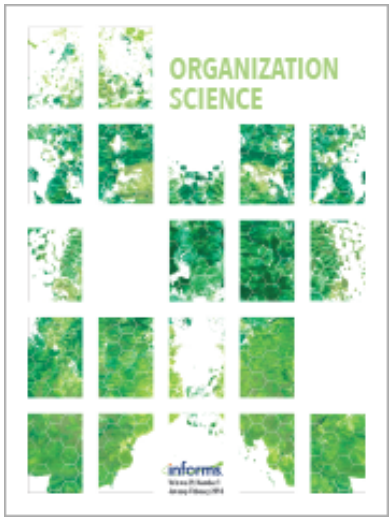


Organization Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

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

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

To cite this article:
Fabrizio Dell'Acqua, Edward McFowland III, Ethan Mollick, Hila Lifshitz, Katherine C. Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, Karim R. Lakhani (2026) Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of Artificial Intelligence on Knowledge Worker Productivity and Quality. Organization Science

Published online in Articles in Advance 11 Mar 2026

. https://doi.org/10.1287/orsc.2025.21838

https://pubsonline.informs.org/journal/orsc

ORGANIZATION SCIENCE
Articles in Advance, pp. 1–21
ISSN 1047-7039 (print), ISSN 1526-5455 (online)

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

Fabrizio Dell'Acqua,[a,*] Edward McFowland III,[a] Ethan Mollick,[b] Hila Lifshitz,[a,c] Katherine C. Kellogg,[d] Saran Rajendran,[e] Lisa Krayer,[e] François Candelon,[e] Karim R. Lakhani[a]

[a] Digital Data Design Institute, Harvard Business School, Boston, Massachusetts 02134; [b] The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104; [c] Artificial Intelligence Innovation Network, Warwick Business School, London CV4 7AL, United Kingdom; [d] Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142; [e] Henderson Institute, Boston Consulting Group, Boston, Massachusetts 02110

*Corresponding author

Contact: fdellacqua@hbs.edu, https://orcid.org/0000-0002-1998-0542 (FD); emcfowland@hbs.edu, https://orcid.org/0000-0001-5249-7117 (EdM); emollick@wharton.upenn.edu, https://orcid.org/0000-0001-6231-496X (EtM); hdiginnovation@gmail.com, https://orcid.org/0000-0002-3461-003X (HL); kkellogg@mit.edu, https://orcid.org/0000-0003-4372-3498 (KCK); saravanan0822@gmail.com (SR); krayer.lisa@bcg.com (LK); francois.candelon@seven2.eu (FC); klakhani@hbs.edu, https://orcid.org/0000-0002-5535-8304 (KRL)

Received: December 18, 2025
Revised: January 23, 2026
Accepted: January 27, 2026
Published Online in Articles in Advance: March 11, 2026

https://doi.org/10.1287/orsc.2025.21838

Abstract. We introduce and study the concept of a "jagged technology frontier" to describe the uneven impact of artificial intelligence (AI) capabilities, where AI assistance improves performance for some tasks but worsens it for others, even within the same knowledge workflow and with a seemingly similar level of difficulty. In collaboration with the global management consulting firm Boston Consulting Group, we have developed realistic management consulting tasks and examined the human performance implications of using AI to perform complex and knowledge-intensive work. The preregistered experiment involved 758 knowledge workers. After establishing a performance baseline on similar tasks, 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. For each one of a set of 18 realistic knowledge tasks within the frontier of AI capabilities ranging from creative to analytical tasks, subjects using AI outperformed those not using AI, completing 12.2% more tasks and completing them 25.1% more quickly on average while also delivering solutions of significantly improved quality. However, for a complex managerial task selected to be outside the frontier, subjects using AI were 19% less likely to produce correct solutions compared with those without AI, pointing to potential limitations of AI supporting knowledge workers. We discuss the positive and negative implications of AI-aided human performance in knowledge-intensive tasks.

Funding: Financial support of the Harvard Business School Digital Data Design Institute and Division of Research and Faculty Development is acknowledged.

Supplemental Material: The online appendix is available at https://doi.org/10.1287/orsc.2025.21838.

Keywords: technology and innovation management • organizational economics • economics and organization • organization and management theory • research design and methods • field experiments • implementation of new technology • organizational processes

1. Introduction

The capabilities of artificial intelligence (AI) systems to aid humans in a variety of tasks have improved rapidly, especially since the release of OpenAI's ChatGPT, which made large language models (LLMs) widely available for public use. Recent research has shown that these systems are unexpectedly capable of augmenting human abilities in the completion of creative, analytical, and writing tasks as well as achieving top scores in academic aptitude tests at graduate and professional levels (Geerling et al. 2023, Noy and Zhang 2023, Boussioux et al. 2024, Brynjolfsson et al. 2025). This represents a significant shift in human-machine augmentation, with hybrid aptitudes increasingly relevant across a broad range of knowledge work—including some highly creative and highly educated professions (Eloundou et al. 2024, Zhou and Lee 2024).

As generative AI's capabilities increasingly overlap with human skill sets, the integration of AI into human work presents new fundamental challenges,

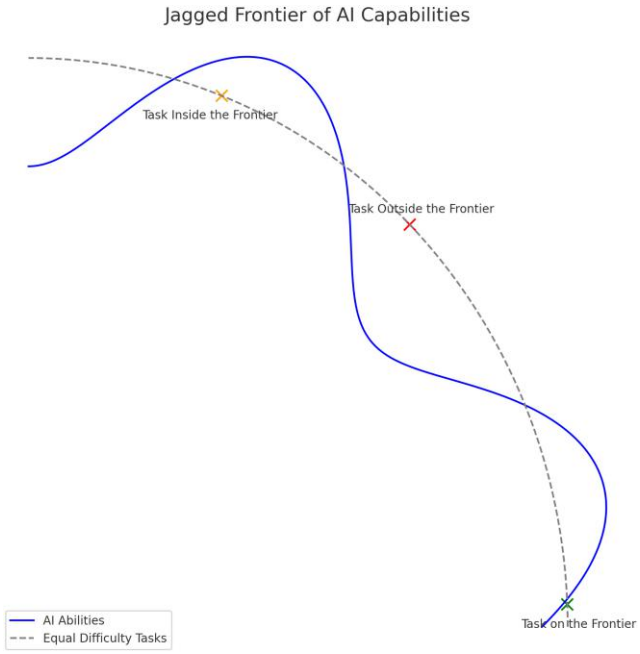
opportunities, and questions, especially in knowledge-intensive scientific and professional fields. In this paper, we investigate the impact of generative AI on management consulting—a crossdisciplinary field where top graduates from leading universities tackle diverse projects across industries. We examine this issue using a randomized, controlled laboratory-in-the-field experiment, in which highly skilled management consultants performed tasks resembling their standard work assignments.

Our results demonstrate that AI capabilities cover an uneven set of knowledge work, and we introduce the concept of a "jagged technological frontier" to characterize the performance of AI when used by knowledge workers. Tasks that appear to human knowledge workers to be of similar difficulty may be performed either better or worse by humans using AI. Within this jagged frontier, AI can complement human work. However, outside the frontier, AI output is inaccurate, is less useful, and can degrade human performance. AI assistance improves human performance only for tasks within current AI capabilities—within the jagged technological frontier—and worsens human performance outside of it. We find that workers who skillfully navigate this frontier in their use of AI systems gain substantial quality and productivity benefits.

Figure 1 provides a conceptual representation of this jagged technological frontier, illustrating how tasks that appear similar in complexity for humans may nonetheless fall on opposite sides of the boundary where AI assistance helps versus harms performance. Notably, jaggedness reflects systematic differences in how well AI capabilities align with different tasks compared with expectations based on human capabilities. Because the capabilities of AI are rapidly evolving and poorly understood, it can be hard ex ante for knowledge workers to grasp exactly what the boundary of this frontier might be at a given moment in time. Indeed, within the same knowledge workflow, some tasks are beyond the frontier, whereas others remain within it, making effective AI use challenging—especially without diligent examination of the workflow to identify deviations from expectations and the potential for redesign.

The potential benefits of AI for knowledge work have been a subject of considerable interest, although much of this research focuses on AI prior to the release of GPT-3.5 (e.g., Brynjolfsson et al. 2018, Çalli et al. 2021, Davies et al. 2021, Monisha et al. 2021). The general availability of GPT-3.5 in the form of ChatGPT since November 2022 has changed both the nature and urgency of the discussion (Berg et al. 2023, Eloundou et al. 2024, Zhou and Lee 2024). Studies of previous generations of AI (e.g., Agrawal et al. 2018, Furman and Seamans 2019, Brynjolfsson et al. 2021) and of recently released LLMs (e.g., Noy and Zhang 2023, Choi and Schwarcz 2024, Brynjolfsson et al. 2025) suggest that these systems can considerably impact worker performance. Our study focuses on complex

Figure 1. The Jagged AI Frontier



Notes. This figure provides a conceptual illustration of the jagged technological frontier of AI. The dashed line represents tasks of roughly equal perceived difficulty to human knowledge workers. Tasks that appear similar in difficulty to humans may fall on opposite sides of the frontier such that AI assistance improves performance for some tasks while degrading it for others. The figure is stylized and not intended to depict specific tasks or empirical effect sizes. ChatGPT produced this image starting from the authors' prompts.

tasks selected and evaluated by industry experts to replicate the real-world task assignments of knowledge workers. Most knowledge work includes a set of interdependent tasks, some of which may be a good fit for current LLMs, whereas others are not. We examine multiple types of interdependent knowledge tasks and build on recent studies to suggest ways of understanding the impact of generative AI on real-world task assignments, under which circumstances knowledge-based organizations may benefit but also face negative outcomes, and we consider how this impact might change as the technology advances.

Three factors make it difficult for knowledge workers to know what the value and downsides of using generative AI may be ahead of time. First, LLMs have surprising capabilities that they were not explicitly or intentionally developed to have—capabilities that are growing rapidly over time as model size, computational capability, and algorithmic performance increase. Although trained as general models, LLMs demonstrate specialist knowledge and abilities as part of their training process and during normal use (Reed et al. 2022, Boiko et al. 2023, Moor et al. 2023, Singhal et al. 2023). Although considerable debate remains on the concept of emergent capabilities from a technological perspective (Schaeffer et al. 2023), the effective capabilities of LLMs and the ways that they are used

are novel and unexpected, widely applicable, and increasing rapidly over short time spans. Recent work has argued that AI performs at a high level in knowledge work contexts ranging from medicine to law (Ali et al. 2023, Lee et al. 2023) and that the human-AI combination outperforms humans alone on many measures of innovation (Boussioux et al. 2024, Meincke et al. 2024). Additionally, scores on various standardized academic tests, albeit an imperfect measure of LLM capabilities, have 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 factor differentiating LLMs from previous forms of AI: their ability to directly improve the performance of knowledge workers who use these systems and more generally, their unprecedented usefulness in knowledge-intensive domains. Early studies of the new generation of AI suggest direct performance increases from using AI, especially for complex work, like writing tasks (Noy and Zhang 2023) and software programming (Peng et al. 2023), as well as for ideation and creative work (Boussioux et al. 2024, Meincke et al. 2024). Consequently, the effects of AI are expected to be higher in knowledge-intensive domains featuring highly educated workers (Felten et al. 2023, Eloundou et al. 2024).

The final relevant characteristic of generative AI is its relative opacity. Although previous generations of AI are often black boxed, which can be frustrating for users (Lebovitz et al. 2022), LLMs have been shown to produce incorrect but plausible results (hallucinations or confabulations) in a way that users find accessible and "believable," leading to inaccurate outputs.[1] The advantages of LLMs, although substantial, are frequently unclear to users. Despite performing well at some tasks, LLMs fail in other circumstances in ways that are difficult to predict in advance. Adding to this opacity, developers provide no guidance regarding the best way to use these AI systems, which appear to be best learned via ongoing user trial and error and the exchange of experiences and heuristics on various online forums.

These three factors together suggest that both the value and downsides of using generative AI may be difficult for knowledge workers to know ahead of time. Some unexpected tasks (like idea generation) are easy for AIs, whereas tasks that seem as though they should be easy for machines to complete are challenging for some LLMs (e.g., basic math). Our research suggests that this feature of LLMs as used by humans creates a "jagged frontier," where tasks that appear to be of similar difficulty are performed better or worse by humans using AI. Because of the "jagged" nature of the frontier, a knowledge worker may face tasks in an assignment that are on both sides of the frontier without a priori knowledge regarding each task's position with respect to that frontier. Knowledge workers may fail to see AI's limitations and become overly reliant on AI's output, potentially overlooking its limitations (Dell'Acqua 2022). 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. This work investigates how humans navigate this jagged frontier and the subsequent performance implications.

For this study, we collaborated with a global management consulting firm (Boston Consulting Group (BCG)), advising them on the design, development, and execution of a preregistered randomized experiment to assess the impact of AI on knowledge workers.[2] We specifically worked with them to develop tasks that closely resemble the workflow of their management consultants. Subsequently, the author team received the data that the company collected for this experiment and conducted the analysis presented in this paper. Online Appendix F shows a detailed analysis using Bureau of Labor Statistics data that contextualizes our experimental occupation within the broader landscape of knowledge work, demonstrating how management consulting responsibilities align with the multifaceted demands of modern professional environments.

We tested two distinct sets of tasks: one situated beyond the frontier of AI capabilities and one situated within the bounds of that frontier. The experiment aimed to understand how AI integration with human-performed tasks might reshape the traditional workflows of highly skilled professionals in a knowledge domain. Both sets of tasks were developed to be realistic and were designed by senior professionals at BCG who had experience in the same roles as the participants during their careers and were currently supervising workers in those roles. A senior managing director and partner at the company commented that these tasks were "very much in line with part of the daily activities of the consultants" involved. Notably, some forms of these tasks are also used by the company to screen job applicants, typically from elite academic backgrounds (including PhDs), for their highly coveted positions. Online Appendix G uses a survey of 11 managing directors and partners at BCG, among the most senior executives at the firm, to confirm that the underlying competencies assessed by our experimental tasks are critical for recruiting, job performance, and career progression at BCG.

Our experimental findings demonstrate mixed results regarding the quality and performance effects of humans integrating AI into their task flow. For inside-the-frontier tasks, we show that this generation of LLMs is highly capable of significantly increasing quality and productivity, leading to substantial productivity benefits across multiple dimensions when used by humans. However, quality drops substantially when knowledge workers deploy AI to solve tasks that are beyond AI's capability

frontier. Moreover, the location of the frontier of AI capability is not immediately obvious to knowledge workers, even though the tasks themselves are quite familiar. Consultants tended to overrely on AI, leading to a decline in combined human-machine performance when closer human supervision was necessary (Athey et al. 2020, Dell'Acqua 2022, Glaeser et al. 2024). This is especially relevant when we consider that our participants were highly specialized knowledge workers working within their domain of expertise, emphasizing the need for knowledge workers to have a nuanced understanding of AI's capabilities as they navigate the evolving landscape of human-AI integration to optimize the balance between AI assistance and human expertise.

Because this jagged frontier is expanding and changing, the overall results suggest that AI will have a profound impact on knowledge work, one that will become more pronounced as LLM capabilities expand. However, the distribution of these impacts will be uneven. Our results not only highlight AI's potential to transform knowledge work but also, underscore the importance of human involvement in machine-assisted knowledge workflows. The effectiveness of AI in knowledge work will critically depend on human judgment—particularly to discern which tasks within the workflow are suited to leveraging AI augmentation and where human expertise should be prioritized.

2. Experimental Context

BCG is a premier global management consulting firm, admitting only around 1% of applicants (Vlamis et al. 2024). Once hired, consultants enter a highly structured career track in which promotions are largely determined by strict performance appraisals. Promotion decisions often hinge on consultants' abilities to excel in diverse tasks from creative idea generation to complex business modeling, making BCG an ideal setting to investigate how AI might affect skill development and performance. BCG also primarily promotes from within.

In close collaboration with senior BCG leaders, we designed tasks that capture core facets of consulting. These task types closely capture the breadth of work that consultants commonly handle—ranging from innovative solution proposals to methodical problem-solving. We crafted each task to reflect realistic business scenarios and crucially, ensured that some tasks were well suited to AI support and others were not as described in Section 4. To validate the real-world importance of these tasks beyond our immediate author team, we surveyed 11 managing directors and partners at BCG. Their feedback affirmed that the tasks that we designed were highly relevant for day-to-day consulting work and aligned with the core competencies that BCG evaluates in both recruiting and ongoing performance reviews (see Online Appendix G).

BCG management consultants represent an ideal population for studying knowledge work, which we define as nonroutine, cognitively demanding activities centered on the creation, application, or integration of specialized intellectual expertise.[3] Online Appendix F shows how management consulting exemplifies knowledge work by analyzing Occupational Information Network (O*NET) occupational data. Although BCG consultants operate at the upper end of the complexity spectrum, our comparative analysis reveals that the diverse tasks that they perform—from analytical problem-solving to creative ideation—mirror those found across other high-skill professional domains that require a minimum of an undergraduate degree for consideration to be eligible for entry into those professions.

3. Methods

We collected data from a randomized experiment to assess the causal impact of AI, specifically GPT-4 (then the most capable of the AI models available (June 2023)) on highly qualified professionals otherwise working traditionally without AI.[4] We preregistered our study, detailing the design structure, the experimental conditions, the dependent variables, and our main analytical approaches.[5] We aimed to determine how introducing this AI into the tasks of highly skilled knowledge workers might augment, disrupt, or influence their performance and workflow. Details about the experimental structure are reported in Online Appendix A.

The study included 758 BCG consultants (~7% of their individual-contributor consulting workforce) performing knowledge work. This is an elite group of knowledge workers with degrees from top universities around the world, such as Harvard, Yale, and Oxford. Tables 1 and 2 display a breakdown of demographic characteristics and experimental performance metrics for subjects across conditions. Table 3 reports the most common educational majors among our participants. Each participant completed the initial survey and was randomly assigned to one of two distinct arms of the experiment, each involving a unique type of task, with no overlap between the groups.[6]

Approximately half of the participants (385 subjects) tackled a series of tasks that prompted them to conceptualize and develop new products, focusing on aspects such as creativity, analytical skills, persuasiveness, and writing. The other half (373 subjects) engaged in business problem-solving tasks using quantitative data, customer and company interviews, and a persuasive writing component. Both sets of tasks were developed to be realistic: that is, representative of the tasks that these workers engaged in regularly. To achieve this, the tasks were designed by BCG professionals from the relevant sectors. Figure 2 provides a visual representation of our experimental design across the two arms of the experiment.

The two arms of the experiment followed a consistent structure. Initially, participants undertook a task without

Table 1. Summary Statistics by Demographics and Experimental Conditions

	GPT + overview (A)			GPT only (B)			Control (C)			*p*-values
Variable	*N*	Mean	SD	*N*	Mean	SD	*N*	Mean	SD	(A = B, A = C, B = C)
Panel A: Inside the frontier										
Female	126	0.27	0.45	129	0.34	0.48	130	0.34	0.48	0.22, 0.23, 0.96
English	126	0.57	0.50	129	0.57	0.50	130	0.58	0.49	0.93, 0.83, 0.76
Tenure	126	0.56	0.50	129	0.53	0.50	130	0.52	0.50	0.56, 0.44, 0.85
Location	126	0.33	0.47	129	0.32	0.47	130	0.32	0.47	0.90, 0.86, 0.97
Tech Openness	126	0.59	0.49	129	0.65	0.48	130	0.61	0.49	0.30, 0.74, 0.47
Panel B: Outside the frontier										
Female	125	0.35	0.48	118	0.33	0.47	129	0.32	0.47	0.73, 0.57, 0.83
English	125	0.57	0.50	118	0.59	0.49	129	0.54	0.50	0.69, 0.69, 0.42
Tenure	125	0.56	0.50	118	0.50	0.50	129	0.55	0.50	0.35, 0.88, 0.43
Location	125	0.29	0.45	118	0.31	0.46	129	0.33	0.47	0.77, 0.52, 0.73
Tech Openness	125	0.57	0.50	118	0.64	0.48	129	0.63	0.49	0.28, 0.33, 0.90
Panel C: Inside = outside; *p*-values										
Female	0.16			0.86			0.72			
English	0.96			0.67			0.50			
Tenure	0.96			0.67			0.57			
Location	0.52			0.83			0.86			
Tech Openness	0.76			0.80			0.74			

Notes. This table provides a breakdown of demographic characteristics for subjects across conditions. For each category, we report total sample size (*N*), mean values, and standard deviations (SDs) to detail the distribution and variability of the data. All *p*-values are from two-sided *t*-tests of equality of means. None are significant at the 5% level.

the aid of AI, establishing a baseline for performance and enabling within-subject analyses.[7] Following this, participants were randomly assigned to one of three conditions to assess the influence of AI on their tasks, with these conditions kept consistent across both arms. The first group (a control condition) proceeded without AI support. The second group ("GPT only") had the assistance of an AI tool based on GPT-4. The third group ("GPT + overview") utilized the same AI tool but also received a supplementary prompt engineering overview, which increased their familiarity with the AI tool. These overview materials included instructional videos and documents that outlined and illustrated effective usage strategies.

Following the baseline assessment, subjects assigned to the various AI conditions had access to a custom-built company platform for every experimental task. This platform, developed using OpenAI's application programming interface (API), facilitated an interactive experience with OpenAI's GPT-4 that mirrored the dynamics of ChatGPT. Although the functionality resembled that of ChatGPT, the platform enabled the collection of all data logs, including participant prompts and the AI's corresponding responses, providing comprehensive insight 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. The same model was used for both sets of tasks.

Other than thematic differences, the two sets of tasks differed in another key way. Although both were designed to be of comparable complexity and realistically represent real-world consulting responsibilities, tasks in

Table 2. Summary Statistics—Performance

	GPT + overview			GPT only			Control		
Variable	*N*	Mean	SD	*N*	Mean	SD	*N*	Mean	SD
Inside the frontier: Performance									
Quality	126	5.86	0.52	129	5.68	0.61	130	4.38	0.45
Completion	126	0.93	0.15	129	0.91	0.16	130	0.82	0.17
Timing	126	3,894	1,281	129	3,635	1,494	130	5,023	827
Outside the frontier: Performance									
Correctness	125	0.60	0.49	119	0.71	0.46	129	0.84	0.36
Timing	125	1,571	920	119	1,853	995	129	2,260	917
Subjective coherence quality	125	7.33	1.62	118	6.90	2.34	129	5.86	2.20

Notes. This table provides a breakdown of experimental performance metrics for subjects across conditions. For each category, we report total sample size (*N*), mean values, and standard deviations (SDs) to detail the distribution and variability of the data.

Table 3. Most Frequently Reported Majors—Bachelor's Degree

Major	Count	Subjects, %
Economics	104	14.10
Finance	42	5.70
Mechanical engineering	41	5.60
Business administration	41	5.60
Chemical engineering	31	4.20
Business	23	3.10
Engineering	17	2.30
Industrial engineering	17	2.30
Civil engineering	15	2.00
Political science	14	1.90
Computer science	12	1.60
Politics	11	1.50
Chemistry	11	1.50
Mathematics	10	1.40

Notes. This table presents the distribution of the most frequently reported majors among our subjects' bachelor's degrees. The majors are listed in descending order of prevalence, with the count and percentage of subjects for each major provided. The percentages reflect the proportions of subjects within our study sample who reported each major as their field of study during their undergraduate education.

the first experiment were designed to be within the potential technological frontier of GPT-4. In contrast, the second experiment was deliberately designed so that human interaction and guidance were necessary to perform successfully when using GPT-4. The second experiment also contained a persuasive component that fell within the frontier.[8] We conducted pretests for both sets of tasks, iterating several times to ensure that they were appropriately situated on either side of the frontier before initiating the experiment.

Participant engagement and effort were encouraged through performance-based incentives designed to reflect BCG's performance evaluation system.[9] As outlined in our preregistration, all participants who demonstrated honest effort received "office contribution" recognition—a designation with direct financial implications for annual bonuses. The bottom 5% of performers forfeited this recognition, whereas the top 20% received additional acknowledgment to their supervisors noting their exceptional performance, potentially further impacting bonus calculations and advancement prospects.

4. Results

4.1. Quality and Productivity Booster—Inside the Frontier

The inside-the-frontier experiment focused on creative product innovation and development. We defined inside-the-frontier tasks by pretesting them to confirm that generative AI could produce useful, high-quality outputs in the BCG client-consulting context from straightforward prompts. 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.

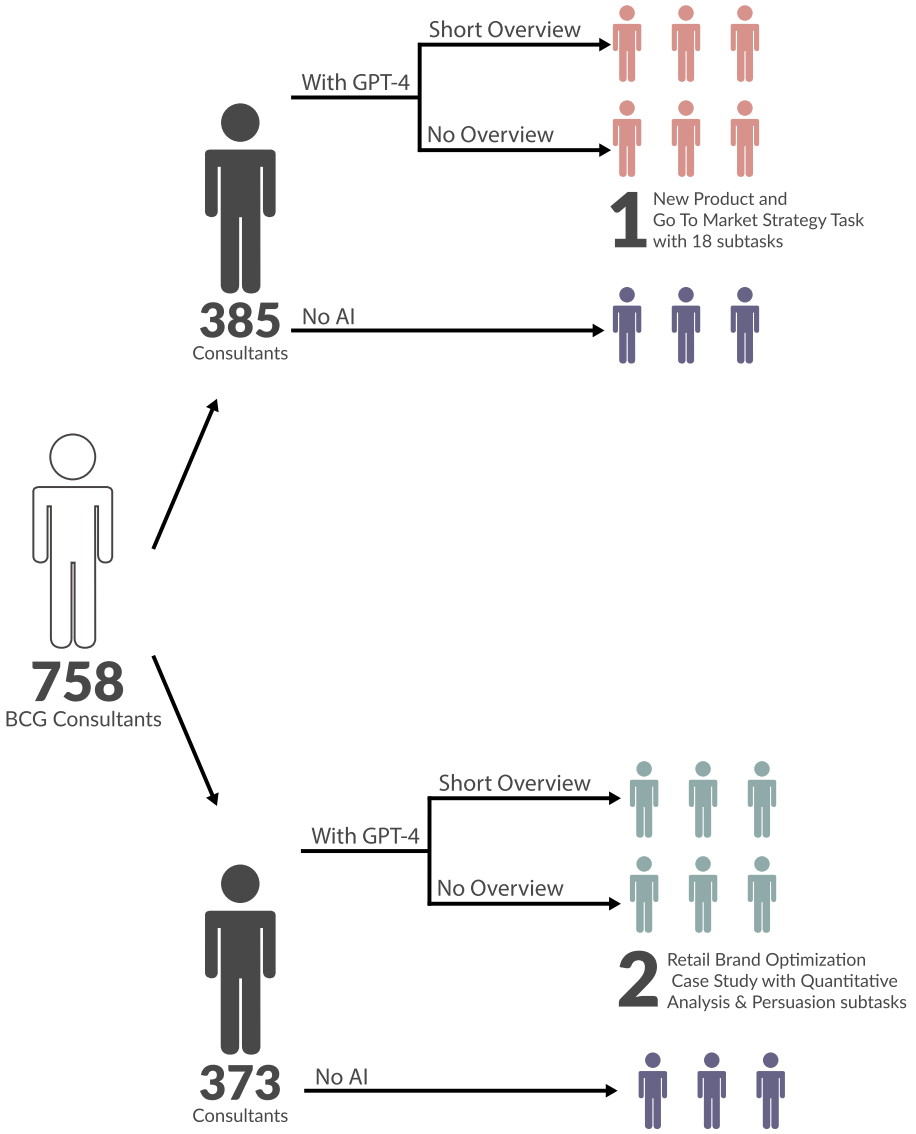
In this experimental task, participants were tasked with conceptualizing a footwear idea for niche markets and delineating every step involved from prototype description to market segmentation to entering the market. An executive from a leading global footwear company verified that task design covered the entire process that their company typically undergoes from ideation to product launch.[10] Participants responded to a total of 18 tasks (or as many as they could within the given time[11]). These tasks spanned various domains, categorizable 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 the Online Appendix.

The primary outcome variable in the experiment is the quality of the subjects' responses. To quantify this quality, we employed a set of human graders to evaluate each question that participants did not leave unanswered.[12] Each response was independently evaluated by three human graders. We then calculated the mean grade assigned by humans to each response. This gave us 18 dependent variables (one per question). We subsequently averaged these scores across all questions to derive a composite "quality" score for use in our main analyses. As an additional assessment, we utilized GPT-4 to independently score subjects' responses.[13] Similarly to the human grades, we produced a score for each of the 18 questions and then, a composite "quality (GPT)" score.

Figure 3 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. Comparing the dashed lines and the overall distributions of the experimental conditions in Figure 3 clearly reveals the significant performance enhancements associated with the use of GPT-4, with the GPT-only and GPT + overview groups both recording superior performances to the control group.

Table 4 presents the results of the analyses using response quality as the dependent variable and highlights the performance implications of using AI.[14] Columns (1), (2), and (3) in Table 4 utilize human-generated grades as the dependent variable, whereas column (4) in Table 4 uses the composite grades generated by GPT-4. Across all specifications, both treatments—GPT + overview and GPT only—demonstrate positive effects.

Figure 2. Experimental Structure



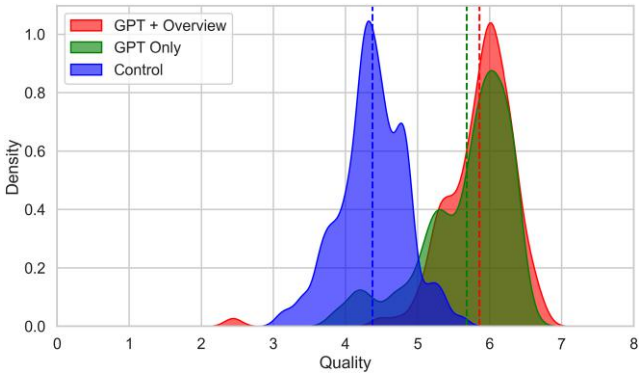
Note. This figure illustrates the experimental setup utilized in our study.

In column (1) in Table 4, GPT + overview leads to a 1.48 increase in scores over the control mean of 4.37, representing a 33.9% increase, with GPT only producing a 1.31 (or 29.9%) increase. Notably, columns (2), (3), and (4) in Table 4 incorporate performance metrics from the assessment task, and the treatment coefficients that they report remain very consistent.[15] Column (4) in Table 4 uses GPT scores as the dependent variable and shows coefficients of 1.35 above the control group for the GPT + overview treatment and 1.22 above the control group for the GPT-only treatment, corresponding to 19% and 17.1% improvements in performance, respectively.[16]

The beneficial impacts of using AI remain consistent across all our specifications.[17] As a robustness check, we isolated the quality of responses within each category (creativity, analytical thinking, writing proficiency, and persuasiveness) and applied the same model detailed in column (1) in Table 4. Each of the regressions demonstrates a significant effect of introducing AI on knowledge workers' performance, with very sizable results for all question categories. We verified that the treatment effects are not driven by superficial spelling or grammar improvements. Adding controls for the number and rate of such errors leaves all main coefficients virtually unchanged (see Online Appendix H). Finally, we tested the reliability of these findings by replicating our analyses on a new set of grades for the ideas generated in the first question. Our results are consistent as discussed in Section 5.

Another key observation from Table 4 is the differential impact of the two AI treatments. Specifically, the GPT + overview treatment consistently exhibits a more pronounced positive effect compared with the

Figure 3. Inside the Frontier—Performance Distribution



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, and blue for subjects in the control condition). Each subject's responses were rated by MBA graders on creativity, analytical thinking, writing proficiency, or persuasiveness depending on the specific question. Graders used a 1–10 scale, where higher scores indicate stronger performance. The *x* axis thus represents the average of these ratings across all applicable tasks for each subject, and the *y* axis shows the kernel density estimate of that distribution.

GPT-only treatment. The bottom of Table 4 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 for enhancing the efficacy of AI assistance, with Table 4 also highlighting various other factors, including gender, native English proficiency, tenure, location, and openness to new technology, none of which changed the significance of the impact of AI on performance.[18]

Table 5 presents the results related to the percentage of tasks completed by subjects, this analysis's dependent variable. Across columns (1), (2), and (3) in Table 5, the two treatment conditions (GPT + overview and GPT) only demonstrate a positive effect on task completion. On average, these coefficients indicate a 12.2% increase in completion rates. Notably, directly comparing the two AI treatments reveals that the difference in their impacts is not statistically significant. The control group completed on average 82% of their tasks; those in the GPT + overview group completed about 93%; and those in the GPT-only group completed about 91%. Column (2) in Table 5 incorporates the performance metric from the assessment, and column (3) in Table 5 extends the analysis by including the same set of controls as in Table 4. The coefficients suggest that the integration of AI tools substantially enhances the rate of task completion while also increasing quality.

Figure 4 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 (2024).[19] 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. Note that the top-half-skill performers also received a significant boost, although not as much as the bottom-half-skill performers.

Our findings gain additional perspective through comparison with recent research in this area. Brynjolfsson et al. (2025) explore customer support roles,

Table 4. Inside the Frontier—Quality

Variable	(1) *Quality*	(2) *Quality*	(3) *Quality*	(4) *Quality (GPT)*
GPT + overview	1.482***	1.490***	1.506***	1.349***
	(0.061)	(0.057)	(0.061)	(0.058)
GPT only	1.307***	1.329***	1.337***	1.216***
	(0.067)	(0.065)	(0.065)	(0.059)
Assessment		0.174***	0.171***	
		(0.058)	(0.060)	
Assessment (GPT)				0.167**
				(0.070)
Female		−0.153***	−0.156***	−0.042
		(0.057)	(0.057)	(0.049)
English Native		0.066	0.070	0.097*
		(0.061)	(0.063)	(0.055)
Low Tenure		0.045	0.045	0.018
		(0.052)	(0.054)	(0.048)
Location		−0.029	−0.029	0.050
		(0.067)	(0.069)	(0.061)
Tech Openness		0.068	0.067	0.065
		(0.056)	(0.063)	(0.058)
AI-exposure controls			Yes	Yes
R^2	0.610	0.642	0.646	0.639
GPT = GPT + overview	0.014	0.025	0.021	0.049
Control mean	4.375	4.375	4.375	7.098
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 as their dependent variable; each response was independently graded by three human evaluators. In contrast, column (4) uses the average response quality in the experimental task as determined by GPT. Columns (2) and (3) incorporate the average response quality from the assessment task as graded by human evaluators, whereas column (4) utilizes the GPT-evaluated metric. Columns (3) and (4) add seven presurvey AI-exposure controls: perceived current task automation (0%–100%); belief that one's job could be automated; Likert-scale items on familiarity with text-generation tools, image-generation tools, prompt engineering, and large language model mechanics; and a five-point measure of prior ChatGPT usage frequency (never to daily). Higher scores on all seven variables indicate greater AI familiarity or usage. The bottom of the table displays *p*-values from an *F*-test comparing the effects of receiving the GPT + overview treatment vs. the GPT-only treatment. All regressions include robust standard errors.

*$p < 0.10$; **$p < 0.05$; ***$p < 0.01$.

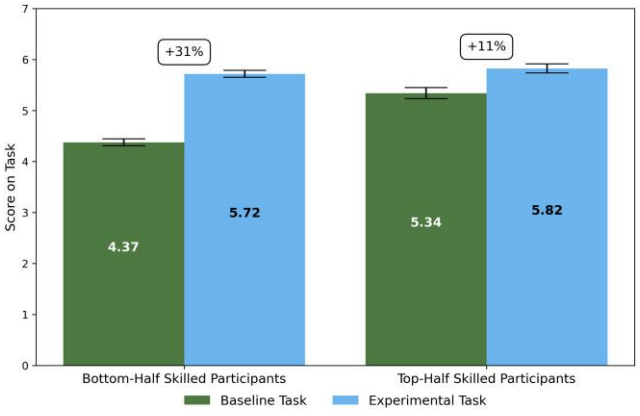
Table 5. Inside the Frontier—Completion

Variable	(1) *Percentage Completion*	(2) *Percentage Completion*	(3) *Percentage Completion*
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)
Percentage Completion (Assess)			0.367***
			(0.052)
AI-exposure controls			Yes
R^2	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 subtasks (of 18) that subjects successfully completed. Columns (2) and (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 seven presurvey AI-exposure controls: perceived current task automation (0%–100%); belief that one's job could be automated; Likert-scale items on familiarity with text-generation tools, image-generation tools, prompt engineering, and large language model mechanics; and a five-point measure of prior ChatGPT usage frequency (never to daily). Higher scores on all seven variables indicate greater AI familiarity or usage. The bottom of the table displays *p*-values from an *F*-test comparing the effects of receiving the GPT + overview treatment vs. the GPT-only treatment. All regressions include robust standard errors.
*$p < 0.10$; ***$p < 0.01$.

identifying productivity gains primarily among lower-skilled workers, with negligible effects observed among their higher-skilled counterparts. Choi and Schwarcz (2024) use law school examinations and find that students at the bottom of the performance distribution benefited enormously from AI assistance, whereas higher-skilled students saw performance declines. Conversely, Otis et al. (2024b) report mixed outcomes from the deployment of generative AI among small-scale entrepreneurs in Kenya. Although high-performing individuals reap the benefits of AI-generated advice, low-performing individuals tend to fare worse when following AI suggestions. In our study, centered on high-end knowledge work, we observe a universal benefit for workers involved in complex, high-level tasks, particularly for tasks within the AI's capability frontier. Within this group of elite professionals, relatively lower-skilled individuals gained the most, suggesting that generative AI may serve as a significant equalizer

Figure 4. Inside the Frontier—Bottom-Half Skills and Top-Half Skills



Notes. This figure shows the performance of AI-treated subjects, comparing those in the bottom half with those in the top half of the assessment task performance. Assessment task results are in green, and experimental task results are in blue.

in the realm of high-end knowledge work, narrowing the performance gap between different skill levels; however, those who were above average also benefited meaningfully through the use of the AI system. Together, these insights highlight the significance of task-specific and contextual considerations for effectively leveraging generative AI.

For tasks within the frontier of AI capacity, we prevented subjects from completing the experiment before the allotted time was over. To achieve this, the final question was especially long, asking participants 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, although 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 6 uses this *Timing* variable as the dependent variable. The GPT + overview treatment reduces time spent on the first 17 questions by 1,129 seconds (18.8 minutes or 22.5% faster than the control), and the GPT-only treatment reduces time spent on the first 17 questions by 1,388 seconds (23.1 minutes or 27.6% faster than the control).

Our results reveal significant effects, underscoring the prowess of AI even on tasks traditionally executed by highly skilled knowledge workers. Not only did using AI increase the number of subtasks completed by more than 12%, it also enhanced the quality of the responses by an average of 32%. 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 major performance benefits.

4.2. Quality Disruptor—Outside the Frontier

In refining the task within the frontier and recognizing the substantial quality and productivity gains enabled by integrating AI, we sought to develop a task that participants could not have AI easily complete by simply copying and pasting our instructions as a prompt. We designed the beyond-the-frontier task in collaboration with BCG, using as a starting point the type of business cases that BCG uses for its highly competitive job interviews. These cases are used to gauge the skill and knowledge fit of these workers in the context of tasks of the kind that they would face in their work with the company. This meant creating a task at which knowledge workers would excel but AI would struggle, at least without extensive guidance and human intervention. We settled on a task based on an existing business case that used data from a spreadsheet and a file presenting interviews with company insiders, which were adjusted, anonymized, 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. Although the spreadsheet data alone were designed to seem to be comprehensive, a careful review of the interview notes revealed crucial details. When considered in totality, this information led to an incorrect conclusion regarding 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

Table 6. Inside the Frontier—Timing

Variable	(1) *Timing*	(2) *Timing*	(3) *Timing*
GPT + overview	−1,129.143***	−1,105.622***	−1,094.640***
	(135.181)	(136.614)	(129.681)
GPT only	−1,388.415***	−1,356.364***	−1,351.998***
	(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)
AI-exposure controls			Yes
R^2	0.196	0.206	0.345
GPT = GPT + overview	0.137	0.155	0.124
Control mean	5,023	5,023	5,023
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 number of seconds that subjects took to reach the final question (question 18). Each column displays the results of a distinct linear regression model. Columns (2) and (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 seven presurvey AI-exposure controls: perceived current task automation (0%–100%); belief that one's job could be automated; Likert-scale items on familiarity with text-generation tools, image-generation tools, prompt engineering, and large language model mechanics; and a five-point measure of prior ChatGPT usage frequency (never to daily). Higher scores on all seven variables indicate greater AI familiarity or usage. The bottom of the table displays *p*-values from an *F*-test comparing the effects of receiving the GPT + overview treatment vs. the GPT-only treatment. All regressions include robust standard errors.

*$p < 0.10$; ***$p < 0.01$.

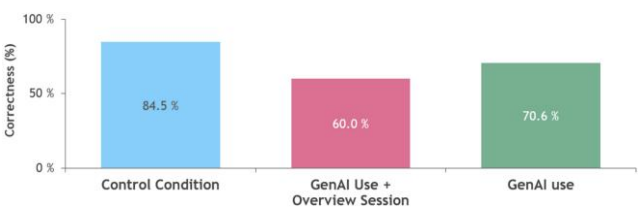
worked on the assessment task, which required them to analyze the company's sales distribution channel performance. Using insights from interviews and financial data, participants were asked to provide information and informed advice to the chief executive officer (CEO). Their recommendations were to pinpoint which channel held the most potential for growth.

Upon completing their assessment task, participants moved to the main experimental task. The focus transitioned from the examination of the company's distribution channels to brand analysis, with subjects having to analyze the company's brand performance. Similarly to the assessment task, participants used insights from interviews and financial data to provide recommendations to the CEO. Their recommendations were designed to pinpoint which brand held the most potential for growth. Participants were also expected to suggest actions to improve the chosen brand, regardless of the exact brand that they had chosen. Details of these tasks are reported in Online Appendix B.

For this task outside the frontier, our primary evaluation metric is "correctness," which is represented as a binary variable, where a value of one indicates that subjects provided an accurate recommendation and zero signifies otherwise.[20] Figure 5 visualizes the correctness percentages for the 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, but the AI conditions scored at 60% and 70.6% (an average decrease of 19 percentage points when combining the AI treatment conditions and comparing them with the control condition).

Table 7 captures 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.[21] Both AI treatments—GPT + overview and GPT only—show a significant negative impact, with the GPT + overview group recording a more pronounced decrease (24.5% versus 13.9%). Column (2) in Table 7 introduces the performance metric from the assessment, whereas column (3) in Table 7 further refines the analysis by incorporating the same set of controls as in Tables 4 and 5. When directly comparing the two treatments, the difference in their impacts is statistically significant at the 10% level across specifications.

Figure 5. Outside the Frontier—Performance



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.

Table 7. Outside the Frontier—Correctness

Variable	(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)
AI-exposure controls			Yes
R^2	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. The dependent variable across all columns is a binary indicator for whether the subject provided the correct strategic recommendation (1 = correct, 0 = incorrect). Columns (2) and (3) include a binary correctness metric from the assessment task as graded by human evaluators and a set of controls. Column (3) additionally includes seven presurvey AI-exposure controls: perceived current task automation (0%–100%); belief that one's job could be automated; Likert-scale items on familiarity with text-generation tools, image-generation tools, prompt engineering, and large language model mechanics; and a five-point measure of prior ChatGPT usage frequency (never to daily). Higher scores on all seven variables indicate greater AI familiarity or usage. The bottom of the table displays *p*-values from an *F*-test comparing the effects of receiving the GPT + overview treatment vs. the GPT-only treatment. All regressions include robust standard errors.

*$p < 0.10$; **$p < 0.05$; ***$p < 0.01$.

Table 8 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*, which represents how long subjects spent on the task calculated in seconds.[22] Column (2) in Table 8 further refines the analysis by incorporating the same set of controls as in Table 4. The findings for both AI treatments—GPT + overview and GPT only—document a reduction in the time spent: more than 11 minutes less for GPT + overview (a 30% decrease in time spent compared with the control mean) and more than 6 minutes less for GPT only (a decrease of 18% in time spent compared with the control mean). Comparing the two

Table 8. Outside the Frontier—Timing

Variable	(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)
AI-exposure controls			Yes
R^2	0.085	0.407	0.414
GPT = GPT + overview	0.022	0.000	0.000
Control mean	2,260	2,260	2,260
Observations	373	373	373

Notes. This table examines the effects of introducing GPT-4 on the time that subjects spent completing the experimental task outside the frontier. Each column displays the results of a distinct linear regression model. The dependent variable across all columns is *Timing*, which is defined as the number of seconds that subjects spent on the task before submission. Columns (2) and (3) include the timing of the assessment task and a set of controls. Column (3) additionally includes seven presurvey AI-exposure controls: perceived current task automation (0%–100%); belief that one's job could be automated; Likert-scale items on familiarity with text-generation tools, image-generation tools, prompt engineering, and large language model mechanics; and a five-point measure of prior ChatGPT usage frequency (never to daily). Higher scores on all seven variables indicate greater AI familiarity or usage. The bottom of the table displays *p*-values from an *F*-test comparing the effects of receiving the GPT + overview treatment vs. the GPT-only treatment. All regressions include robust standard errors.

***$p < 0.01$.

coefficients reveals that GPT + overview recorded a more substantial—and statistically significant—decrease in time spent compared with the GPT-only group.

Table 9 further examines the persuasiveness and internal coherence of recommendations in the outside-the-frontier experiment, independent of whether the underlying solution was correct. The dependent variable, *Subjective coherence quality*, captures how well arguments were structured and supported as assessed on a 1–10 scale. Two independent sets of graders evaluated each response: BCG consultants not involved in the experiment and business school students with prior grading experience. Both groups followed a rubric developed by the authors based on workplace evaluation practices (see Online Appendix E for the full rubric). Importantly, graders were not informed of the correct strategic solution; their task was to assess the clarity, persuasiveness, and coherence of reasoning. Our subjective coherence quality metric uses the average of these two grades. Column (1) in Table 9 shows that the treatment GPT + overview leads to a 1.47-point-greater score (a 25.1% increase over the control mean), with GPT only increasing the score by 1.05 points (a 17.9% increase over the control mean). Across all specifications, subjects using AI (whether GPT + overview or GPT only) consistently outperformed those not using AI in terms of subjective coherence quality, regardless of the correctness of their answer. When we control for various factors in columns (2), (3), and (4) in Table 9, the positive impact of AI remains robust. Column (3) in Table 9 repeats the specification in column (2) in Table 9 for the subset of participants with incorrect answers, and column (4) in Table 9 does the same for those with correct answers. In both instances, the effects of using AI are positive.

Figure 6 illustrates subjective coherence quality conditional on whether participants ultimately solved the problem correctly. Across both the "correct" and "incorrect" subgroups, AI-supported participants produced recommendations that were more coherent and persuasive, underscoring that AI improves presentation and argumentation even when the underlying analysis is flawed. This finding underscores the multifaceted ways in which AI can influence performance in the workflow of highly skilled professionals.

Table 10 summarizes our results and presents the test statistics, *p*-values, and standardized effect sizes comparing the treatment conditions with the control group across our key outcome measures. Depending on the outcome variables, we use different statistical tests (*t*-test, chi-squared test, and MWU), which consistently show statistical significance across all our main results. Notably, all of these results remain statistically significant after applying the Bonferroni correction to adjust the significance threshold for multiple comparisons. We are additionally reporting standardized effect sizes (Cohen *d*, Cohen *h*, and rank biserial) to convey the magnitude of any significant differences. The effect sizes for our main findings range from moderate to large, with most falling in the medium to large range, underscoring the substantive impact of our interventions.

5. Robustness Checks with Shoe Design Experts

5.1. Individual Shoe Design Ideas

To further test the reliability of the findings presented in the main paper, we replicate our analysis on a new set of grading data. Specifically, we carry out this replication using the grades that artist evaluators assigned to shoe design ideas from the first question of the field experiment.

Table 9. Outside the Frontier—Subjective Coherence Quality

Variable	(1) *Subjective Coherence Quality*	(2) *Subjective Coherence Quality*	(3) *Subjective Coherence Quality*	(4) *Subjective Coherence Quality*
GPT + overview	1.475***	1.330***	1.448**	1.493***
	(0.242)	(0.229)	(0.718)	(0.248)
GPT only	1.046***	0.926***	1.777**	0.724**
	(0.289)	(0.263)	(0.805)	(0.285)
Assessment—Subjective Coherence Quality		0.370***	0.337***	0.401***
		(0.053)	(0.125)	(0.056)
Female		−0.437*	−0.884	−0.224
		(0.238)	(0.569)	(0.255)
English Native		−0.102	−0.655	0.223
		(0.218)	(0.534)	(0.238)
Low Tenure		−0.038	−0.610	0.229
		(0.211)	(0.556)	(0.218)
Location		0.297	−0.009	0.465*
		(0.231)	(0.515)	(0.245)
Tech Openness		0.219	−0.150	0.344
		(0.220)	(0.508)	(0.229)
AI-exposure controls		Yes	Yes	Yes
R^2	0.085	0.246	0.222	0.328
GPT = GPT + overview	0.098	0.103	0.531	0.007
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 subjective coherence quality of the recommendations provided in the experimental task outside the frontier. Each column displays the results of a distinct linear regression model. The dependent variable across all columns is *Subjective Coherence Quality*, which is measured on a 1–10 scale. Column (2) includes the *Subjective Coherence Quality* of the assessment task, a set of controls, and seven presurvey AI-exposure controls: perceived current task automation (0%–100%); belief that one's job could be automated; Likert-scale items on familiarity with text-generation tools, image-generation tools, prompt engineering, and large language model mechanics; and a five-point measure of prior ChatGPT usage frequency (never to daily). Higher scores on all seven variables indicate greater AI familiarity or usage. 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 vs. the GPT-only treatment. All regressions include robust standard errors.

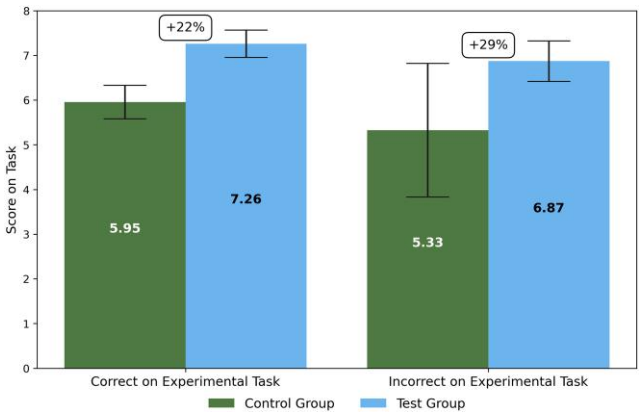
*$p < 0.10$; **$p < 0.05$; ***$p < 0.01$.

5.1.1. Data Collection. The data set of shoe design ideas is derived from the first question of the "inside-the-frontier" task. Participants were asked to "generate ideas for a new shoe aimed at a specific market or sport that is underserved." To conduct this analysis, we disaggregated each list into individual shoe design ideas, anonymized participant identifications, and presented the ideas to our graders in a randomized order.[23] Graders were instructed to rate the creativity of each idea on a scale from 1 to 10, disregarding grammatical and spelling issues. The graders were not informed about the study's purpose or that some ideas were generated with the help of AI.

Five graders evaluated ideas. They were selected from a pool of applicants through resume screening and interviews, and they were filtered based on their relevant experience. All evaluators were students majoring in product design or industrial design at top programs with demonstrated experience in footwear design. Three graders evaluated all shoe design ideas, and two of the five graders only evaluated half of the ideas. To account for this constraint, each of these two graders was assigned a mutually exclusive, randomly selected subset comprising 50% of the data set. In total, every shoe design idea received four different evaluations by four graders.

These new grading data differ from the initial data in four important ways. First, instead of distributing 18 grades across 18 different questions as in the original "inside-the-frontier" analysis, this approach leverages at least 10 independent grades on just 1 question, thereby limiting the analysis's scope but preserving its robustness. Second, the grades are based on creativity rather than general quality, allowing us to directly test the notion that AI tools enhance human creativity. Third, the evaluators have extensive experience in footwear design rather than a general business background, suggesting that their creativity scores hold greater validity within this specific field. Fourth, this data set uses four different evaluations instead of three.

Figure 6. Outside the Frontier—Subjective Coherence 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 subjective coherence quality of the control group, whereas the blue bars indicate the average subjective coherence quality of the treatment groups. The y axis denotes the average scores, ranging from 1 to 10.

5.1.2. Results. The replication results are shown in Table 11.[24] As expected, both treatments—GPT + overview and GPT only—demonstrate positive effects across all specifications. As with the original analysis, the GPT + overview treatment exhibits a more pronounced positive effect than GPT only, although this difference is not statistically significant in any of the specifications. The magnitude of the effect is lower than in the main analysis.[25] In column (1) in Table 11, the GPT + overview and GPT-only treatment groups had creativity scores that were 14% and 11.7% higher than the control group, respectively, as compared with 33.9% and 29.9% in the main analysis, respectively.

5.2. Full Replication with Shoe Design Experts

To further validate our findings and address feedback on the subjectivity of our evaluation metrics, we expanded our robustness checks with the shoe design experts. Although the earlier analysis focused on the creativity of ideas from the first question, this expanded analysis covers all 18 questions of the "inside-the-frontier" task.

5.2.1. Methods. We worked with three independent shoe design experts among those who evaluated ideas in Section 5.1 to evaluate responses for all 18 questions. To ensure high-quality and thorough assessments, each question for every participant was graded by two of the three experts.[26] The experts were compensated on an hourly basis, and just like with our original graders, they were given unlimited time to complete their evaluations, removing any time pressure that might compromise the quality of their assessments. As with the initial analysis, graders were blind to the experimental conditions.

5.2.2. Results. The results from this expanded expert evaluation are presented in Table 12. The findings are entirely consistent with the results reported in the main paper and in fact, demonstrate even stronger positive

Table 10. Test Statistics and Effect Sizes—Treatments vs. Control

		Test statistic			Effect size	
Variable	Treatment effect	Name	Value	p-value	Name	Value
Panel A: GPT only vs. control group						
Inside the frontier						
Quality	1.307	MWU	15,674	<0.000	Rank biserial	0.869
Completion	0.090	t-test	4.349	<0.000	Cohen d	0.528
Timing	−1,388.415	t-test	−9.244	<0.000	Cohen d	−1.679
Outside the frontier						
Correctness	−0.139	Chi squared	9.481	0.002	Cohen h	−0.337
Timing	−407.329	t-test	−3.343	0.001	Cohen d	−0.444
Subjective Coherence Quality	1.046	MWU	10,021	<0.000	Rank biserial	0.317
Panel B: GPT + overview vs. control group						
Inside the frontier						
Quality	1.483	MWU	16,079	<0.000	Rank biserial	0.963
Completion	0.111	t-test	5.583	<0.000	Cohen d	0.651
Timing	−1,129.143	t-test	−8.352	<0.000	Cohen d	−1.365
Outside the frontier						
Correctness	−0.245	Chi squared	17.874	<0.000	Cohen h	−0.560
Timing	−689.191	t-test	−5.980	<0.000	Cohen d	−0.752
Subjective Coherence Quality	1.475	MWU	11,400	<0.000	Rank biserial	0.414

Notes. This table shows test statistics and effect sizes for the treated conditions with respect to the control group. For rows with an ordinal outcome variable, a Mann–Whitney U test (MWU) is used for the test statistic, and rank biserial is used for the effect size. For rows with a binary outcome variable, a chi-squared test is used for the test statistic, and Cohen h is used for the effect size. All other rows use t-tests and Cohen d.

Table 11. Inside the Frontier—Quality (Individual Ideas)

Variable	(1) *Quality*	(2) *Quality*	(3) *Quality*
GPT + overview	0.566***	0.584***	0.581***
	(0.143)	(0.138)	(0.131)
GPT only	0.471***	0.491***	0.483***
	(0.077)	(0.073)	(0.073)
Assessment		0.075***	0.081***
		(0.024)	(0.023)
Female		−0.002	−0.020
		(0.067)	(0.069)
English Native		−0.007	−0.007
		(0.058)	(0.060)
Low Tenure		−0.175***	−0.177***
		(0.036)	(0.036)
Location		0.049	0.042
		(0.065)	(0.065)
Tech Openness		−0.073**	−0.001
		(0.031)	(0.037)
AI-exposure controls			Yes
R^2	0.015	0.018	0.019
GPT = GPT + overview	0.290	0.302	0.216
Control mean	4.043	4.043	4.043
Observations	15,404	15,404	15,404

Notes. This table reports results for the inside-the-frontier quality outcome using only responses to question 1. The dependent variable is overall quality of the idea measured on a standardized scale and evaluated by human graders. Each column displays the results of a distinct linear regression model. Columns (2) and (3) incorporate the average response quality from the assessment task and a set of controls. Column (3) includes seven presurvey AI-exposure controls. The bottom of the table displays *p*-values from an *F*-test comparing the effects of receiving the GPT + overview treatment vs. the GPT-only treatment. All regressions include robust standard errors.
**$p < 0.05$; ***$p < 0.01$.

effects of AI on performance. Across all specifications, both the GPT + overview and GPT-only treatments show a significant and substantial improvement in the quality of work produced by participants. The effect sizes are larger than those observed in the main analysis, reinforcing our conclusion that AI assistance enhances the quality of knowledge work on tasks inside the frontier. These positive effects remain statistically significant when the analysis is disaggregated by question type (creativity, analytical thinking, writing proficiency, and persuasiveness).

These robustness checks support our main findings. Additionally, this convergence across diverse evaluator types and methodologies provides strong evidence against systematic evaluator bias and the hypothesis that graders are systematically fooled by AI-generated content.

6. Discussion

In this paper, we study the use of AI with a jagged capability frontier by humans performing high-end knowledge work. Using a preregistered, randomized laboratory-in-the-field experiment, we test the potential dual role of AI in augmenting the efforts of knowledge workers, recognizing that it functions as both a booster—enhancing efficiency and productivity on tasks within its capability frontier—and a disruptor, which negatively impacts performance on tasks beyond its 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 nature of our experimental subjects. Specifically, we were able to recruit from a highly skilled and motivated population, with participants who engaged in tasks that were developed in collaboration with our partner company. These tasks closely mirrored their professional activities as knowledge workers, allowing us to assess AI's impact on tasks that approximate real-world workflows while recognizing that they do not fully capture every complexity of on-the-job work.

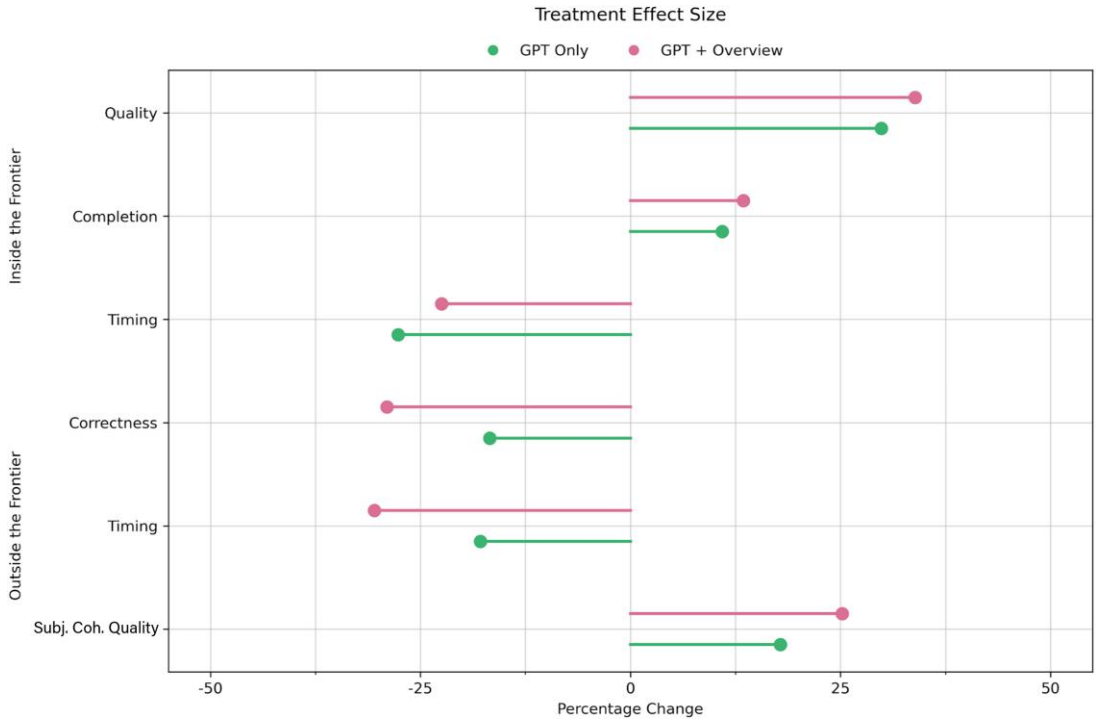
We found that the utility of AI can fluctuate across a professional's workflow, with some tasks falling inside the frontier, whereas others fall outside of it. Figure 7 displays the treatment effect sizes in percentage change for both inside and outside the frontier, showing the magnitude of our treatment effects conditions across all

Table 12. Inside the Frontier—Quality (Design Experts)

Variable	(1) *Quality*	(2) *Quality*	(3) *Quality*
GPT + overview	1.748***	1.753***	1.775***
	(0.067)	(0.063)	(0.066)
GPT only	1.613***	1.616***	1.629***
	(0.072)	(0.071)	(0.071)
Assessment		0.182**	0.175**
		(0.079)	(0.081)
Female		−0.121*	−0.130**
		(0.063)	(0.064)
English Native		0.103	0.113
		(0.068)	(0.071)
Low Tenure		0.039	0.038
		(0.057)	(0.058)
Location		0.054	0.050
		(0.073)	(0.075)
Tech Openness		0.078	0.074
		(0.060)	(0.067)
AI-exposure controls			Yes
R^2	0.663	0.685	0.690
GPT = GPT + overview	0.070	0.060	0.049
Control mean	4.498	4.498	4.498
Observations	385	385	385

Notes. This table replicates the inside-the-frontier quality analysis using evaluations provided by a new set of design expert graders. The dependent variable is overall quality measured on a standardized scale. The specifications parallel those in the main inside-the-frontier quality analysis. Robust standard errors are reported in parentheses.
*$p < 0.10$; **$p < 0.05$; ***$p < 0.01$.

Figure 7. Summary—Effect Sizes



Notes. The figure displays the treatment effect size (in percentage change) for several metrics in two experimental conditions: GPT only (green dots) and GPT + overview (red dots). Subj. Coh., subjective coherence.

of our key outcome measures. For tasks inside the frontier, our findings have substantial and positive performance implications. Across 18 realistic business tasks—ranging from creative to analytical tasks—AI significantly improved performance and quality for every model specification, increasing speed by more than 25%, performance by more than 30%, and task completion by more than 12%. However, for a task outside the frontier, subjects using AI were 19 percentage points less likely to produce correct solutions. Although this design highlights a single outside-the-frontier task—an inherent limitation—it underscores that even a single complex challenge beyond AI's current capabilities can disproportionately reduce overall performance if users rely too heavily on AI's output. These results suggest a need to understand the frontier's shape and position as well as how it is perceived by knowledge workers to better determine AI's influence on work.

Our design employs different evaluation methods by task type, directly mirroring how consulting work is assessed in practice. For writing-intensive activities inside the frontier, performance is evaluated through subjective rubrics, matching BCG's standard procedures because no single correct answer exists. Conversely, the strategy case outside the frontier features an objectively correct recommendation that is measured dichotomously and supplemented with separate persuasiveness assessments. This hybrid approach reflects again the incommensurability of knowledge work outputs. Nevertheless, subjective metrics may contain evaluation bias, representing a key limitation of our study and suggesting important avenues for developing additional objective performance proxies in future research.

Our findings should be interpreted within important boundary conditions regarding incentive structures. Our results apply most directly to knowledge work contexts where quality within allocated time frames is prioritized over speed optimization, such as is the case at BCG and similar professional service environments. In settings where efficiency gains translate directly into rewards or where explicit accuracy-speed trade-offs exist—such as manufacturing, customer service with call-time metrics, or piecework environments—the effects of AI deployment may differ substantially. Future research should investigate how the jagged frontier manifests across different incentive structures in knowledge work.

A further caveat concerns who actually chooses to adopt generative AI when its use is discretionary rather than strongly encouraged, such as in our setting. Recent evidence shows sizable heterogeneity; early adopters are disproportionately male and more technologically confident, and the resulting performance gains can widen pre-existing gender- and skill-gaps rather than close them (Otis et al. 2024a, Humlum and Vestergaard 2025, Blandin et al. 2026). If such adoption gaps persist, the technology could widen existing promotion

differentials. Additionally, beyond unequal gains, AI may substitute some workers entirely. Recent evidence from online labor markets indicates that generative AI has reduced both employment and earnings for freelancers in affected occupations (Hui et al. 2024), underscoring that AI's impact on work encompasses not only how effectively workers perform tasks but also, their continued employment prospects.

Although our analysis focuses on individuals, the scale of our experiment within a single organization surfaces implications for how firms might integrate AI into knowledge work. Our findings suggest that organizations should focus on the knowledge workflow and the tasks within it, and for each task, they should evaluate the value of using different configurations of humans and AI. This will require rethinking collaboration between humans and AI (Raisch and Krakowski 2021, Faraj and Leonardi 2022, Feuerriegel et al. 2022, Lebovitz et al. 2022, Anthony et al. 2023, Choudhary et al. 2023, Dell'Acqua et al. 2025), human training (Beane 2019, Cowgill et al. 2020, Kellogg et al. 2021, Gaessler and Piezunka 2023), how new roles will emerge and be created (Barrett et al. 2012, Sergeeva et al. 2020, Allen and Choudhury 2022, Kellogg 2022), new capabilities and strategies (Iansiti and Lakhani 2020), and new forms of organizing (Bailey et al. 2022, Beane and Leonardi 2025). In this context, assessing AI capabilities relative to tasks within workflows becomes especially important (Lebovitz et al. 2021, Eloundou et al. 2024).

This paper contributes to the emerging literature on the impact of AI on work by showing its nuanced impact on knowledge work. Studies such as Noy and Zhang (2023) have documented that providing AI assistance boosts productivity and efficiency among college-educated professionals. Our paper adds substantially to this finding in several key respects. First, our laboratory-in-the-field experiment conducted with management consultants offers a real-world setting that mirrors the complexities of actual professional workflows, lending high ecological validity to our results. Second, whereas Noy and Zhang (2023) focused primarily on writing tasks, we include a broader range of activities. Third, we designed a dual-task framework with tasks outside current AI capabilities. This "jagged frontier" approach reveals that although AI significantly improves performance on tasks within its capabilities, it can hinder performance on those requiring complex analysis or reasoning.

Other papers have illustrated AI's heterogeneous effects. For example, Brynjolfsson et al. (2025) have highlighted AI's varying impact based on skill levels, showing enhancements primarily among lower-skilled workers, whereas Otis et al. (2024b) have found no overall effect from AI usage but noted benefits disproportionately affecting higher-skilled individuals. Our study, the first to measure the impact of AI on knowledge workers in a setting resembling their standard work activities, addresses the tension of the contrasting results in the literature by demonstrating AI's both positive and negative effects on knowledge work. These enhancements show that our work not only confirms the productivity benefits of introducing AI but also, expands our understanding of AI's limitations and the best ways to integrate it in professional environments.

Our findings also connect to several related literatures. Prior work shows that AI systems tend to generate larger performance gains in ideation and content creation tasks than in judgment or decision-making tasks (e.g., Vaccaro et al. 2024). Although our findings are consistent with this pattern, they show that such divergent effects can arise within the same knowledge workflow and for the same professionals, complicating knowledge workers' and managers' ability to anticipate when AI assistance will be beneficial versus harmful. Our findings also relate to research on automation bias, which documents how human performance can deteriorate when automated systems provide imperfect guidance in decision-making contexts (e.g., Skitka et al. 1999, Buçinca et al. 2021). We highlight how such performance losses can emerge when AI systems perform exceptionally well on adjacent tasks, underscoring the challenge posed by AI's jagged capabilities.

Our findings offer multiple avenues for interpretation in the context of future implementations of human-AI collaboration. First, our results lend support to optimism about AI capabilities for important high-end knowledge work tasks. In our study, the human-AI combination proved surprisingly capable, demonstrating the potential of AI-aided professional work. However, we also show the challenges resulting from AI's jagged capabilities; experienced and incentivized knowledge professionals, engaged in tasks akin to some of their daily responsibilities, performed worse when given access to AI. This happened even though they did not need to rely on AI's output to complete the task. The jagged nature of the knowledge frontier suggests that in real-world settings, knowledge-intensive workflows will most likely straddle both sides, creating situations where knowledge workers—unaware of their position relative to AI's capabilities—perform worse as a result of becoming overly reliant on AI ("falling asleep at the wheel" (Dell'Acqua 2022)), a risk that is reinforced by the persuasive capabilities of generative AI (Randazzo et al. 2025a).

Since the release of our working paper in September 2023, the notion of AI as a jagged technology has been acknowledged by industry leaders and practitioners. Sundar Pichai, the CEO of Google, noted in an interview with Lex Fridman (Pichai 2025): "Have you heard AJI, the artificial jagged intelligence? Sometimes feels that way, both their progress and you see what they can do and then you can trivially find they make numerical

errors or counting R's in strawberry or something, which seems to trip up most models or whatever it is. So maybe we should throw that term in there. I feel like we are in the AJI phase where dramatic progress, some things don't work well, but overall you're seeing lots of progress." Similarly, Andrej Karpathy, one of the leading AI systems developers, emphasized the practical implications of this phenomenon, advising practitioners to "use LLMs for the tasks they are good at but be on a lookout for jagged edges, and keep a human in the loop" (Karpathy 2024). In a subsequent blog post, Karpathy (2025) elaborated on this concept: "As verifiable domains allow for [reinforcement learning from verifiable rewards], LLMs 'spike' in capability in the vicinity of these domains and overall display amusingly jagged performance characteristics—they are at the same time a genius polymath and a confused and cognitively challenged grade schooler, seconds away from getting tricked by a jailbreak to exfiltrate your data."

As the boundaries of AI capabilities continue to expand, human professionals may need to recalibrate their understanding of the frontier. The frontier is jagged, but it is not fixed, and any given task may move inside with the release of a new model. Overall, AI appears poised to significantly affect human cognition and problem-solving but in uneven ways. AI may lower—and raise—the cognitive costs associated with knowledge acquisition and application across tasks. Effectively leveraging AI in professional settings will require understanding how work is positioned relative to AI's jagged frontier, even while subject to quadratic and exponential increases in benchmark performance on a periodic basis.

Acknowledgments

The authors thank Corey Gelb-Bicknell and Annika Hildebrandt for insightful input and excellent research assistance. They thank Saud Almutairi, Michael Bervell, John Cheng, Pallavi Deshpande, Patrick Healy, Maxim Ledovskiy, John Kalil, Yogesh Kumaar, Kelly Kung, Rick Lacerda, Paula Marin Sariego, Quoc-Anh Nguyen, Rafael Noriega, Alejandro Ortega, Rahul Phanse, Nitya Rajgopal, Ogbemi Rewane, Anahita Sahu, Kyle Schirmann, Andrew Seo, Tanay Tiwari, Elliot Tobin, and Aaron Zheng for helpful research assistance. They also thank Kevin Dai for outstanding support with data and visualizations. Additionally, the authors thank MarcAntonio Awada, Charles Ayoubi, Maxime Courtaux, Clement Dumas, Dietmar Harhoff, Gaurav Jha, Jackie Lane, Jessie Li, Max Männig, Michael Menietti, Rachel Mural, Steven Randazzo, Zahra Rasouli, Esther Yoon, Leonid Zhukov, and David Zuluaga Martínez for helpful feedback. Seminar participants at Harvard, the Massachusetts Institute of Technology, Columbia, Stanford, the University of Toronto, INSEAD, Bocconi, Chicago Booth, Georgia Tech, Wharton, New York University, University of Southern California, Google, Microsoft, OpenAI, The Organisation for Economic Co-operation and Development, *Strategy Science*, and the Academy of Management provided helpful feedback. The authors are grateful for the guidance of Lamar Pierce and two referees at *Organization Science* in helping to improve this paper. K. R. Lakhani thanks Martha Wells, Anne Leckie, Iain Banks, and Alastair Reynolds for inspiring artificial intelligence futures. The authors used Poe, Claude, Manus, and ChatGPT for light copyediting and graphic creations. All errors are the authors' own. The first working paper of this study was released on September 18, 2023.

Author Contributions

FDA and KRL established the collaboration, designed the experiment, and oversaw its execution. FDA and KRL led the analyses, framing, and write-up. FC was instrumental in initiating the collaboration. EMcF contributed to variable construction and all analyses. EM contributed to the design, the prompt training and the framing. HL and KK led the qualitative component of the study and contributed to the framing of the paper. FC, LK and SR facilitated organizational access, enabled the execution of the experiment, and supported selected analyses. All coauthors contributed to the experimental design and to revisions of the manuscript.

Endnotes

[1] For example, lawyers in New York submitted legal briefs with six fictitious cases (Reuters). Moreover, greater algorithmic interpretability may lead to poorer decision-making performance (DeStefano et al. 2022).

[2] Although we describe our study as a laboratory-in-the-field experiment, our design extends beyond the standard paradigm in two key ways. (a) Participants were active members of their firm's workforce with real professional consequences at stake, and (b) the tasks were codeveloped with company leadership to capture core competencies evaluated in actual client work rather than standardized laboratory tasks.

[3] By extension, a knowledge worker is someone hired primarily for their ability to complete such knowledge-based tasks, relying on their intellectual capital to successfully carry out their work. These definitions align with knowledge work as described by Drucker (1959), Acemoglu and Autor (2011), and Aral et al. (2012).

[4] The project received institutional review board approval.

[5] Preregistration was completed with Open Science Framework (OSF). The preregistration did not include the conceptual framing of a "jagged frontier" nor a detailed description of the specific tasks used in the experiment.

[6] Notably, participants were in early stages of their consulting careers; our findings may differ for more experienced managers in higher-level roles.

[7] Dell'Acqua et al. (2025) adopts a comparable experimental framework to evaluate subjects' competencies.

[8] See Online Appendices A and B for additional details.

[9] As confirmed by a senior Head of Office at BCG, "[t]he experimental incentive structure reflects our performance evaluation system. Crucially, once work is delivered within the agreed timeline, the actual number of days, hours, or minutes spent is not part of performance evaluation."

[10] The executive noted that these tasks almost exactly matched their process for developing new footwear. The only step missing from

this exercise was an evaluation of how the new product would integrate with the company's existing product lines. Because our experiment used a fictional company, we did not require participants to present their product suggestions in relation to existing ones.

[11] See Online Appendix B for the timing of each task.

[12] We hired 10 graders from among the MBA students at a top business school program. Each of them scored a separate set of questions based on simple rubrics. These rubrics specifically asked our graders to focus on creativity, analytical thinking, writing proficiency, or persuasiveness depending on the question.

[13] We report the grading prompts in the Online Appendix. These results are robust against alternative prompts.

[14] Table 4 use a composite quality score averaged across all questions. Each question was individually analyzed as a dependent variable, and results are entirely consistent across all questions.

[15] The additional "AI-exposure" controls enter only in column (3) in Table 4 (and column (4) in Table 4) of each regression table and consist of seven pre-experiment survey items. Two variables capture perceived task automation risk: (i) the respondent's estimate of what share of their current tasks is already performed by AI (0% = 1, 100% = 7) and (ii) agreement with the statement "my job, in principle, could be automated by AI" (1 = strongly disagree, 7 = strongly agree). Four variables gauge prior generative AI familiarity, each measured on the same one to seven Likert scale: familiarity with (1) text-generation tools, such as ChatGPT or Bard; (2) image-generation tools, such as Midjourney or DALL-E; (3) prompt-engineering techniques for eliciting better answers; and (4) the underlying workings of large language models. Finally, ChatGPT usage frequency ranges from 1 = "never used" to 5 = "daily or almost daily." Higher scores on every variable denote greater exposure to AI technologies.

[16] These percentage improvements are also relatively lower because GPT-4 scores our control baseline higher.

[17] Almost every participant (97.6%) assigned to treatment used AI at least once during the experimental task. In separate qualitative work, we examine heterogeneity in how knowledge workers interact with generative AI systems and develop a typology of human-generative AI collaboration modes (Randazzo et al. 2025b).

[18] We employ binary variables for all of these factors. "Female" is set to one if a subject identifies as female and zero otherwise. "English Native" is set to one if a subject claims native proficiency in English and zero otherwise. "Low tenure" is set to one if a subject has been with BCG for a year or less and zero otherwise. "Location" is set to one if a subject's office is located in Europe or the Middle East and zero otherwise. "Tech openness" is set to one if the subject expressed high receptivity to technology in their enrollment survey and zero otherwise.

[19] It is important to note that "higher skill" and "lower skill" are relative here. All of these knowledge workers would appear to be extremely highly skilled in most other real-world contexts.

[20] Each response was graded for correctness by BCG consultants. Additionally, we used GPT-4 for grading as we did for inside-the-frontier tasks. GPT-4 grades were consistent with those of humans in more than 97% of cases.

[21] We chose a linear probability model for its interpretability and ease of comparing treatment effects across conditions. Results remain consistent when employing a logistic regression model.

[22] Participants were allowed to proceed to the subsequent phase of the experiment upon completing their task without waiting for the allotted time to expire.

[23] As participants had to submit "at least 10 ideas." All ideas were retained for participants who submitted more than 10.

[24] In the analysis, we apply a similar regression framework to that used in Table 4 but now, focusing on the design grades. Each grader's evaluation of an individual shoe idea constitutes a separate observation, and we cluster standard errors by grader identification and participant identification. Results are robust to alternative estimation approaches.

[25] We should note that these graders assessed each shoe idea independently, focusing exclusively on creativity from a design-oriented perspective (rather than, e.g., business viability).

[26] A small number of responses (20 in total or 0.3% of all answers) were graded by only one expert rather than two.

References

Acemoglu D, Autor D (2011) Skills, tasks and technologies: Implications for employment and earnings. Card D, Ashenfelter O, eds. *Handbook of Labor Economics*, vol. 4 (Elsevier, Amsterdam), 1043–1171.

Agrawal A, Gans J, Goldfarb A (2018) *Prediction Machines: The Simple Economics of Artificial Intelligence* (Harvard Business Press, Boston).

Ali R, Tang OY, Connolly ID, Fridley JS, Shin JH, Zadnik Sullivan PL, Cielo D, et al. (2023) Performance of ChatGPT, GPT-4, and Google Bard on a neurosurgery oral boards preparation question bank. *Neurosurgery* 93(5):1090–1098.

Allen R, Choudhury P (2022) Algorithm-augmented work and domain experience: The countervailing forces of ability and aversion. *Organ. Sci.* 33(1):149–169.

Anthony C, Bechky BA, Fayard A-L (2023) "Collaborating" with AI: Taking a system view to explore the future of work. *Organ. Sci.* 34(5):1672–1694.

Aral S, Brynjolfsson E, Van Alstyne M (2012) Information, technology, and information worker productivity. *Inform. Systems Res.* 23(3 Part 2):849–867.

Athey S, Bryan K, Gans J (2020) The allocation of decision authority to human and artificial intelligence. *AEA Papers Proc.* 110(1):80–84.

Bailey DE, Faraj S, Hinds PJ, Leonardi PM, von Krogh G (2022) We are all theorists of technology now: A relational perspective on emerging technology and organizing. *Organ. Sci.* 33(1):1–18.

Barrett M, Oborn E, Orlikowski WJ, Yates J (2012) Reconfiguring boundary relations: Robotic innovations in pharmacy work. *Organ. Sci.* 23(5):1448–1466.

Beane M (2019) Shadow learning: Building robotic surgical skill when approved means fail. *Admin. Sci. Quart.* 64(1):87–123.

Beane MI, Leonardi PM (2025) Pace layering as a metaphor for organizing in the age of intelligent technologies: Considering the future of work by theorizing the future of organizing. *J. Management Stud.* 62(5):2025–2052.

Berg JM, Raj M, Seamans R (2023) Capturing value from artificial intelligence. *Acad. Management Discoveries* 9(4):424–428.

Blandin A, Bick A, Deming DJ (2026) The rapid adoption of generative AI. *Management Sci.*, ePub ahead of print January 20, https://doi.org/10.1287/mnsc.2025.02523.

Boiko DA, MacKnight R, Kline B, Gomes G (2023) Autonomous chemical research with large language models. *Nature* 624(7992):570–578.

Boussioux L, Lane JN, Zhang M, Jacimovic V, Lakhani KR (2024) The crowdless future? Generative AI and creative problem-solving. *Organ. Sci.* 35(5):1589–1607.

Brynjolfsson E, Jin W, McElheran K (2021) The power of prediction: Predictive analytics, workplace complements, and business performance. *Bus. Econom.* 56(4):217–239.

Brynjolfsson E, Li D, Raymond LR (2025) Generative AI at work. *Quart. J. Econom.* 140(2):889–942.

Brynjolfsson E, Mitchell T, Rock D (2018) What can machines learn and what does it mean for occupations and the economy? *AEA Papers Proc.* 108(1):43–47.

Buçinca Z, Malaya MB, Gajos KZ (2021) To trust or to think: Cognitive forcing functions can reduce overreliance on AI in

AI-assisted decision-making. *Proc. ACM Human-Comput. Interaction* 5(CSCW1):188.

Çalli E, Sogancioglu E, van Ginneken B, van Leeuwen KG, Murphy K (2021) Deep learning for chest X-ray analysis: A survey. *Medical Image Anal.* 73(2):102125.

Choi JH, Schwarcz D (2024) AI assistance in legal analysis: An empirical study. *J. Legal Ed.* 73(2):384–420.

Choudhary V, Marchetti A, Shrestha YR, Puranam P (2023) Human-AI ensembles: When can they work? *J. Management* 51(2): 536–569.

Cowgill B, Dell'Acqua F, Deng S, Hsu D, Verma N, Chaintreau A (2020) Biased programmers? Or biased data? A field experiment in operationalizing AI ethics. *Proc. 21st ACM Conf. Econom. Comput. (EC '20)* (Association for Computing Machinery, New York), 679–681.

Davies A, Veličković P, Buesing L, Blackwell S, Zheng D, Tomašev N, Tanburn R, et al. (2021) Advancing mathematics by guiding human intuition with AI. *Nature* 600(7887):70–74.

Dell'Acqua F (2022) Falling asleep at the wheel: Human/AI collaboration in a field experiment on HR recruiters. Working paper, Laboratory for Innovation Science, Harvard Business School, Boston.

Dell'Acqua F, Kogut B, Perkowski P (2025) Super Mario meets AI: Experimental effects of automation and skills on team performance and coordination. *Rev. Econom. Statist.* 107(4):951–966.

DeStefano T, Kellogg K, Menietti M, Vendraminelli L (2022) Why providing humans with interpretable algorithms may, counter-intuitively, lead to lower decision-making performance. MIT Sloan Research Paper No. 6797, Massachusetts Institute of Technology, Cambridge.

Drucker PF (1959) *Landmarks of Tomorrow: A Report on the New Post Modern World* (Routledge, Abingdon-on-Thames, UK).

Eloundou T, Manning S, Mishkin P, Rock D (2024) GPTs are GPTs: Labor market impact potential of LLMs. *Science* 384(6702): 1306–1308.

Faraj S, Leonardi PM (2022) Strategic organization in the digital age: Rethinking the concept of technology. *Strategic Organ.* 20(4): 771–785.

Felten EW, Raj M, Seamans R (2023) Occupational heterogeneity in exposure to generative AI. Preprint, submitted April 10, https://doi.org/10.2139/ssrn.4414065.

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

Furman J, Seamans R (2019) AI and the economy. *Innovation Policy Econom.* 19(1):161–191.

Gaessler F, Piezunka H (2023) Training with AI: Evidence from chess computers. *Strategic Management J.* 44(11):2724–2750.

Geerling W, Mateer GD, Wooten J, Damodaran N (2023) ChatGPT has aced the test of understanding in college economics: Now what? *Amer. Economist* 68(2):233–245.

Glaeser E, Hillis A, Kim H, Kominers SD, Luca M (2024) Decision authority and the returns to algorithms. *Strategic Management J.* 45(4):619–648.

Hui X, Reshef O, Zhou L (2024) The short-term effects of generative artificial intelligence on employment: Evidence from an online labor market. *Organ. Sci.* 35(6):1977–1989.

Humlum A, Vestergaard E (2025) The unequal adoption of ChatGPT exacerbates existing inequalities among workers. *Proc. Natl. Acad. Sci. USA* 122(1):e2414972121.

Iansiti M, Lakhani KR (2020) *Competing in the Age of AI: Strategy and Leadership When Algorithms and Networks Run the World* (Harvard Business Press, Boston).

Karpathy A (2024) Commentary on jagged capabilities of large language models. *Twitter* (July 25, 2024), https://x.com/karpathy/status/1816531576228053133.

Karpathy A (2025) 2025 LLM year in review. Accessed February 3, 2026, https://karpathy.bearblog.dev/year-in-review-2025/.

Kellogg KC (2022) Local adaptation without work intensification: Experimentalist governance of digital technology for mutually beneficial role reconfiguration in organizations. *Organ. Sci.* 33(2):571–599.

Kellogg KC, Myers JE, Gainer L, Singer SJ (2021) Moving violations: Pairing an illegitimate learning hierarchy with trainee status mobility for acquiring new skills when traditional expertise erodes. *Organ. Sci.* 32(1):181–209.

Lebovitz S, Levina N, Lifshitz-Assaf H (2021) Is AI ground truth really true? The dangers of training and evaluating AI tools based on experts' know-what. *MIS Quart.* 45(3):1501–1526.

Lebovitz S, Lifshitz-Assaf H, Levina N (2022) To engage or not to engage with AI for critical judgments: How professionals deal with opacity when using AI for medical diagnosis. *Organ. Sci.* 33(1):126–148.

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

Meincke L, Girotra K, Nave G, Terwiesch C, Ulrich KT (2024) Using large language models for idea generation in innovation. Working paper, Operations, Information and Decisions, The Wharton School, University of Pennsylvania, Philadelphia.

Monisha R, Sen S, Davangeri RU, Sri Lakshmi KS, Dey S (2021) An approach toward design and implementation of distributed framework for astronomical big data processing. Udgata SK, Sethi S, Gao XZ, eds. *Intelligent Systems*, Lecture Notes in Networks and Systems, vol. 431 (Springer, Singapore), 267–275.

Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ, Rajpurkar P (2023) Foundation models for generalist medical artificial intelligence. *Nature* 616(7956):259–265.

Noy S, Zhang W (2023) Experimental evidence on the productivity effects of generative artificial intelligence. *Science* 381(6654): 187–192.

OpenAI (2023) GPT-4 technical report. Preprint, submitted March 15, https://arxiv.org/abs/2303.08774.

Otis NG, Delecourt S, Cranney K, Koning R (2024a) Global evidence on gender gaps and generative AI. Working Paper No. 25-023, Harvard Business School, Boston.

Otis N, Clarke R, Delecourt S, Holtz D, Koning R (2024b) The uneven impact of generative AI on entrepreneurial performance. Preprint, submitted February 27, https://doi.org/10.2139/ssrn.4671369.

Peng S, Kalliamvakou E, Cihon P, Demirer M (2023) The impact of AI on developer productivity: Evidence from GitHub Copilot. Preprint, submitted February 13, https://arxiv.org/abs/2302.06590.

Pichai S (2025) Interview with Lex Fridman. Lex Fridman Podcast #471. Transcript. Accessed February 3, 2026, https://lexfridman.com/sundar-pichai-transcript.

Raisch S, Krakowski S (2021) Artificial intelligence and management: The automation–augmentation paradox. *Acad. Management Rev.* 46(1):192–210.

Randazzo S, Joshi A, Kellogg KC, Lifshitz H, Dell'Acqua F, Lakhani KR (2025a) GenAI as a power persuader: How professionals get persuasion bombed when they attempt to validate LLMs. Working Paper No. 26-021, Harvard Business School, Boston.

Randazzo S, Lifshitz H, Kellogg KC, Dell Acqua F, Mollick E, Candelon F, Lakhani KR (2025b) Cyborgs, centaurs and self-automators: The three modes of human–GenAI knowledge work and their implications for skilling and the future of expertise. Working Paper No. 26-036, Harvard Business School, Boston.

Reed S, Zolna K, Parisotto E, Gomez Colmenarejo S, Novikov A, Barth-Maron G, Gimenez M, et al. (2022) A generalist agent. Preprint, submitted November 11, https://arxiv.org/abs/2205.06175.

Schaeffer R, Miranda B, Koyejo S (2023) Are emergent abilities of large language models a mirage? *Adv. Neural Inform. Processing Systems*, vol. 36 (Curran Associates Inc., Red Hook, NY), 55565–55581.

Sergeeva AV, Faraj S, Huysman M (2020) Losing touch: An embodiment perspective on coordination in robotic surgery. *Organ. Sci.* 31(5):1248–1271.

Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, Scales N, et al. (2023) Large language models encode clinical knowledge. *Nature* 620(7972):172–180.

Skitka LJ, Mosier KL, Burdick M (1999) Does automation bias decision-making? *Internat. J. Human-Comput. Stud.* 51(5):991–1006.

Vaccaro M, Almaatouq A, Malone TW (2024) When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behav.* 8(12):2293–2303.

Vlamis K, Varanasi L, Paradis T (2024) MBB explained: How hard it is to get hired and what it's like to work for the prestigious strategy consulting firms, McKinsey, Bain, and BCG. *Bus. Insider Africa* (November 29), https://africa.businessinsider.com/careers/mbb-explained-how-hard-it-is-to-get-hired-and-what-its-like-to-work-for-the/v6gf0lp.

Zhou E, Lee D (2024) Generative artificial intelligence, human creativity, and art. *PNAS Nexus* 3(3):pgae052.

Fabrizio Dell'Acqua is a postdoctoral researcher at Harvard Business School and Harvard's Digital Data Design Institute. He received his PhD in management from Columbia Business School. His research examines how human-AI collaboration reshapes knowledge work at the individual, team, and organizational levels. Prior to his PhD, he received degrees in economics from Bocconi University and London Business School.

Edward McFowland III is an assistant professor of business administration in the Technology and Operations Management Unit at Harvard Business School. He also co-leads the Data Science & AI Operations Lab at Harvard's Digital Data Design Institute. His research develops statistical machine learning methods—especially for anomalous pattern detection and causal inference—to improve managerial decision-making. He earned his PhD in Information Systems and Management from Carnegie Mellon University.

Ethan Mollick is the Ralph J. Roberts Distinguished Faculty Scholar, a Rowan Fellow, and an associate professor of management at the Wharton School of the University of Pennsylvania. He received his PhD and MBA from the Massachusetts Institute of Technology Sloan School of Management. His research interests include the effects of artificial intelligence on work and education, with particular emphasis on how emerging technologies transform organizational processes and individual performance.

Hila Lifshitz is a professor of management at Warwick Business School and affiliated faculty at Harvard's Digital Data Design Institute. She is the head of the Artificial Intelligence Innovation Network at Warwick University. She conducts field studies exploring the transformation of day-to-day knowledge work processes and the use of AI for innovation processes as well as for critical decision-making processes. She earned her doctorate from Harvard Business School.

Katherine C. Kellogg is the David J. McGrath Jr Professor of Management and Innovation at the Massachusetts Institute of Technology Sloan School of Management. Her research focuses on helping knowledge workers and organizations develop and implement predictive and generative artificial intelligence solutions on the ground in everyday work, with a particular focus on worker voice, learning, decision making, collaboration, and innovation.

Saran Rajendran is currently a Director of AI Strategy at Palo Alto Networks. Previously, he was a project leader at Boston Consulting Group, where he worked at the Henderson Institute while working on this paper.

Lisa Krayer is a principal at Boston Consulting Group, where she focuses on the business and workforce implications of emerging technologies, including generative artificial intelligence (AI). Her recent research and engagement with clients examine the impact of generative AI on talent development, skills, and learning, with an emphasis on large-scale upskilling. She holds a PhD in electrical engineering from the University of Maryland, with a research focus on photonics and optoelectronics.

François Candelon is a Partner at Seven2, a private equity firm, and formerly served as Global Director of the BCG Henderson Institute. He is currently an Executive Fellow at Harvard's Digital Data Design Institute. His research interests include AI adoption and human-AI collaboration in corporations and their impact on value creation. He holds an MSc from the Ecole Polytechnique and a predoctorate degree in Industrial Economy from University Paris-Dauphine.

Karim R. Lakhani is the Dorothy & Michael Hintze Professor of Business Administration at Harvard Business School, specializing in technology management, open innovation and AI strategy and transformation. He is the founding chair of Harvard's Digital Data Design Institute, and the Laboratory for Innovation Science. His work includes pioneering field experiments with organizations like NASA, Harvard Medical School, Broad Institute, and Procter & Gamble. Lakhani holds a PhD in Management from MIT.