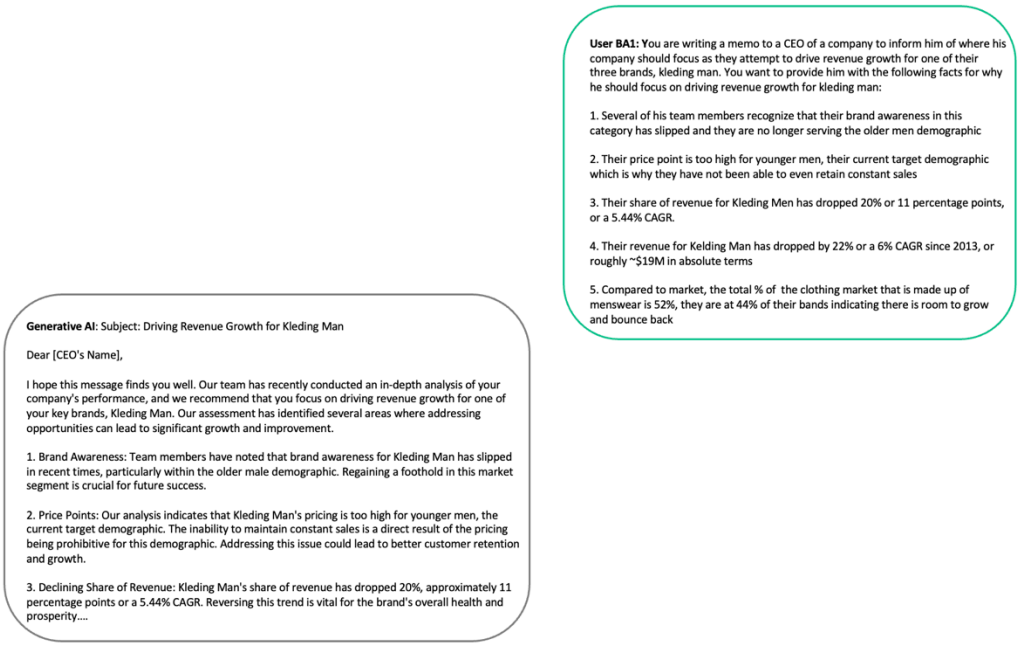


Figure 2: Cyborg Practices – Using AI to Validate Work and Requesting Editorial Changes