Microsoft

Generative AI in Real-World Workplaces: Microsoft's Second Research Report on AI and Productivity

Editors: Sonia Jaffe, Neha Parikh Shah, Jenna Butler, Alex Farach, Alexia Cambon, Brent Hecht, Michael Schwarz, and Jaime Teevan

Full report URL: aka.ms/productivity-report2

© Copyright Microsoft Corporation. All rights reserved.

Generative AI in Real-World Workplaces: Microsoft's Second Research Report on AI and Productivity

Microsoft released a new research report focusing on real-world impact of generative AI tools as workers integrate them into everyday complex workflows

Results suggest that the positive productivity effects observed in a lab setting are indeed beginning to manifest in real-world work, and, perhaps even more importantly, they show how generative AI can provide even greater impact as work practices evolve

Summarizes over a dozen recent Microsoft studies from 50+ researchers on the application of generative AI in real-world workplaces – including the first large-scale randomized trial of Copilot with thousands of users from over 60 customers

Link: http://aka.ms/productivity-report-pdf

Key Learnings

Generative AI is already helping people be measurably more productive in their day-to-day jobs

The productivity story in real-world workflows is more complex than observed in lab studies (as expected)

Productivity gains associated with generative AI, including time and accuracy, vary by role, function and organization

Variance in adoption and utilization influences AI's impact

Early studies suggest generative AI may affect the cognitive effort required for task completion

Learning 1: AI is already helping people be measurably more productive

Early Access Telemetry Study

Likely largest randomized, controlled study ("clinical trial") of LLMs in the workplace, compares employees with and without Copilot

Examines over 60 tenants and 6k employees

	Copilot Users
Documents	• 10% more documents created and edited
Emails	• 11% fewer emails read • 4% less time interacting with emails
Meetings	• Effects differed by company

Work Trend Index Survey

Expansive survey aimed to capture user sentiments and experiences with generative AI, broadly

Includes 31K global information workers across 31 countries

- 29% of AI users report being familiar with generative AI, using it at work at least several times a week, and saving more than 30 minutes a day by using it

Bing Chat Log Analysis (link)

Study to understand how people use AI-augmented search differently from traditional search, by task complexity and type of work domain

Analyzes 80k randomly selected, de-identified conversations from Bing Copilot and Bing searches

	High Complexity Tasks	Knowledge Work Domains
Bing Copilot (AI augmented)	**37%**	**73%**
Bing Search (traditional search)	13%	37%

Learning 2: The productivity story in real-world workflows is more complex than observed in lab studies

Lab studies show straightforward gains

Designed for tasks hypothesized to be amenable for generative AI, generally showing straightforward and substantial improvements to productivity

As usual, real world has more complexity

Real workflows more complex than controlled lab environments and include tasks at which Copilot is less good or not yet assisting. There are costs to integrating AI into existing workflows

We do see meaningful impacts within real, complex workflows

Impressive to see metrics like documents edited and e-mail time moving significant. Effect sizes lower than some expectations, higher than others

Some benefits are hard to capture

- People report using Copilot much less mentally demanding, but do not differ in a test of cognitive load (Stroop test)
- People using GitHub Copilot report they don't want to do tasks without it, but we find no effect on Engineering Systems NSAT, likely due to the limited time developers spend coding and the influence of other tools and activities
- Perceived time saved is generally larger than actual savings – echoes long line of research suggesting an unmeasured element makes tasks with Copilot more enjoyable (ex: elevator rides with mirrors feel like they go faster)

Learning 3: Productivity gains associated with generative AI, including time and accuracy, vary by role, function, organization

Work Trend Index Survey: By Job Function

Perceived Benefits of AI Integration

(31k+ respondents)

- **Task Automation**: Benefits depend on the extent of tasks in function that can be automated or assisted by AI.
- **Repetitive Tasks**: Most considerable improvements in roles with repetitive tasks or extensive content creation.
- **High Interaction Roles**: Largest gains in productivity, quality, and efficiency in customer-facing roles (Customer Service, Sales) and content creation (Marketing, Creative).
- **Specialized Fields**: More modest improvements in technical fields (Legal, Engineering) due to the complexity and nuance of tasks

Software Developers

- **Automation preference**: *(800+ MS developers)* Most want to see AI help with automating routine tasks, like generating unit tests and writing documentation
- **Engineering Satisfaction**: *(30k+ developers)* GitHub Copilot use did not seem to affect Engineering systems satisfaction, likely due to the limited time developers spend coding and the influence of other tools and activities
- **Lab study**: GitHub Copilot improved speed and accuracy on a familiar task, but not on one with unfamiliar concepts and components

Customer Service Agents

- (Lab study) Sellers with licensing chatbot answered faster higher accuracy, more completeness than those without

Security Analysts

- (Lab study) Security analysts with Copilot more accurate and complete than those without. Gains were smaller for experts than novices, but still substantial

Multilingual teams

- (Lab study) Copilot improved accuracy for Japanese speakers trying to understanding an English meeting recording

	Accuracy among Japanese listeners recapping a	
	Japanese Meeting	English Meeting
With Copilot Meeting Recap	96.8%	97.5%
Standard tools	94.8%	83.8%

Learning 4: Variance in adoption and utilization influences AI's impact

Consistent Copilot use leads to job benefits

M365 Copilot user survey (n=885)

- Using Copilot for more than 10 weeks report more benefits compared to those with shorter usage (job enjoyment, reduced meeting attendance)
- Users who report saving over 11 min/day also report better work-life balance, more interesting and fulfilling work, more motivated and productive

Power Users regularly experiment with different ways of using AI

WTI Survey (n=31k)

- What defines a power user?

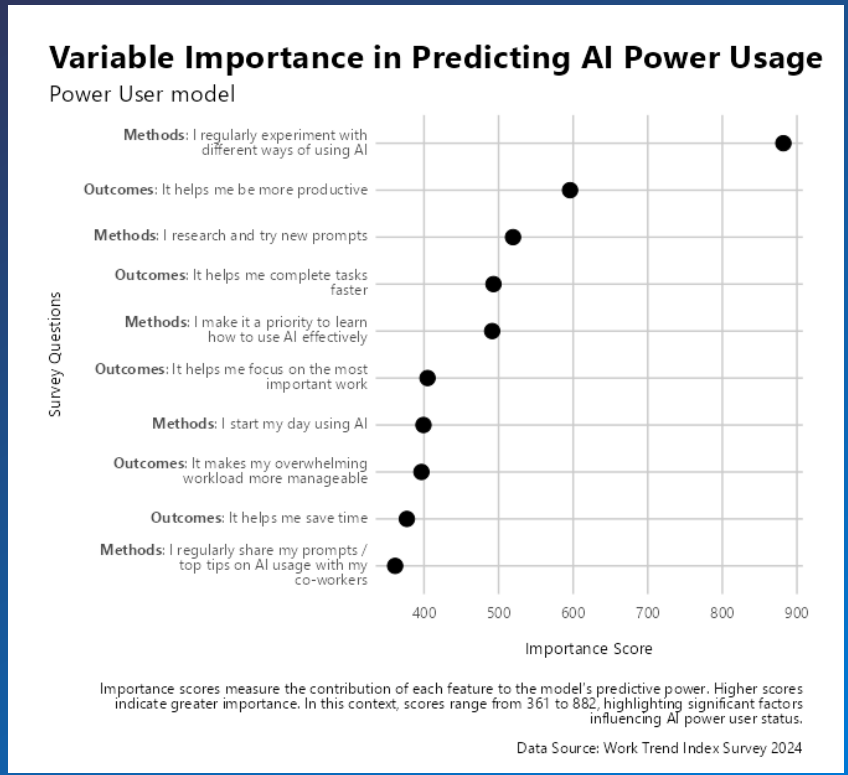
  Most importantly, they regularly experiment with using AI. Also, less 'unsanctioned' AI use and learning focus

Effects of Copilot differ based on usage

Early Access Telemetry study (randomized)

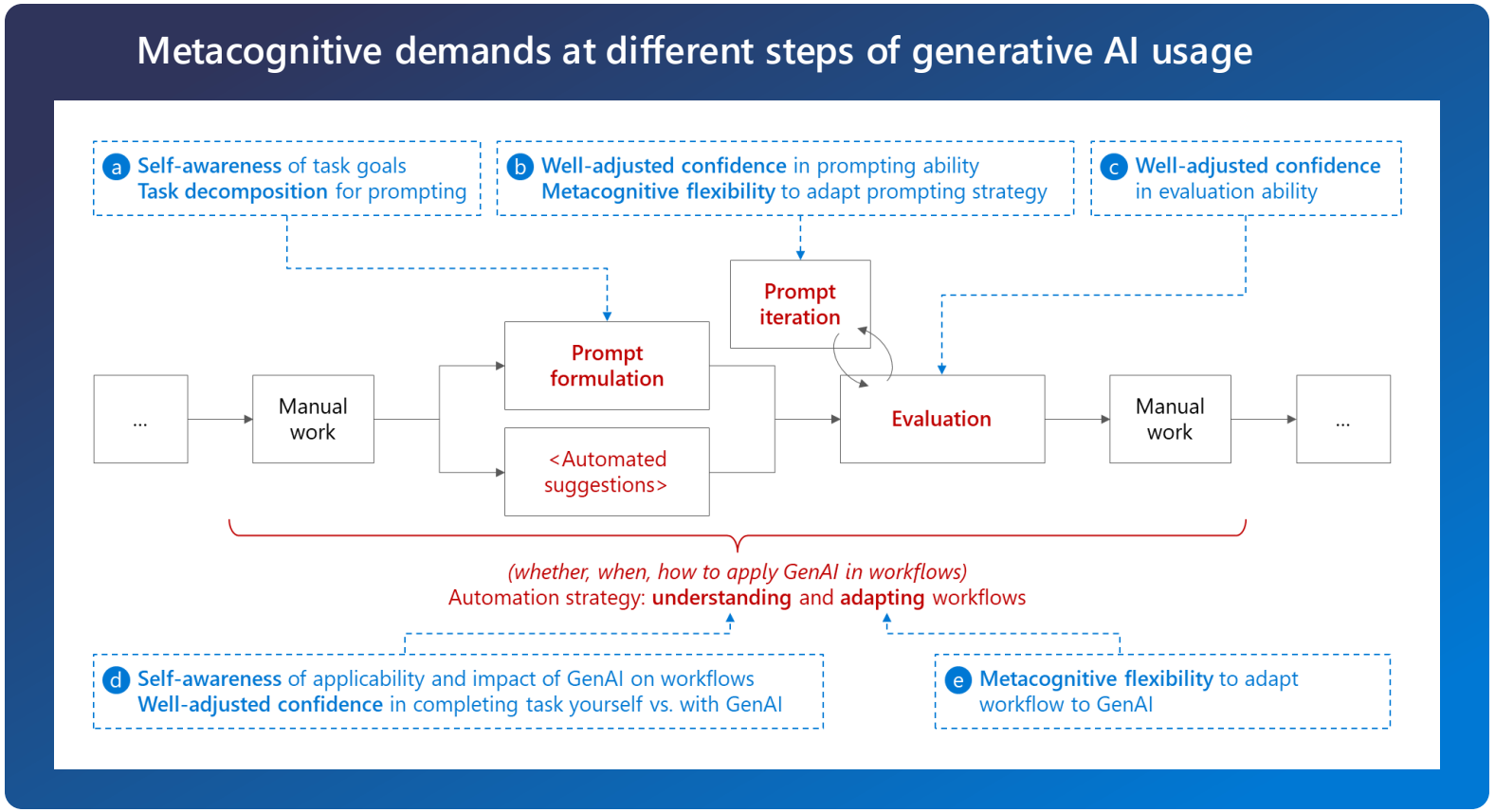
- Substantial variation across tenants in adoption/usage rates. Substantial difference between effect of being given a Copilot *license* and effect of Copilot *usage*

Responses most predictive of being a Power User



Learning 5: Early studies suggest generative AI may affect the cognitive effort required for task completion

High-impact paper on generative AI and metacognition discusses how generative AI requires users to "think about thinking" E.g., Task goals, task decomposition, and confidence in one's ability to evaluate the output



Indicates need for more metacognitive support in generative AI systems/tools, including:

- Breaking down complex tasks into sub-tasks
- Explainability and helping people learn an accurate mental model of what the systems can do
- Providing evaluation support
- New non-text-forward interfaces

A small, internal lab study found people reported Copilot lightened the cognitive load, but tests did not show lessened cognitive fatigue

This the second Microsoft report covering research on AI's impact on work

FIRST REPORT

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

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)

Dec 2023

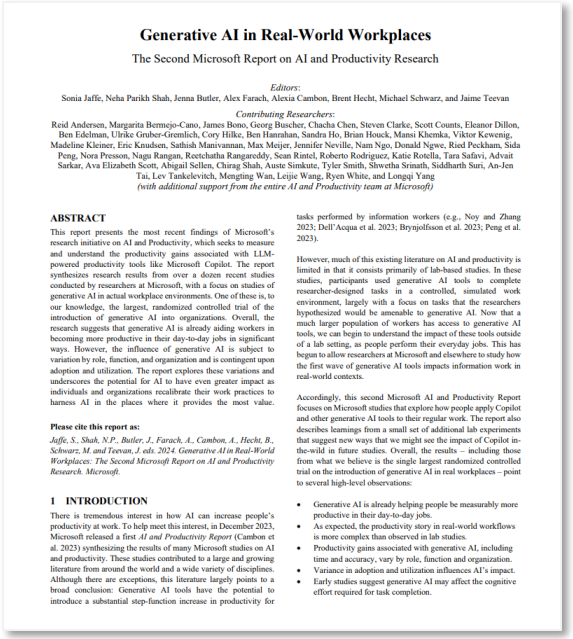
The first report focused on understanding the **potential impact** of generative AI tools as people begin to use AI for work.

- Synthesized eight Microsoft-run research studies on AI's impact on productivity.
- Most studies were lab-based, and explored tasks for which existing literature already suggested AI would do well.
- The included studies consistently found meaningful increase in speed of execution without a significant decrease in quality.
- Available at aka.ms/productivity-report1

SECOND REPORT

Generative AI in Real-World Workplaces

New



Generative AI in Real-World Workplaces

The Second Microsoft Report on AI and Productivity Research

Editors:
Sonia Jaffe, Neha Parikh Shah, Jenna Butler, Alex Farach, Alexia Cambon, Brent Hecht, Michael Schwarz, and Jaime Teevan

Contributing Researchers:
Reid Andersen, Margarita Bermejo-Cano, James Bono, Georg Buscher, Chacha Chen, Steven Clarke, Scott Counts, Eleanor Dillon, Ben Edelman, Ulrike Gruber-Gremlich, Cory Hilke, Ben Hanrahan, Sandra Ho, Brian Houck, Mansi Khemka, Viktor Kewenig, Madeline Kleiner, Eric Knudsen, Sathish Manivannan, Max Meijer, Jennifer Neville, Nam Ngo, Donald Ngwe, Ried Peckham, Sida Peng, Nora Presson, Nagu Rangan, Reetchatha Rangareddy, Sean Rintel, Roberto Rodriguez, Katie Rotella, Tara Safavi, Advait Sarkar, Ava Elizabeth Scott, Abigail Sellen, Chirag Shah, Auste Simkute, Tyler Smith, Shwetha Srinath, Siddharth Suri, An-Jen Tai, Lev Tankelevitch, Mengting Wan, Leijie Wang, Ryen White, and Longqi Yang
(with additional support from the entire AI and Productivity team at Microsoft)

ABSTRACT

This report presents the most recent findings of Microsoft's research initiative on AI and Productivity, which seeks to measure and understand the productivity gains associated with LLM-powered productivity tools like Microsoft Copilot. The report synthesizes research results from over a dozen recent studies conducted by researchers at Microsoft, with a focus on studies of generative AI in actual workplace environments. One of these is, to our knowledge, the largest, randomized controlled trial of the introduction of generative AI into organizations. Overall, the research suggests that generative AI is already aiding workers in becoming more productive in their day-to-day jobs in significant ways. However, the influence of generative AI is subject to variation by role, function, and organization and is contingent upon adoption and utilization. The report explores these variations and underscores the potential for AI to have even greater impact as individuals and organizations recalibrate their work practices to harness AI in the places where it provides the most value.

Please cite this report as:

Jaffe, S., Shah, N.P., Butler, J., Farach, A., Cambon, A., Hecht, B., Schwarz, M. and Teevan, J. eds. 2024. Generative AI in Real-World Workplaces: The Second Microsoft Report on AI and Productivity Research. Microsoft.

1 INTRODUCTION

There is tremendous interest in how AI can increase people's productivity at work. To help meet this interest, in December 2023, Microsoft released a first *AI and Productivity Report* (Cambon et al. 2023) synthesizing the results of many Microsoft studies on AI and productivity. These studies contributed to a large and growing literature from around the world and a wide variety of disciplines. Although there are exceptions, this literature largely points to a broad conclusion: Generative AI tools have the potential to introduce a substantial step-function increase in productivity for tasks performed by information workers (e.g., Noy and Zhang 2023; Dell'Acqua et al. 2023; Brynjolfsson et al. 2023; Peng et al. 2023).

However, much of this existing literature on AI and productivity is limited in that it consists primarily of lab-based studies. In these studies, participants used generative AI tools to complete researcher-designed tasks in a controlled, simulated work environment, largely with a focus on tasks that the researchers hypothesized would be amenable to generative AI. Now that a much larger population of workers has access to generative AI tools, we can begin to understand the impact of these tools outside of a lab setting, as people perform their everyday jobs. This has begun to allow researchers at Microsoft and elsewhere to study how the first wave of generative AI tools impacts information work in real-world contexts.

Accordingly, this second Microsoft AI and Productivity Report focuses on Microsoft studies that explore how people apply Copilot and other generative AI tools to their regular work. The report also describes learnings from a small set of additional lab experiments that suggest new ways that we might see the impact of Copilot in-the-wild in future studies. Overall, the results – including those from what we believe is the single largest randomized controlled trial on the introduction of generative AI in real workplaces – point to several high-level observations:

- Generative AI is already helping people be measurably more productive in their day-to-day jobs.
- As expected, the productivity story in real-world workflows is more complex than observed in lab studies.
- Productivity gains associated with generative AI, including time and accuracy, vary by role, function and organization.
- Variance in adoption and utilization influences AI's impact.
- Early studies suggest generative AI may affect the cognitive effort required for task completion.

July 2024

The new report focuses on **real-world impact** of generative AI tools as workers integrate them into everyday complex workflows.

- Synthesizes fourteen Microsoft studies, including the largest randomized controlled trial on generative AI in organizations.
- Focused on field research but includes some lab studies.
- Results suggest positive productivity effects are beginning to manifest in real-world work, and show how generative AI can provide even greater impact as work practices evolve.
- Available at aka.ms/productivity-report2

Appendix

The new report synthesizes results from 14 recent studies

Studies of workers using AI on the job

- Early Access Program Telemetry Study
- Work Trend Index Study
- Copilot Usage in the Workplace Survey
- Study on Generative Search Engines and Task Complexity

Specific Roles and Functions

- Comparing across Roles in Copilot Usage in the Workplace Survey
- Towards Effective AI Support for Developers: A Survey of Desires and Concerns
- Problem-Solving Styles and Confidence Generating Prompts for GitHub Copilot
- GitHub Copilot and Engineering System Satisfaction

A Selection of New Lab Studies

- Comparing the Effect of Different Task Types on Effective Use of GitHub Copilot
- Understanding the Impact Copilot for Security Has for Security Professionals
- Experiment with Licensing Chatbot for Sellers
- The Effect of Copilot in a Multi-lingual Context
- Impact of Generative AI on Metacognition
- Impact of Copilot on Cognitive Load

Additional details from three of the studies in the new report

Studies of workers using AI on the job

Early Access Program Telemetry Study:
Compares employees with/without Copilot

Over 60 tenants and 6k employees

Largest Controlled Study of Generative AI Productivity Impacts to date

Large-scale randomized controlled field experiment of Copilot for Microsoft 365, focusing on real-world use

	Copilot Users
Documents	• 10% more documents created and edited
Emails	• 11% fewer emails read • 4% less time interacting with emails
Meetings	• Effects differed by company

Substantial variation across tenants in adoption/usage rates. Substantial difference between effect of being given a Copilot *license* and effect of Copilot *usage*

Specific Roles and Functions

Work Trend Index Survey:
Expansive survey aimed to capture user sentiments and experiences with generative AI, broadly

31K global information workers across 31 countries

Perceived Benefits of AI Integration

- **Task Automation:** Benefits depend on the extent of tasks in function that can be automated or assisted by AI
- **Repetitive Tasks:** Most considerable improvements in roles with repetitive tasks or extensive content creation
- **High Interaction Roles:** Largest gains in productivity, quality, and efficiency in customer-facing roles (Customer Service, Sales) and content creation (Marketing, Creative)
- **Specialized Fields:** More modest improvements in technical fields (Legal, Engineering) due to the complexity and nuance of tasks

What makes a Power User?

They regularly experiment with different ways of using AI

A Selection of New Lab Studies

Generative AI in Multilingual Teams

(Lab study) Copilot improved accuracy for Japanese speakers trying to understanding an English meeting recording

	Accuracy among Japanese listeners recapping a	
	Japanese Meeting	English Meeting
With Copilot Meeting Recap	96.8%	97.5%
Standard tools	94.8%	83.8%