Measuring the impact of AI tools on worker productivity through lab experiments

Ben Edelman
Alex Farach
Donald Ngwe
Lars Roemheld

October 2024 NABE-TEC

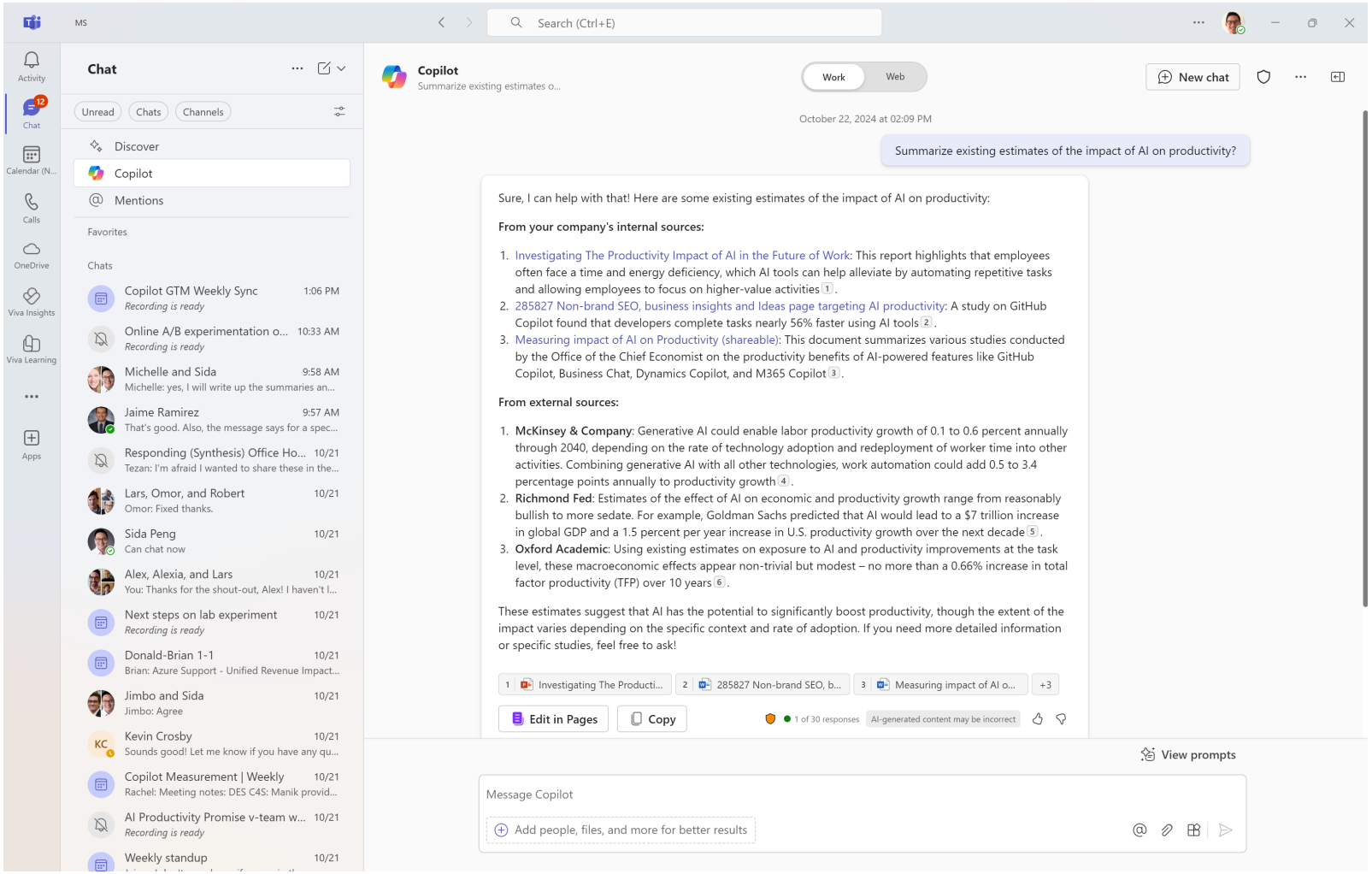
Overview

We ran a series of lab experiments in which subjects were assigned a variety of **information retrieval, synthesis, and writing** tasks.

In each experiment, we randomly assigned subjects to control and treatment conditions. Subjects in both conditions had access to Microsoft Office and subjects in the treatment group had access to **Microsoft Copilot**.

Broadly, subjects in the treatment group performed tasks **faster** and with higher **accuracy** and **quality**.

Microsoft Copilot



Example: New Product Development

We designed four distinct tasks related to a fictional product (the Osiris I laptop) to give a realistic, if idealized, view of the kind of work common when bringing new products to market.

	Analyzing a survey of market needs	**Find information from prior market research**	**Draft an announcement about the new product**	**Brainstorm possible product names**
	Excel Copilot	M365 Chat	M365 Chat or Word Copilot	M365 Chat
Sample questions	What is the most commonly mentioned unmet need perceived by laptop users? How many respondents mentioned the screen as their favorite feature?	Which processor will power the Osiris I? Which ports will the Osiris I have? Which blockbuster game will the Osiris I be pre-loaded with?	Please produce a 400-500 word blog post about the release of a new gaming laptop, Osiris I, based on the information in the Contoso SharePoint. Include details its specifications and pricing.	Propose ten possible product names for a new laptop currently known by code-named Osiris I, with the features in the company SharePoint (folder "New laptop intro").

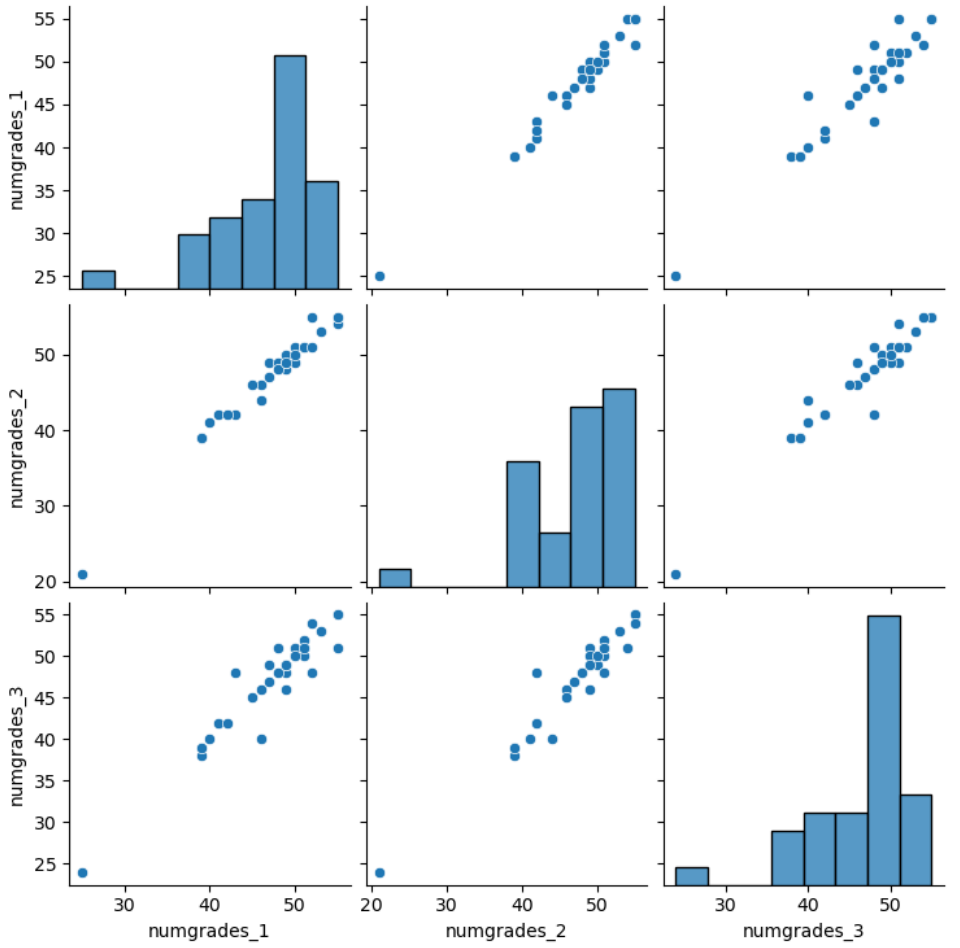
Grading freeform text using GPT

We use the OpenAI API to grade each freeform written response on accuracy.

Sample questions

1. Does the text mention that Alice needs to submit her report by Friday?
2. How well-written is the text?
3. How suitable is the text for relaying to a customer?

Correlation of AI-assigned grades with human-assigned grades ≈0.68.



QQ plot of three independent draws with temp = 0 shows high consistency of LLM grades

Example: New Product Development

	M365 Copilot Users Better than M365 Base Users	
	Accuracy/Quality	**Speed**
Overall	na	42.8%**
Analyze survey	10.9%**	30.5%**
Market research	3.8%	-0.7%
Product announcement (drafting)	Facts: -4.5% Quality: 5.4%**	65.4%**
Propose product name (ideation)	3.2%**	43.7%**

List of selected lab experiments

Features	Results	Population
Microsoft Copilot	Information retrieval (multiple-choice) 27% faster	Upwork freelancers
Microsoft Copilot	Japanese speakers match speed and accuracy of English speakers	Upwork freelancers
Microsoft Copilot	Legal work completed 32% faster with 20% higher accuracy	Microsoft legal professionals
Microsoft Copilot	M&A analysis completed 54% faster with 10% higher accuracy	Visa employees
Microsoft Copilot	40% faster on client advisory	Clifford Chance legal professionals
Microsoft Copilot	Completed marketing brief 42% faster with 12% higher quality	Unilever marketing professionals
Copilot for Azure	Tasks in Azure Portal completed with 10% higher accuracy	Upwork freelance IT admins
Licensing Q&A Skill	~50% improvement in quality and speed of response to licensing Qs	Microsoft sellers

Caveats & Limitations

Assigned tasks designed to be representative of typical office tasks, but also chosen to correspond with specific AI tool functionality

Results do not speak directly to overall productivity gains for workers or firms

Conclusion

Lab experiments consistently demonstrate increased speed, accuracy, and quality for a variety of tasks across several information worker populations

Results point to which types of tasks we can expect AI to help with, rather than generalizable productivity improvement