Working Paper 24-013

Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge

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

Navigating the Jagged Technological Frontier:

Abstract

The public release of Large Language Models (LLMs) has sparked tremendous interest in how humans will use Artificial Intelligence (AI) to accomplish a variety of

1 Introduction

The capabilities of Artificial Intelligence to produce human-like work have improved rapidly, especially since the release of OpenAI's ChatGPT, one of several Large Language

capabilities that they were not specifically created to have, and ones that are growing rapidly over time as model size and quality improve. Trained as general models, LLMs nonetheless demonstrate specialist knowledge and abilities as part of their training

Taken together, these three factors – the surprising abilities of LLMs, their ability to

2 Methods

We collected data from two randomized experiments to assess the causal impact of AI, specifically GPT-4 – the most capable of the AI models at the time of the experiments (Spring 2023) – on high human capital professionals wsuringtraditsionally iethutgAI.2

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

typically goes through, from ideation to product launch.[5] Participants responded to a total of 18 tasks (or as many as they could within the given time frame). These tasks spanned various domains. Specifically, they can be categorized into four types: creativity (e.g., “Propose at least 10 ideas for a new shoe targeting an underserved

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

Table 2 presents the results related to the percentage of task completion by subjects, which is the dependent variable in this analysis. Across Columns 1, 2, and 3, both treatments — GPT + Overview and GPT Only — demonstrate a positive effect on task completion. On average, these coefficients indicate a 12.2% increase in completion rates.[10]

3.2 Quality Disruptor - Outside the frontier

In refining the task within the frontier and recognizing the substantial quality and productivity gains from AI integration, we sought a task that AI couldn't easily complete

the outside-the-frontier experiment. The dependent variable, ‘Recommendation Quality,’

References

Acemoglu, Daron and Pascual Restrepo

Berg, Justin, Manav Raj, and Robert Seamans, *"Capturing Value from Artificial*

Eloundou, Tyna, Sam Manning, Pamela Mishkin, and Daniel Rock, *"Gpts are gpts: An early look at the labor market impact potential of large language models," arXiv preprint arXiv:2303.10130, 2023.*

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

Felten, Edward W., Manav Raj, and Robert Seamans

Soto, Christopher J. and Oliver P. John, *"Short and extra-short forms of the Big Five*

Figure 1:

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

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

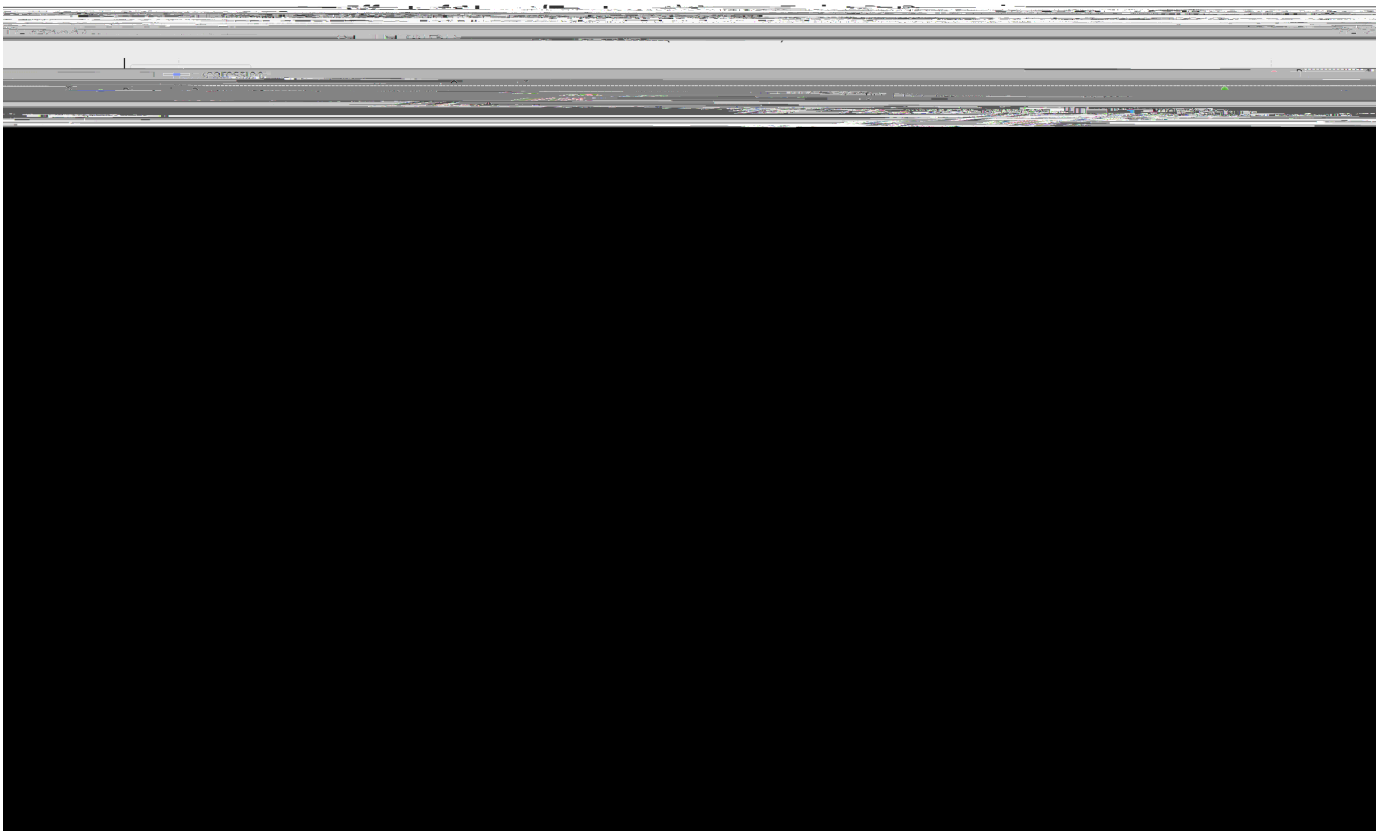

Figure 5: **Bottom-Half Skills and Top-Half Skills - Inside the Frontier**

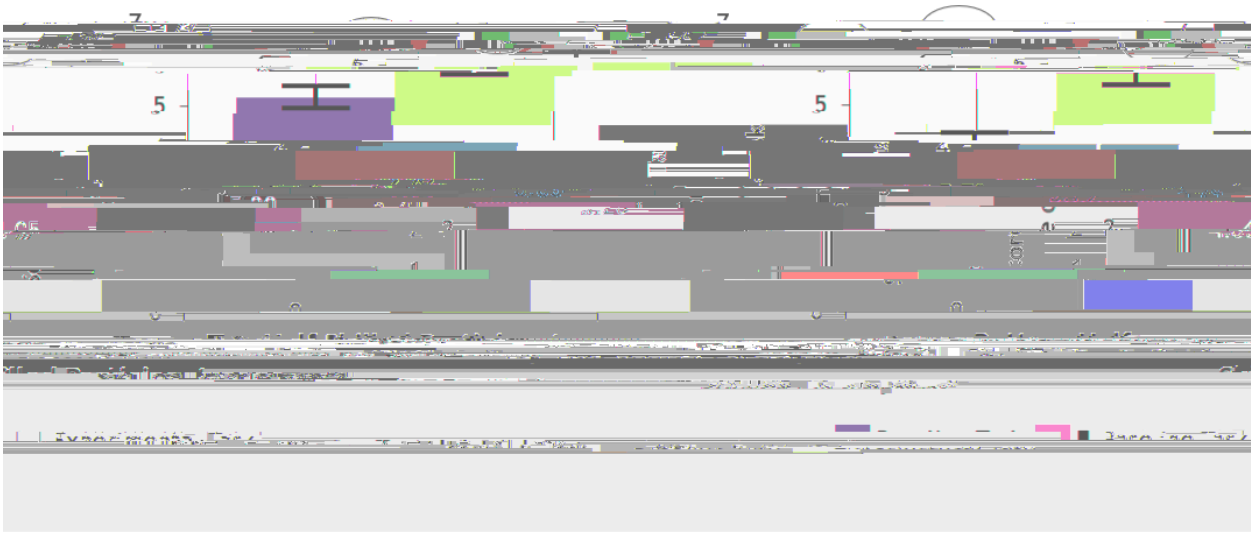

Figure 7: **Performance - Outside the Frontier**

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

Figure 8: **Recommendation Quality**

Notes: This figure displays the average performance of subjects who were correct in the experimental task outside the frontier a8qrcentage

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

	(1) Quality	(2) Quality	(3) Quality	(4) Quality (GPT)
GPT + Overview	1.746	1.752		

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

	(1) Timing	(2) Timing	(3) Timing
GPT + Overview	-1129.143		

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

	(1) Timing	(2) Timing	(3) Timing
GPT + Overview	-689.191		

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

	(1) Rec. Quality	(2) Rec. Quality	(3) Rec. Quality	(4) Rec. Quality

Appendix

A Tasks

submitted—and display it

subjects are prone to abdicating

D Collective Variation

In this Appendix

We consider the between semantic similarity measured by

!!

' ! ! $ % * & + (!) T C 2 0 T 2 1 1 . 0 4 T < 0 0 0 3 0 0 1 C > !

- P-Q1-:/9.7)-39/6290;)8E0.7-:)/6)'&)61/>1/#$%&'()*$%+$3-$20'/$/4(3-.(01$ <5683 2.3B4 21(3-$)-1(2)235($)-1(/)1-936931(3-$)%$