Early LLM-based Tools for Enterprise Information Workers Likely Provide Meaningful Boosts to Productivity

A first update from Microsoft's research initiative on AI and Productivity

Alexia Cambon[1], Brent Hecht[1], Ben Edelman, Donald Ngwe, Sonia Jaffe, Amy Heger, Mihaela Vorvoreanu, Sida Peng, Jake Hofman, Alex Farach, Margarita Bermejo-Cano, Eric Knudsen, James Bono, Hardik Sanghavi, Sofia Spatharioti, David Rothschild, Daniel G. Goldstein, Eirini Kalliamvakou, Peter Cihon, Mert Demirer, Michael Schwarz, and Jaime Teevan

(with additional support from the entire AI and Productivity team at Microsoft)

ABSTRACT

This report presents the initial findings of Microsoft's research initiative on "AI and Productivity", which seeks to measure and accelerate the productivity gains created by LLM-powered productivity tools like Microsoft's Copilot. The many studies summarized in this report, the initiative's first, focus on common enterprise information worker tasks for which LLMs are most likely to provide significant value. Results from the studies support the hypothesis that the first versions of Copilot tools substantially increase productivity on these tasks. This productivity boost usually appeared in the studies as a meaningful increase in speed of execution without a significant decrease in quality. Furthermore, we observed that the willingness-to-pay for LLM-based tools is higher for people who have used the tools than those who have not, suggesting that the tools provide value above initial expectations. The report also highlights future directions for the AI and Productivity initiative, including an emphasis on approaches that capture a wider range of tasks and roles.

1 Introduction

One of the most significant ways that technological advances help improve quality of life is by enabling step-function improvements in labor productivity [6,8]. Many have hypothesized that recent improvements in large language models (LLMs) and applications built on LLMs would provide such a productivity boost, and a historically large one at that (e.g., [4,6]). Early evidence mostly supported this hypothesis, with productivity gains seen in various types of lab studies when workers were provided with LLM-based tools [7,10,18,22,23,25].

Microsoft has invested significantly in building productivity tools based on LLMs (branded "Copilot"), in large part based on the hypothesis that such tools would substantially increase the productivity of information workers. In concert with building these tools, Microsoft also formed a cross-company research team (the "AI and Productivity" research team) that seeks to measure and accelerate the productivity gains provided by these tools. As Copilot has started to become a reality, this team – from which this paper emerges – began to explore one of the first research questions that needs to be answered to advance our broader mandate:

> *RQ: What impact does Copilot have on productivity for common enterprise information worker tasks for which LLMs have been hypothesized to provide significant value?*

Importantly, for this stage of work, we did not seek to evaluate Copilot's (or LLM's) overall increase in information worker productivity using a fully representative set of tasks, nor did we attempt to run experiments in fully ecologically valid scenarios. We are moving in these directions in ongoing research. Rather, for this phase of research, we sought to explore whether Copilot provides meaningful productivity boosts on some common tasks for which we believed it was likely to do so based on the properties of the technology. In other words, in the language of Dell'Acqua et al. [10], we primarily explored tasks on the AI-friendly side of the "jagged technological frontier."

The dozens of researchers in Microsoft's "AI and Productivity" research team have launched over 30 studies looking at this first question, as well as more advanced ones that will be the subject of follow-up reports. While most of the studies remain active, a sufficient number have returned results to justify a synthesis of the research we have conducted to address this first question. This report covers these initial findings. In future manuscripts, we plan to report on research with substantially increased ecological validity (e.g., via RCTs deployed in real organizations), added diversity in methodology (e.g., via privacy-preserving analyses of chat logs [24]), targeted studies of potential productivity barriers, analyses of new and improved capabilities, and other means of more holistically understanding the productivity impacts of LLM-based tools.

At a high level, the results thus far support the hypothesis that the first versions of Microsoft's Copilot tools do substantially increase productivity on some common tasks performed by enterprise

[1] Contact authors: Alexia Cambon (alexia.cambon@microsoft.com) and Brent Hecht (brent.hecht@microsoft.com)

information workers. This productivity boost most often appears as a meaningful increase in speed of execution without a significant decrease in quality, although there is some variation how the greater productivity manifests.

We also saw several other trends in the first wave of studies. For instance, the willingness-to-pay for LLM-based tools is higher for people who have used the tools, suggesting that the tools provide value above initial expectations. We also see evidence of high self-reported productivity when using LLM-based tools, with perceived time saved substantially exceeding actual time saved. Finally, although not the focus of this phase of the research, we observed early evidence that there are tasks for which the productivity gains are more complex and may be harder to actualize in certain cases, with current LLM-based tools providing a new set of options rather than simply accelerating existing ways of working.

Below, we first provide a brief description of the studies whose results are included in this first research report. We then present cross-cutting findings of the studies, focusing on three dimensions of productivity: speed, quality, and effort. Finally, we close with Discussion and Future Work.

2 Study Descriptions

In this section we provide a brief overview of the studies included in this first report. Many are or will be supported by separate dedicated manuscripts, and we leave methodological detail to those manuscripts, providing links where possible. While none of these studies have yet gone through peer review, several are currently under review with more to be submitted soon. Note also that in addition to the eight studies presented here, the AI and Productivity research team has many others in progress. These will be discussed in future reports as their findings develop. When we discuss results of the studies that involve comparisons (e.g. between a treatment and control), we only report results that were determined to be significant by a traditional statistical test (unless otherwise noted).

Copilot Common Task Study: This experiment assessed productivity gains from Copilot across multiple tasks common among office workers: email information retrieval, intranet (SharePoint) information retrieval, content creation, and meeting summarization. We recruited 147 participants via the Upwork platform. We randomly divided participants into treatment (Copilot) and control (no Copilot) groups. To assess the quality of participants' information retrieval, we asked them questions about the information they found, and we scored their accuracy. To assess the quality of the content they drafted, we asked a LLM both to check for key facts and to assess quality along multiple dimensions. To assess speed, we measured the time taken for each subtask and for the series of tasks as a whole. In a post-task survey, we also asked participants how they felt about the task and the products they used. Details are in the full manuscript [13].

Primary study authors: Benjamin Edelman, Donald Ngwe

Copilot Usage in the Workplace Survey: Microsoft first launched Copilot to the public via an "Early Access Program." Only a small number of people in each organization invited to join the program had access to Copilot. We surveyed these early adopters to understand their experiences and perceived productivity gains using Copilot. We excluded responses from anyone who said they had been using Copilot for less than three weeks, with a final sample consisting of 297 responses. The survey was 10 minutes in length and was deployed globally from October 3 to November 2, 2023, with a cross-functional representation in its sample. The survey was fully anonymous.

Primary study authors: Alexia Cambon, Alex Farach, Margarita Bermejo-Cano, Eric Knudsen

Copilot in Teams Meeting Study: This study sought to test productivity effects of Copilot for catching up on missed meetings. We recruited 57 Microsoft employees for the study. Access to Copilot was deployed across Microsoft largely on a division-by-division basis, with certain divisions gaining access before others. Thirty-three of the participants in the study had access to Copilot, while the rest did not. We provided participants with a recording and transcript of a scripted 35-minute Teams meeting between four employees planning a fictional team offsite. We then asked participants to write a 200- to 300-word email summarizing details from the "missed" meeting. Participants completed the study unmoderated and online. They recorded their computer screens after providing consent. To assess speed, we used time on the task. To measure quality, two independent (human) raters scored email contents for key details from the meeting. After the task was complete, we asked participants how productive they felt and how draining they perceived the task to be, among other questions.

Primary study authors: Amy Heger, Mihaela Vorvoreanu, Alexia Cambon

Copilot Information Retrieval Study: This experiment examined productivity gains in enterprise information retrieval. We recruited 163 participants via Upwork, randomly split them into treatment (Copilot) and control (no Copilot) groups, and we brought them into a simulated enterprise environment. The environment included an organizational procurement policy customized from a public-domain government procurement policy, as well as a large number of files that were irrelevant to the task. We also prepopulated users' email inboxes and calendars with a variety of messages and meetings. We asked participants to answer questions about information in the files, emails, and meeting invites. We scored their work for accuracy, and we measured the time required for completion. In a post-task survey, we also asked participants how they felt about the task and the products they used. Details are in the full manuscript [13].

Primary study authors: Ben Edelman, Donald Ngwe, Sida Peng

LLM-based Search Study: This early-stage study used an LLM-based search tool based on GPT-3.5 and compared the tool with traditional search for complex information retrieval tasks. Participants were recruited via Amazon Mechanical Turk. We ran two experiments: one that examined speed and accuracy of the task (90 participants) and another focused on the effectiveness of a

simple intervention for when the LLM produced incorrect results (120 participants). Details are in the full manuscript [25].

Primary study authors: Jake Hofman, Sofia Spatharioti, David Rothschild, Dan Goldstein

Outlook Email Study: In this study, we assessed the quality of emails written using Copilot in Outlook versus ordinary emails written without Copilot. We recruited 62 participants from Upwork, all native speakers of US English. For a corpus of actual emails among business professionals and executives, we selected authors and messages from the Enron email corpus, a common dataset for e-mail research [9,20]. For each original email we selected, we then used the Outlook "Sound Like Me" feature to create a variant rewritten by Copilot in the style of the underlying author. We repeated this process for numerous emails from numerous authors. We then showed the human-written and Copilot-written messages to participants in various arrangements and combinations, and we asked them to evaluate both the human- and AI-written emails on factors such as clarity, conciseness, and the degree to which an email "sounded like" the original author. Details are in the full manuscript [12].

Primary study authors: Ben Edelman, Donald Ngwe

M365 Defender Security Copilot Study: This experiment studied productivity gains in enterprise security operations. We recruited 149 participants from Upwork, security novices with standard IT skills but no specific security expertise. We gave them all access to M365 Defender, and we randomly granted half of them access to Security Copilot, which provides AI-based interpretation of security incidents, recommends responses, and interprets attack scripts. We asked participants to explore the tool, then answer questions about what they found. Details are in the full manuscript [11].

Primary study authors: Ben Edelman, James Bono, Sida Peng

GitHub Copilot Study: The first study in the research program whose results have already been widely disseminated, this study recruited 95 developers from Upwork to implement an HTTP server in JavaScript and gave half of them access to GitHub Copilot. Participants were incentivized to complete the task quickly. The task was considered complete when a participant's code passed twelve programmed tests. Researchers measured both the completion rate and average time to completion for participants with and without Copilot. Details are in the full manuscript [23].

Primary study authors: Sida Peng, Eirini Kalliamvakou, Peter Cihon, Mert Demirer

3 Findings

Defining productivity

Choosing a definition of productivity that is relevant to a wide variety of information worker contexts is notoriously difficult [16,21]. For this work, we opted to use a three-part framework that aims to capture both short- and long-term productivity effects that could result from the introduction of LLM-based tools for information workers. The three parts of this framework are (1) *speed*, (2) *quality*, and (3) *effort*.

The first two metrics are motivated by the basic economic understanding of productivity as output for a given input. We track how quickly participants finish a task, the output per unit time (*speed*). Since the lab studies asked participants to do a fixed set (quantity) of tasks, we look at the *quality* of their work as a measure of output. The definition of quality varies from study to study, but accuracy on the task in question is the most common.

Our centering of *effort* as a core metric is less traditional, but aims to capture well-known long-term productivity effects that can result from changes to working practices, often enabled by increased automation. In particular, scholars of Taylorism and related production strategies have often noted that boosts to productivity through new working practices are sometimes offset in the long-term by increases in turnover, growing worker dissatisfaction, and related effects [3,5]. Relatedly, there is some evidence that broad-based job satisfaction is currently quite low (e.g. [17]); as such, it is also possible that LLM-based technologies might actually help mitigate these types of issues, e.g. by reducing effort on tasks perceived as draining. The studies in this paper that track effort operationalize it differently, and usually do so through survey-based approaches that seek to understand the degree of exhaustion experienced or perceived energy expended by a participant. Future studies will seek to evaluate effort using other techniques, including functional neuroimaging.

Speed

For almost all of the tasks across all of the studies, we observed a significant increase in the speed of performing tasks when using an LLM-based tool relative to performing the tasks without the tool. Figure 1 summarizes these observed speed increases. Copilot users completed the task in 26% to 73% of the time, on average, when compared with people not using Copilot, with the largest difference being in the Copilot in Teams Meeting Study. Of course, the overall current speed increase from using Copilot is likely to be much lower than this, given the tasks studied were ones we believed were particularly likely to benefit from AI support.

Our surveys showed that participants also perceived substantial time savings. In the Copilot Common Tasks Study, we asked participants with Copilot to estimate how much time that tool saved them. They guessed 36 minutes on average, when actual time savings were 12 minutes on average, indicating their perception of significant time reduction on the tasks. Similarly, 73% of employees surveyed in the Early Access Program research agreed that Copilot helped them complete tasks faster, and 85% said it would help them get to a good first draft faster. When asked to estimate how much time Copilot saved them on a daily basis, respondents most often reported 11-30 minutes (35%) and another 22% reported greater than that. Average daily time saved was 14 minutes a day, or 1.2 hours a week when calculated using the lower

end of the time range bins ("0 minutes", "<5 minutes", "5-10 minutes", "11-30 minutes", "31-60 minutes", and ">1 hour").

Quality

Looking across the set of studies, there was a strong trend for the increases in speed to come without costs to quality – participants using LLM-based tools achieved quality levels that were not statistically distinguishable from those who were not using LLM-based tools. For example, in both the Copilot Common Task Study and the Copilot Information Retrieval Study, we did not see a statistically significant effect on task accuracy, despite the substantial increases in speed discussed above. Relatedly, the Copilot Usage in the Workplace Survey found that self-reported quality assessments were also good: 68% agreed that Copilot actually improved quality of their work.

We did observe a few exceptions to the general trend of "free" increases in speed without costs to quality. In the Copilot in Teams Meetings Study, there was a slight decrease in summarization comprehensiveness relative to the comparison group. While Copilot users took much less time, their summaries on average included 11.06 out of 15 specific pieces of information in the assessment rubric versus the 12.40 of 15 for users who did not have access to Copilot.

The Outlook Email Study offered mixed evidence with respect to quality. Emails written with Copilot were rated 18% clearer and 19% more concise. Participants also scored emails written with Copilot 25% higher on, "Couldn't have said it better myself." However, in some framings, subjects rated messages by Copilot as less likely to be written by a human. That said, in other framings, subjects were worse than random at identifying the human-written messages. We see this as mixed evidence for the quality of Copilot text. Copilot users appeared to like many aspects of the text it generated, but could sometimes tell the difference between Copilot writing versus human writing.

We also saw lower quality in one of the tasks in the LLM-based Search Study [25] that was designed to be particularly complex so the LLM would get the answer wrong. We found that the LLM's accuracy substantially impacted participants' performance: when the LLM got the answer right, there was no significant difference between the treatment (LLM-based search) and the control group (LLM-based search), but when the LLM got the answer wrong accuracy dropped to under 50% in the treatment group. The study also identified that straightforward UX-based uncertainty visualizations can help address this issue. In particular, a simple color-coded highlighting scheme, similar to spelling or grammar check, was found to be effective for helping people identify potentially unreliable information in LLM responses.

The search question the AI got wrong is an example of a task on the other side of the "technological frontier" [10], unlike most of the tasks considered in this report for which we hypothesized the early version of these tools presented strong aptitude for AI assistance. Future work that examines a broader set of tasks and more representative work environments will likely involve more exploration on the other side of this frontier, as well as of course research into solutions that will move the frontier.

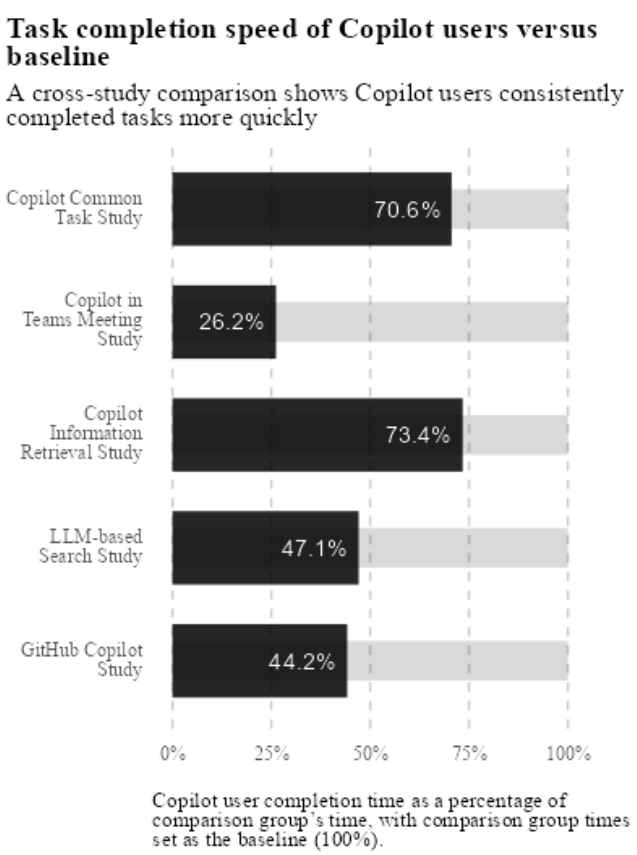

Figure 1: A summary of *speed* results from some of the studies. The gray bars, all normalized to 100%, indicate performance of the study groups that did not have access to an LLM-based productivity tool. The black bars indicate the percentage of time those with access to a tool took relative to those without access.

Finally, in the M365 Defender Security Copilot Study we saw a significant and substantial increase in quality, although that was almost certainly in part a product of the participant pool being made up of security novices. Security novices with Copilot were 44% more accurate in answering questions about the security incidents they examined. We also asked subjects to write an essay about a security incident, and those with Copilot included 151% more key facts (based on the experts' assessment rubric) and scored 32% higher on summary quality, both as graded by LLM. Future work should seek to assess related effects for those with security experience.

Effort

When we were able to measure effort, we did not see any major warning signs regarding long-term productivity impact (e.g., burn-out, exhaustion, or reduction in motivation). In fact, we typically saw the opposite. For instance, in the Copilot in Teams Meeting Study, participants with access to Copilot found the task to be 58%

less draining than participants without access. In the M365 Defender Security Copilot Study, subjects with Copilot reported 16% lower agreement with the statement "This task was a lot of effort" (compared to control users doing the same task without Copilot), and Copilot users reported 83% agreement with the statement "Copilot reduced my effort on this task." Similarly, in the Copilot Usage in the Workplace Survey, Copilot users overall agreed that Copilot helped them spend less mental effort on mundane or repetitive tasks (72%).

The one exception to this trend related not to the capabilities of the AI, but rather challenges with using pre-release software. For instance, in the Copilot Common Task Study, technical disruptions hindered some AI functions, and participants who experienced this problem reported notably lower scores on metrics associated with effort.

Perceived value

As noted earlier, this report should not be read as an assessment of overall productivity gains on a representative set of tasks in wide variety of ecologically valid work environments. Rather, in this wave of work, we sought to test whether we saw productivity boosts for some common tasks that enterprise information workers perform. However, the Copilot Usage in the Workplace Survey provides a preview of what we might expect when we examine Copilot in more ecologically valid conditions. The survey indicated that the majority (70%) of respondents believed that Copilot increased their productivity. Given that the tools were only available to survey respondents in pre-release form, these results suggest more general productivity benefits as we broaden our research scope, even if these initial results are likely skewed by some novelty effects [14].

We also saw evidence for how users valued the tool in what they were willing to give up for it. In the Copilot Usage in the Workplace Survey, 77% of respondents said they would choose Copilot over a weekly free lunch worth $40 / month. In the lab studies, people had exposure to Copilot for only the one study, but were still willing to pay more for it than those who had not used it. In the Copilot Common Task Study, the reported willingness-to-pay for Copilot was 35% greater than for those who had used Copilot in the study versus those who received only a description of the tool. Similarly, those in the treatment group of the Copilot Information Retrieval Study were 40% more likely to say they would pay >$20 USD for the tool relative to those in the control group.

4 Discussion and Future Work

With any new "general purpose technology" like LLMs [15], research shows that it often takes both time and complementary innovations to realize significant productivity gains [8]. The results in this paper suggest we are already moving steadily along this journey. That said, it is clear that there is much more to do.

Perhaps most notably, nearly all research seeking to measure potential productivity gains from LLMs (including the research in this paper) has focused on task-level productivity, but the literature suggests one of the most critical complementary innovations that will be needed maximize productivity gains will be entirely reinvented workflows with new and rearranged tasks [8]. Tools that are chat-centered (vs. app-centered) and agent-like technologies may mark the beginning of these new workflows, but future research is needed to assess whether that can and will bear out. Different productivity measurement strategies will likely also be needed if people are bypassing existing tasks and doing new ones instead, rather than simply doing existing tasks faster or better.

It is also important to note that the tools studied in this paper – and the underlying language models – are under active development. As such, the results in this paper should be viewed as capturing a moment in time versus providing evidence for strong claims about LLM-based productivity tools in a long-term persistent steady state. This is a key reason we put the word "early" in the title of this paper. Running the same studies on the same tools in a year may yield different results. Similarly, user expectations are changing quickly as well, with any novelty effects likely moderating relatively quickly. This paper must also be viewed with awareness of the conflicts of interest involved: the authors are Microsoft researchers studying Microsoft products. While the studies were done using standard scientific practices and many are being submitted for peer review, it is useful to make these conflicts explicit.

Another limitation of the research presented is that it was all done with English speakers doing tasks in English. It is likely that for some languages, especially low-resource ones, the observed productivity gains will be less due to a decrease in quality of the tools' output. Researchers at Microsoft and elsewhere are working to improve the "hyperlingual" [19] capabilities of LLMs and LLM-based productivity tools, e.g. in non-English and low-resource languages [1,2].

Though the early findings above suggest LLM-based productivity tools largely provide a "free" speed increase without decreasing quality, we also saw some evidence that for certain tasks, LLMs will likely pose a tradeoff. In doing so, LLM-based productivity tools may sometimes provide a new option for information workers that did not exist before: the ability to do a certain set of tasks far faster but with marginally lower quality, potentially freeing up time for other activities. We plan to explore this other side of the "frontier" [10] in upcoming work.

More generally, there have been a number of major disruptions to work practices in recent years: not just LLMs, but also remote work and hybrid work. One way to view the impact of all these changes is through the lens of the new dimensions of choice they have created for workers. Remote work created new opportunities for where work can be done (what *space* is used), each with its own pros and cons. Improved tools for asynchronous collaboration (like the ability to record meetings) allow people more options for when work can be done (what *time* it gets done). And current LLMs now create new possibilities for the way that work is accomplished

(what *intelligence* is brought to the fore). Five years ago, for example, most content creation in the enterprise was done in an office between 9am and 5pm, with a human writing every word. Now there are many additional ways to do that work, each with its relative strengths and weaknesses. While the expanded choices bring new uncertainty and risks, they also create substantial new opportunities for making work overall far better than it was before.

REFERENCES

1. Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Maxamed Axmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual Evaluation of Generative AI. https://doi.org/10.48550/arXiv.2303.12528
2. Sanchit Ahuja, Divyanshu Aggarwal, Varun Gumma, Ishaan Watts, Ashutosh Sathe, Millicent Ochieng, Rishav Hada, Prachi Jain, Maxamed Axmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGAVERSE: Benchmarking Large Language Models Across Languages, Modalities, Models and Tasks. https://doi.org/10.48550/arXiv.2311.07463
3. Hugh G. J. Aitken. 2014. *Scientific Management in Action: Taylorism at Watertown Arsenal, 1908-1915*. Princeton University Press.
4. Sam Altman. 2021. Moore's Law for Everything. Retrieved November 13, 2023 from https://moores.samaltman.com/
5. Peter Bain, Aileen Watson, Gareth Mulvey, Phil Taylor, and Gregor Gall. 2002. Taylorism, targets and the pursuit of quantity and quality by call centre management. *New Technology, Work and Employment* 17, 3: 170–185. https://doi.org/10.1111/1468-005X.00103
6. Erik Brynjolfsson. 2022. The Turing Trap: The Promise & Peril of Human-Like Artificial Intelligence. https://doi.org/10.48550/arXiv.2201.04200
7. Erik Brynjolfsson, Danielle Li, and Lindsey R. Raymond. 2023. Generative AI at Work. https://doi.org/10.3386/w31161
8. Erik Brynjolfsson and Andrew McAfee. 2014. *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*. W. W. Norton & Company, New York.
9. William W. Cohen. 2015. Enron Email Dataset. Retrieved from https://www.cs.cmu.edu/~enron/
10. Fabrizio Dell'Acqua, Edward McFowland, Ethan R. Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R. Lakhani. 2023. Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality. https://doi.org/10.2139/ssrn.4573321
11. Benjamin G. Edelman, James Bono, Sida Peng, Roberto Rodriguez, and Sandra Ho. 2023. Randomized Controlled Trial for Microsoft Security Copilot. Retrieved November 30, 2023 from https://papers.ssrn.com/abstract=4648700
12. Benjamin G. Edelman and Donald Ngwe. 2023. Sound Like Me: Findings from a Randomized Experiment. Retrieved November 30, 2023 from https://papers.ssrn.com/abstract=4648689
13. Benjamin G. Edelman, Donald Ngwe, and Sida Peng. 2023. Measuring the Impact of AI on Information Worker Productivity. Retrieved November 30, 2023 from https://papers.ssrn.com/abstract=4648686
14. Steven M. Edwards and Harshavardhan Gangadharbatla. 2001. The Novelty of 3D Product Presentations Online. *Journal of Interactive Advertising* 2, 1: 10–18. https://doi.org/10.1080/15252019.2001.10722054
15. Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. 2023. GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models. https://doi.org/10.48550/arXiv.2303.10130
16. Nicole Forsgren, Margaret-Anne Storey, Chandra Maddila, Tom Zimmermann, Brian Houck, and Jenna Butler. 2021. The SPACE of Developer Productivity: There's more to it than you think. *ACM Queue* 19, 1: 20–48.
17. Gallup Inc. *State of the Global Workplace: 2023 Report*. Retrieved November 30, 2023 from https://www.gallup.com/workplace/349484/state-of-the-global-workplace.aspx
18. Jennifer Haase and Paul H. P. Hanel. 2023. Artificial muses: Generative Artificial Intelligence Chatbots Have Risen to Human-Level Creativity. *Journal of Creativity* 33, 3: 100066. https://doi.org/10.1016/j.yjoc.2023.100066
19. Brent Hecht and Darren Gergle. 2010. The Tower of Babel Meets Web 2.0: User-Generated Content and Its Applications in a Multilingual Context. In *CHI '10: 28th International Conference on Human Factors in Computing Systems* (CHI '10), 291–300. https://doi.org/10.1145/1753326.1753370
20. Bryan Klimt and Yiming Yang. 2004. The enron corpus: A new dataset for email classification research. In *European conference on machine learning*, 217–226.
21. Gloria Mark. 2023. *Attention Span: A Groundbreaking Way to Restore Balance, Happiness and Productivity*. Hanover Square Press.
22. Shakked Noy and Whitney Zhang. 2023. Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence. https://doi.org/10.2139/ssrn.4375283
23. Sida Peng, Eirini Kalliamvakou, Peter Cihon, and Mert Demirer. 2023. The Impact of AI on Developer Productivity: Evidence from GitHub Copilot. https://doi.org/10.48550/arXiv.2302.06590
24. Chirag Shah, Ryen W. White, Reid Andersen, Georg Buscher, Scott Counts, Sarathi Das, Ali Montazer, Sathish Manivannan, Jennifer Neville, Xiaochuan Ni, Nagu Rangan, Tara Safavi, Siddharth Suri, Mengting Wan, and Longqi Yang. 2023. Using Large Language Models to Generate, Validate, and Apply User Intent Taxonomies. Retrieved November 30, 2023 from https://www.microsoft.com/en-us/research/publication/using-large-language-models-to-generate-validate-and-apply-user-intent-taxonomies/
25. Sofia Eleni Spatharioti, David M. Rothschild, Daniel G. Goldstein, and Jake M. Hofman. 2023. Comparing Traditional and LLM-based Search for Consumer Choice: A Randomized Experiment. https://doi.org/10.48550/arXiv.2307.03744