# GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models

Tyna Eloundou,[1] Sam Manning[1,2], Pamela Mishkin[1*], and Daniel Rock[3]

[1]OpenAI
[2]OpenResearch
[3]University of Pennsylvania – Wharton School
[*]To whom correspondence should be addressed; E-mail: pamela@openai.com

**We investigate the potential impact of large language models (LLMs) on the U.S. labor market. We develop and leverage a new rubric, assessing occupations based on their alignment with LLM capabilities, using both human and GPT-4 ratings. Our findings reveal that around 80% of the U.S. workforce have at least 10% of their work tasks exposed to the capabilities of LLMs, and approximately 19% of workers have at least 50% of their tasks exposed. We do not make predictions about the development or adoption timeline of such LLMs. LLM impact potential is pervasive, LLMs improve over time, and complementary investments will be necessary to unlock their full potential. This suggests LLMs are general-purpose technologies (*1*). As such, LLMs could have considerable economic, societal, and policy implications.**

# Introduction

Recent years, months, and weeks have seen remarkable progress in the field of generative AI and large language models (LLMs). While the public often associates LLMs with various iterations of the Generative Pre-trained Transformer (GPT), LLMs can be trained using a range of architectures, and are not limited to transformer-based models (*2*). LLMs can process and produce various forms of sequential data, including assembly language, protein sequences and chess games, extending beyond natural language applications alone. Furthermore, LLMs can enable custom search applications through embeddings and can perform tasks like summarization and classification, leveraging context provided in a prompt.

Our study considers the progress of these models' capabilities and the breadth of the complementary technologies that they spawn, underscoring that maximizing the impact of LLMs requires their integration within broader systems (*3–5*). We propose an evaluative framework for understanding the capabilities of LLMs and associated technologies and assessing their potential effects on the tasks workers perform in their jobs. This rubric (Supplementary Materials Section 3.1.2) measures task exposure to LLMs, following the spirit of prior work on quantifying exposure to machine learning (*6–8*). Following prior literature, we use the concept of exposure as a proxy for potential economic impact, irrespective of labor-augmentation or displacement effects (a detailed discussion of relevant literature can be found in Section 2 of the Supplementary Materials). We define exposure as a measure of the potential for task-level productivity to increase by 100 percent or more through the use of an LLM without a drop in quality.[1] We deploy both human annotators and GPT-4 to apply our exposure rubric to U.S. occupational data, primarily sourced from the O*NET database.[2]

For our primary exposure dataset, we collected both human annotations and GPT-4 classi-

[1] in other words, a 50 percent decrease in the time needed to complete a task.

[2] This diverges from recent social science research that makes use of LLMs to simulate human behavior (*?, 9*)

fications, using a prompt tuned for agreement with a validation subsample of labels generated by the authors.[3] When we aggregate exposure at the task level, we observe similar exposure score annotations between human and machine annotations. Overall, they agree on the task-level exposure label 65.6 percent of the time. We transform these annotations into three primary outcome measures, representing task exposure to LLMs alone ("$E1$"), partial or hybrid task exposure contingent on development of complementary systems ("$E1 + 0.5 * E2$"), and integrated exposure assuming full development of complementary systems ("$E1 + E2$").

Our findings suggest that around 19 percent of occupations exhibit at least 50 percent task exposure to LLMs, considering both present capabilities and plausible future advancements in complementary systems. However, human evaluations suggest that when confined to existing language and code abilities without further software integrations ($E1$), just 3 percent of U.S. workers have over half of their tasks exposed to LLMs. Despite these estimates, our measure of exposure as a reflection of the technical capacity of an LLM to effect the completion of a task will be mitigated in its real world impact by numerous other socioeconomic, regulatory, and adoption related factors.

Both human and GPT-4 annotations consistently show that nearly all occupations present some degree of exposure to LLMs, with the extent of exposure differing across various job types. Roles with higher wages typically show a greater degree of exposure. Our analysis projects a pervasive impact of LLMs, such as GPT-4, across sectors even without further advances in the capabilities of LLMs themselves. We also find that the potential impact of LLMs broadens significantly when we consider the evolution of complementary technologies. These collective attributes strongly suggest that LLMs such as Generative Pre-trained Transformers (GPTs) possess key characteristics of general-purpose technologies (GPTs) (*1, 10*).[4]

[3]Importantly, the GPT-4 annotations are quite sensitive to edits to the exposure rubric. For this reason we highly recommend prompt-tuning and validation against a subsample of human-generated labels when using LLMs for similar purposes.

[4]For the remainder of the paper we spell out general-purpose technologies when it is used outside of stating

# 1 Methods and Data Collection

## 1.1 Data on Activities and Tasks Performed by Occupation in the US

We draw from the O*NET 27.2 database (*11*), which provides data on 1,016 occupations, including their respective Detailed Work Activities (DWAs) and tasks. A DWA is a comprehensive set f steps required to complete a task, such as "Study scripts to determine project requirements." A task, on the other hand, is an occupation-specific unit of work that may be associated with zero, one, or multiple DWAs. For example one task linked to the preceding DWA for Actors is "Learn about characters in scripts and their relationships to each other to develop role interpretations".

We obtain employment and wage data from the 2020 and 2021 Occupational Employment series as provided by the Bureau of Labor Statistics (*12–14*). Details and examples for O*NET and BLS labor data collection are described in the Supplementary Materials Section 4.

## 1.2 Exposure

Our results are collected by applying our exposure rubric, where we define **exposure** as the capacity of an LLM or LLM-powered system to reduce the time required for a human to perform a specific DWA or complete a task by at least 50 percent while preserving or improving quality. While the term "LLM" commonly refers to a broad array of data modalities, we confine our definition of LLM-powered software to principally apply to text and code. A summary of our rubric is presented in Figure 1 Panel A, with the full rubric available in Section 3.1.2 of the Supplementary Materials.[5] In processing labels for DWAs, we first consolidate them at the task level, followed by an aggregation at the occupation level. "Core" tasks as defined by O*NET are given equal weight and "Supplemental" tasks are weighed as half a "Core" task. Results are robust to alternative weightings, including using task importance scores in O*NET.

"GPTs are GPTs."

[5]We also have an automation-focused rubric prompt available in the Supplementary Materials Section 3.2.

We set the exposure threshold at an estimated 50 percent reduction in time necessary to complete a DWA or task while maintaining quality. Though somewhat arbitrary, this threshold was selected for ease of interpretation by human annotators. It's likely that actual task time reduction in real-world scenarios would fall below our estimates due to the many other non-technical determinants of productivity impacts, hence our decision to choose a relatively high threshold for our exposure metric.

We collected both human and GPT-4-generated annotations using the exposure rubric, which underlie the bulk of the analyses in this paper.

- *Human Ratings:* We obtained human annotations by applying the rubric to each O*NET Detailed Worker Activity (DWA) and a subset of all O*NET tasks and then aggregated those DWA and task scores[6] at the task and occupation levels. The authors personally labeled a large sample of tasks and DWAs and enlisted experienced human annotators who have reviewed GPT-3, GPT-3.5 and GPT-4 outputs as part of OpenAI's alignment work (*15*). It is important to note that while the team that annotated tasks has experience working with and assessing the outputs of LLMs, they do not have experience performing many of the work tasks in the O*NET database. Expanding the sample of human raters to those in a larger range of occupations is a priority area for future research and validation of these exposure annotations.

- *GPT-4 Ratings:* We administered a similar rubric to an early version of GPT-4 (*16*) but on all task/occupation pairs rather than DWAs. We made slight modifications to the rubric (which was used as a "prompt" to the model in this case) to enhance agreement with a set of human labels. A visualization of full agreement rates is in Panel B and a Table with agreement details is in Panel D of Figure 1.

[6]The authors annotated DWAs that clearly required a high degree of physicality or manual dexterity, and the contracted annotators labeled the remaining activities, along with a subset of tasks including those without associated DWAs and those for which there was no clear task-level annotation after aggregating the DWA annotations.

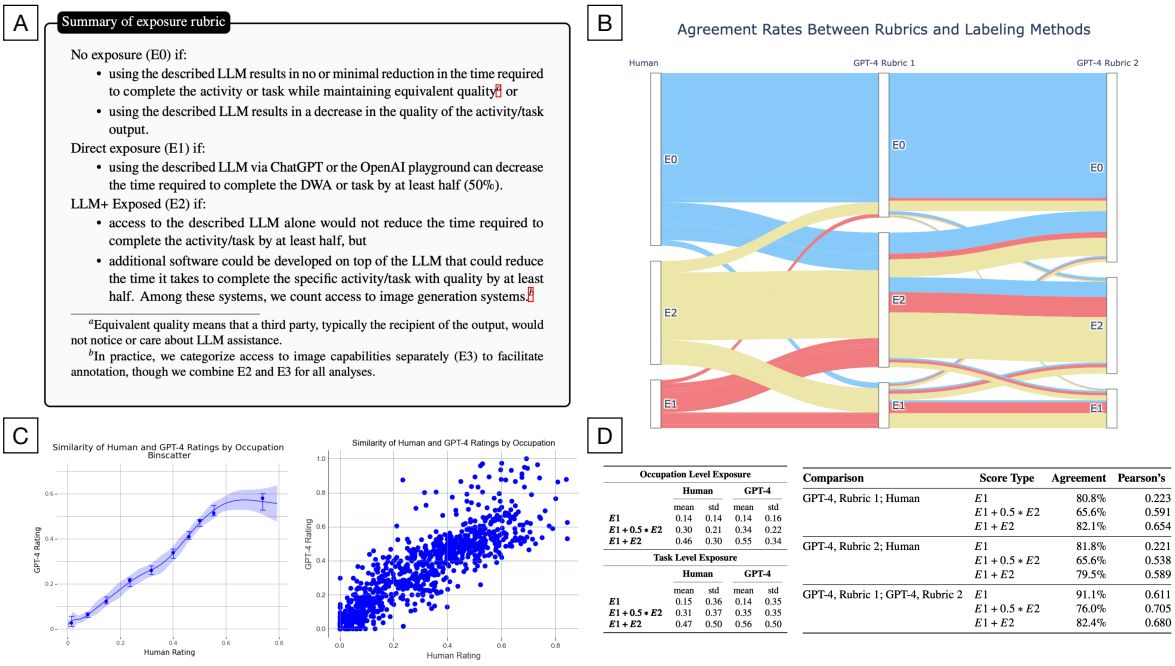
Our analysis focuses on three primary measures as dependent variables, designed to reflect different assumptions about conditions for exposure: (i) $E1$, signifies E1 from the exposure rubric, representing the lower bound of the proportion of exposed tasks within an occupation, (ii) $E1 + 0.5 * E2$, with the 0.5 weight on E2 adjusting for exposure that requires additional investment in complementary software, and (iii) $E1 + E2$, offering an upper bound of exposure via LLMs assuming full development of complementary LLM-powered software. Agreement between annotation groups and measures is detailed in Panels B and D of Figure 1 and in the Supplementary Materials Section 2.1.3. For the remainder of the analysis, unless otherwise noted, the reader may assume that "exposure" refers to $E1 + 0.5 * E2$ exposure. Panel C of Figure 1 displays a binscatter and ordinary scatterplot of the $E1 + 0.5 * E2$ measures from human raters and GPT-4. Both annotation approaches yield similar results, as is also demonstrated in Panel D. These plots include overall summary statistics for occupation-level exposures, as well as task-level ratings frequency by annotator group. Humans rated slightly more tasks as E1, fewer tasks as E2, and more tasks as E0 relative to the GPT-4 annotation approach.

# 2 Results

In the following sections, we discuss results across various roles and wage structures. Additional results on the relative exposure of industries within the U.S. economy can be found in Supplementary Materials Section 7.

## 2.1 Summary Statistics

Panel D of Figure 1 displays occupation and task summary statistics. Both human and GPT-4 annotations reveal average occupation-level $E1$ values ranging between 0.14 and 0.15, implying that approximately 15 percent of tasks within an occupation are directly exposed to LLMs, on average. This percentage increases to 30-34 percent for $E1 + 0.5 * E2$ and 46-55 percent for




**Occupation Level Exposure**

| | Human | | GPT-4 | |
|---|---|---|---|---|
| | mean | std | mean | std |
| ***E1*** | 0.14 | 0.14 | 0.14 | 0.16 |
| ***E1 + 0.5 * E2*** | 0.30 | 0.21 | 0.34 | 0.22 |
| ***E1 + E2*** | 0.46 | 0.30 | 0.55 | 0.34 |

**Task Level Exposure**

| | Human | | GPT-4 | |
|---|---|---|---|---|
| | mean | std | mean | std |
| ***E1*** | 0.15 | 0.36 | 0.14 | 0.35 |
| ***E1 + 0.5 * E2*** | 0.31 | 0.37 | 0.35 | 0.35 |
| ***E1 + E2*** | 0.47 | 0.50 | 0.56 | 0.50 |

| Comparison | Score Type | Agreement | Pearson's |
|---|---|---|---|
| GPT-4, Rubric 1; Human | $E1$ | 80.8% | 0.223 |
| | $E1 + 0.5 * E2$ | 65.6% | 0.591 |
| | $E1 + E2$ | 82.1% | 0.654 |
| GPT-4, Rubric 2; Human | $E1$ | 81.8% | 0.221 |
| | $E1 + 0.5 * E2$ | 65.6% | 0.538 |
| | $E1 + E2$ | 79.5% | 0.589 |
| GPT-4, Rubric 1; GPT-4, Rubric 2 | $E1$ | 91.1% | 0.611 |
| | $E1 + 0.5 * E2$ | 76.0% | 0.705 |
| | $E1 + E2$ | 82.4% | 0.680 |

Figure 1: Deploying the LLM Exposure Rubric

Panel A: A summary of the exposure rubric used to classify the level of LLM exposure faced by each task in the O*NET database. Panel B: Sankey diagram showing agreement between Human and GPT-4 labels in both prompt versions. Blue flows indicate human E0 labels, Yellow flows indicate human E2 labels, and Red flows are human E1 labels. There is variation across each approach. GPT-4 prompts differ primarily with respect to where the E0 label instructions were placed. Panel C: Human ratings (x-axis) and GPT-4 ratings (y-axis) on task level exposure, aggregated to the occupation level and scaled to between 0 and 1. The left plot shows a binscatter, while the right plot has an ordinary scatter. Both rating sources show a high degree of agreement about LLM exposure by occupation. Panel D: Summary statistics for exposure scores generated by humans and GPT-4, presented at the task and occupation level. Task-level exposure means represent the share of tasks across all occupations exposed at each level (E1, E1+0.5*E2, E1+E2). Occupation-level exposure means represent the average share of tasks within an occupation exposed at each level given the label scores ($E1$, $E1 + 0.5 * E2$, or $E1 + E2$). The mean and standard deviation of ratings from human and GPT-4 labels are highly similar. As expected, the variation in labels is higher for tasks than it is for occupations. Occupations, as bundles of tasks, have lower variance. The comparison table shows the agreement between two sets of scores at a time (human and GPT-4 for two separate runs of the exposure prompt and a comparison of GPT-4 in both prompts). Agreement indicates the percentage of tasks for which the labels are the same. Pearson's refers to the Pearson's correlation between task-level scores.

$E1 + E2$. Both human and GPT-4 annotations tag approximately 14-15 percent of total tasks in the dataset as exposed to LLMs. We estimate that 80 percent of workers are engaged in an occupation with at least 10 percent of tasks exposed to LLMs, contingent on the partial development of complementary systems ($E1 + 0.5 * E2$), while approximately 19 percent of workers are in an occupation where over half of its tasks are exposed (See Panel C of Figure 2 for a visualization of exposure across occupations).

Panel A of Figure 2 shows a list of the most exposed occupations under our different measures. Exposure is prevalent in occupations that generate written text or code, as well as those with routine information processing tasks. In Panel D of Figure 2, we plot the exposure by job zone. The x-axis is the percentage of tasks exposed, and the y-axis is the percentage of occupations at or above the x-axis level of exposure. Job zones 4 and 5 are the most exposed, indicating that occupations needing "extensive preparation" (e.g. lawyers, pharmacists, database administrators) are more exposed than those with lower entry barriers (dishwashers, floor sanders, baristas).[7]

Figure 2 Panel C visualizes the possible economic impact of complementary software built atop LLMs. The difference in the y-axis (the share of all occupations) between $E1$ and $E1 + E2$ at a given point along the x-axis (the share of tasks within an occupation that are exposed) yields the aggregate within-occupation exposure potential attributable to tools and software over and above direct exposure from LLMs on their own. The difference in means across all tasks between $E1$ and $E1 + E2$ of 0.42 using the GPT-4 annotations and 0.32 using the human annotations (see Figure 1 Panel D), suggests that the average effect of LLM-powered software on task-exposure could be more than double the exposure from LLMs alone (mean $E1$ of 0.14 based on both human annotations and GPT-4 annotations). While these models are already relevant to a significant portion of workers and tasks, our findings suggest the software innovations they engender could have a far-reaching impact.

[7]Job Zone data is available here: https://www.onetonline.org/help/online/zones

## 2.2 Wages and Employment

Panel C of Figure 2 illustrates exposure intensity across the economy, in terms of occupations (first plot) and total workers (second plot). Each graph point corresponds to the estimated percentage of tasks within occupations (and the share of total workers) exposed on the y-axis with a specified exposure level ($E1$, $E1 + 0.5 * E2$, and $E1 + E2$) indicated on the x-axis. For instance, human annotators determined that 2.4 percent, 18.6 percent, and 49.6 percent of workers are $E1_{50}$, $(E1 + 0.5 * E2)_{50}$, and $(E1 + E2)_{50}$ exposed (e.g. work in occupations with 50 percent of tasks exposed), respectively. The threshold of 50 percent comes from the x-axis and the percentage of workers comes from the y-axis in the right plot of Figure 2. The vertical distance between the $E1$ and the $E1 + E2$ levels at any point along the x-axis represents the exposure potential attributable to tools and applications beyond direct exposure to LLMs like ChatGPT on their own. The distribution of exposure is similar for both workers and occupations, indicating limited correlation between worker concentration in occupations and occupational exposure to LLMs and LLM-powered software.

The bottom two binscatter plots in Panel B of Figure 2 display the relationship between occupation-level exposure and wages. Despite the presence of high-exposure, low-wage and low-exposure, high-wage occupations, the overall trend in the binscatter plots suggests that higher wages are associated with increased LLM exposure. On the other hand, The correlation between LLM exposure and employment levels appears minimal. The top two binscatters show human and GPT-4 exposure ratings, aggregated at the occupation level (y-axis), plotted against the log of total employment (x-axis) in these occupations. No significant exposure variations across different employment levels are observed in either plot.

A

| Group | Occupations with highest exposure | % Exposure |
|---|---|---|
| **Human $E1$** | Interpreters and Translators | 76.5 |
| | Survey Researchers | 75.0 |
| | Poets, Lyricists and Creative Writers | 68.8 |
| | Animal Scientists | 66.7 |
| | Public Relations Specialists | 66.7 |
| **Human $E1 + 0.5 * E2$** | Survey Researchers | 84.4 |
| | Writers and Authors | 82.5 |
| | Interpreters and Translators | 82.4 |
| | Public Relations Specialists | 80.6 |
| | Animal Scientists | 77.8 |
| **Human $E1 + E2$** | Mathematicians | 100.0 |
| | Tax Preparers | 100.0 |
| | Financial Quantitative Analysts | 100.0 |
| | Writers and Authors | 100.0 |
| | Web and Digital Interface Designers | 100.0 |
| | *Humans labeled 15 occupations as "fully exposed."* | |
| **Model $E1$** | Mathematicians | 100.0 |
| | Correspondence Clerks | 95.2 |
| | Blockchain Engineers | 94.1 |
| | Court Reporters and Simultaneous Captioners | 92.9 |
| | Proofreaders and Copy Markers | 90.9 |
| **Model $E1 + 0.5 * E2$** | Mathematicians | 100.0 |
| | Blockchain Engineers | 97.1 |
| | Court Reporters and Simultaneous Captioners | 96.4 |
| | Proofreaders and Copy Markers | 95.5 |
| | Correspondence Clerks | 95.2 |
| **Model $E1 + E2$** | Accountants and Auditors | 100.0 |
| | News Analysts, Reporters, and Journalists | 100.0 |
| | Legal Secretaries and Administrative Assistants | 100.0 |
| | Clinical Data Managers | 100.0 |
| | Climate Change Policy Analysts | 100.0 |
| | *The model labeled 86 occupations as "fully exposed."* | |
| **Highest variance** | Search Marketing Strategists | 14.5 |
| | Graphic Designers | 13.4 |
| | Investment Fund Managers | 13.0 |
| | Financial Managers | 13.0 |
| | Insurance Appraisers, Auto Damage | 12.6 |

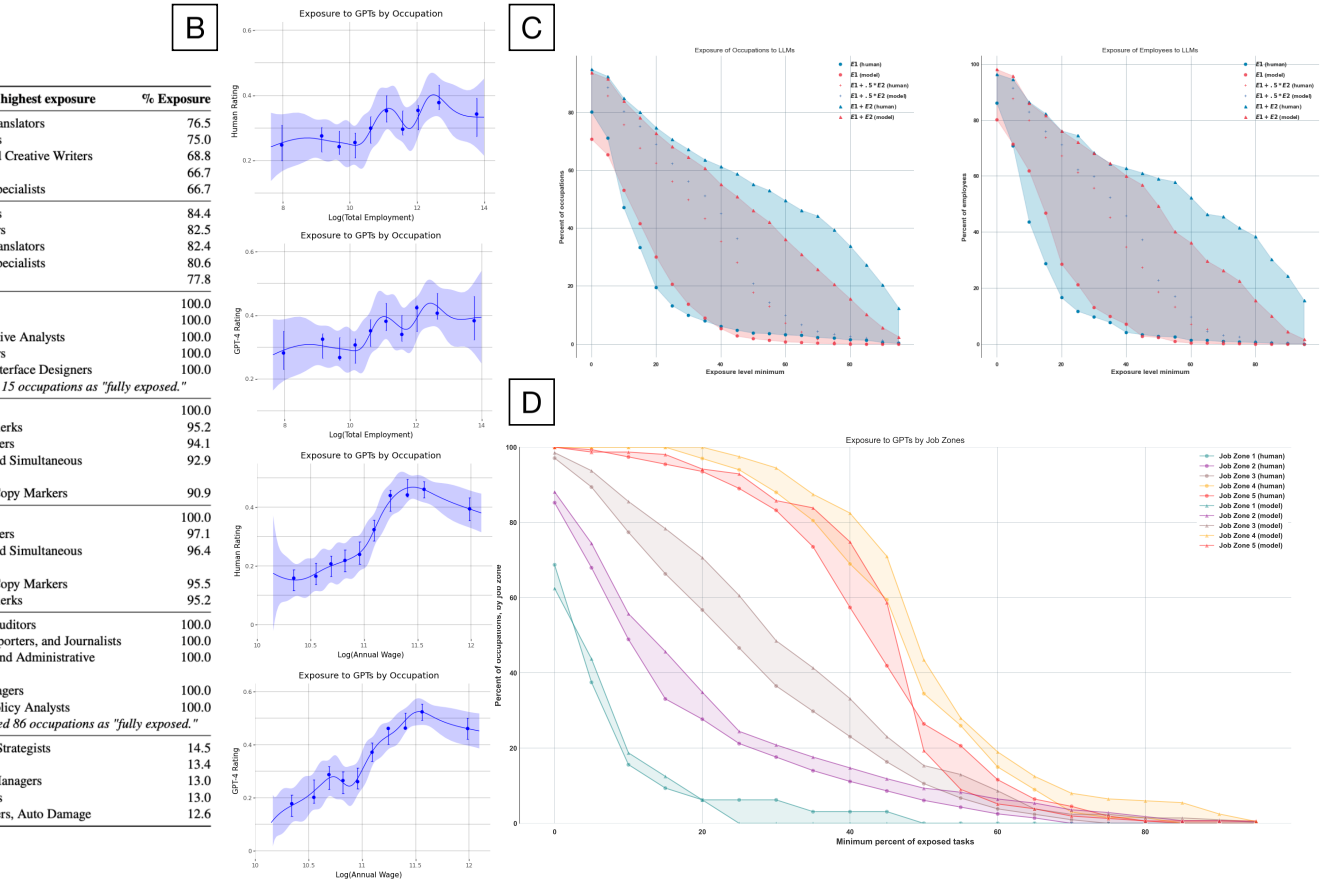



Figure 2: LLM Exposure Rubric Results

Panel A: List of occupations with the highest share of work tasks exposed to LLMs, by rating source and exposure level. Highest variance occupations are those with the highest variance in task-level scores within the occupation. Panel B: Binscatter plots depicting the exposure to LLMs in various occupations, as assessed by both humans and GPT-4. These plots compare exposure to LLMs and partial LLM-powered software (E1+0.5*E2) at the occupation level against the log of total employment within the occupation and the log of the median annual wage for the occupation. While some discrepancies exist, both human and GPT-4 assessments indicate that higher wage occupations tend to be more exposed to LLMs. Tasks classified as "core" in the O*NET database receive twice the weight of "supplemental" tasks within occupations when calculating occupation level exposure in these plots. Employment and wage data are sourced from the BLS-OES survey released in May 2022. Panel C: Exposure intensity across the economy, displayed on the left in terms of percent of affected occupations and on the right as percent of affected workers. The x-axis represents the share of tasks within an occupation exposed at the different exposure levels denoted in the figure legend. The distribution of exposure is similar across occupations and across workers, suggesting that worker concentration in occupations is not highly correlated with occupational exposure to LLMs or LLM-powered software. Panel D: E1+0.5*E2 exposure ratings aggregated to the occupation level in the five Job Zones, which are groups of similar occupations that are classified according to the level of education, experience, and on-the-job training needed to perform them. Exposure increases from Job Zone 1 to Job Zone 4, and either remains similar or decreases at Job Zone 5. Job Zones are discussed in the Supplemental Materials Section 6.1. Higher numbered job zones have greater barriers to entry in terms of training and access.

## 2.3 Robustness Checks

### 2.3.1 Comparison to Prior Measures

While our rubric-based results are consistent across human and machine ratings, we attempt to validate these findings by comparing them to similar measures from previous work. These prior measures broadly fit into two different categories: 1) data from earlier efforts to measure the potential impact of AI and related technologies on occupations and 2) observational data documenting the adoption and deployment of LLMs in the recent past.

We find that earlier measures from (*6–8, 17–19*) combined are predictive of LLM exposure, explaining between 60 and 72 percent of the variance in our $E1 + 0.5 * E2$ measures. Individually, scores related to routine manual or robotics exposures are negatively correlated with LLM exposure, while AI-related exposure scores are positively correlated with LLM exposure. A one standard deviation increase in the AI score from (*8*) or the Suitability for Machine Learning score described in (*6, 19*) is associated with a respective 0.024 and 0.055 standard deviation increase in LLM exposure from the GPT-4 Ratings conditional on other scores. The AI Occupational Exposure Score from (*7*) is strongly positively correlated with LLM scores from humans and GPT-4 in a bivariate regression as well. Panel C in Figure 3 shows the table of standardized multiple regression results for each of the $E1 + 0.5 * E2$ measures. Panel D displays comparative bivariate coefficient estimates and standard errors. The Supplementary Materials Section 9.2 contains additional details on measure comparisons as well. Results indicate that LLM exposure variation is only partially accounted for by earlier forms of technology exposure, but nevertheless a composite score of prior efforts has some predictive power. This form of data offers a hybrid between internal and external validation. Prior efforts from other researchers are consistent, but are not tied explicitly to ongoing changes in the labor market or use cases.

While targeted external validation of the rubric measurements is challenging given the nascent stage of LLM technologies, we can compare our results to what people currently use GPTs for

at work. Also, by using data on developer activity scraped from GitHub, we can validate the hypothesis that development is proceeding at a pace consistent with technology improvement over time.

### 2.3.2 Comparison to Self-Reported ChatGPT Use Cases

In February 2023 OpenAI invited members of the OpenAI Discord Channel to participate in a voluntary survey to gather information about how they are using ChatGPT in their work. Respondents who opted in to participating in the survey, represent both a variety of occupations and a range of countries across the globe (see Supplementary Materials Table 15). The survey asked members whether they used ChatGPT in their current jobs and specifically asked them, "What are some ways that you use ChatGPT in your work? Please list and describe up to three." Rating the collected responses using our GPT-4 approach, and comparing ratings for reported GPT use cases to our own ratings can provide a view into the rubric's external validity. Figure 3 Panel A shows a two-dimensional t-SNE embedding plot where each task in O*NET and each task sourced from the Discord survey are visualized. We find very few Discord-sourced tasks (shown as black circles) in the region of the plot with many unexposed (E0) tasks. On the other hand, the Discord users commonly report using GPTs for tasks that we rate as E1 or E2. After matching Discord response tasks to their closest cosine similarity-based match in O*NET tasks, 3 percent of those O*NET tasks were rated as E0, 62 percent were E1, and 35 percent as E2. We describe the survey and results in the Supplementary Materials Section 9.4.

### 2.3.3 Impact of LLM Deployment on Software Development

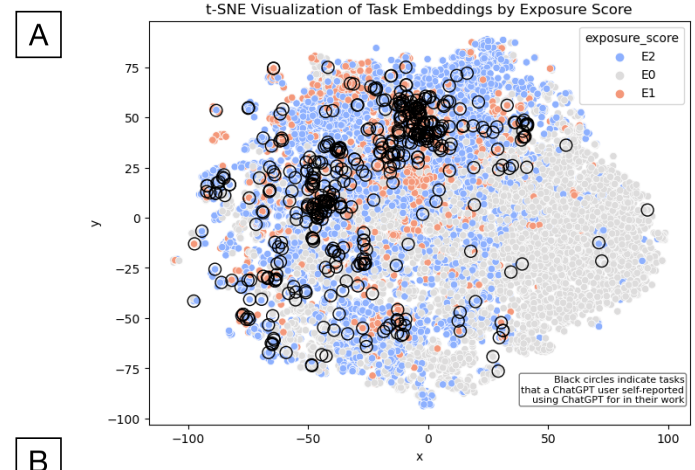
Understanding the impact of LLM deployment on software development is critical to assessing their potential as GPTs. To test the hypothesis that the release of ChatGPT caused an increase in development activity, we obtained data on forks over time from 1,167 public Github repositories that, based on the content of their Readme files, descriptions, and code, can be reasonable

assumed to be differentially impacted by the release of a LLM like ChatGPT. Applying multiple difference-in-difference specifications, results show that the launch of ChatGPT caused an average increase of approximately 7-8 forks per week in LLM-related repositories in the 20 weeks after it was released. Comparing before (fewer than 10,000 forks monthly across LLM-exposed repositories) to after (more than 50,000 forks monthly), there is a clear boom in generative AI development. We report the growth of open-source development efforts as a result of ChatGPT in the Supplementary Materials Section 9.3, but Panel D of Figure 3 shows time-series and synthetic difference-in-difference results indicating an extremely rapid development ramp in the open-source toolkit for software development related to LLMs. While GitHub forks measure input effort more than output capabilities, this effect of ChatGPT on GitHub activity provides evidence for the general-purpose technology criterion that we should expect improvement over time and that we should expect the development of complimentary innovations (in this case related software products).

# 3 Discussion

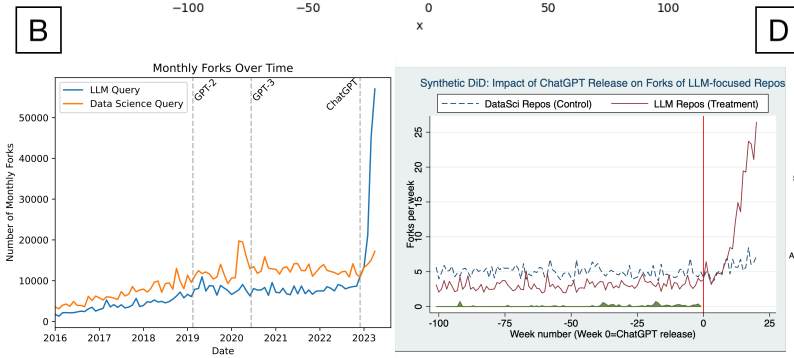
## 3.1 GPTs as General-Purpose Technologies

Considering large language models as general-purpose technologies requires that they meet three main criteria: improvement over time, pervasiveness throughout the economy, and the ability to spawn complementary innovations (*10*). Evidence from the AI and machine learning literature demonstrates that LLMs undeniably meet the first criterion – they are improving in capabilities over time, capable of handling an expanding set of complex tasks and use-cases (see Supplementary Materials Section 2.1). We primarily present evidence to support the last two criteria, showing that LLMs on their own can have a pervasive effect across the economy, and that complementary innovations enabled by LLMs – particularly via software and digital tools – can have widespread application to economic activity.




C

| | GPT-4 Exposure Rating 1 | | GPT-4 Exposure Rating 2 | | Human Exposure Rating | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Software (Webb) | 0.03406*** | 0.03705*** | 0.03347*** | 0.03580*** | 0.02893*** | 0.03032*** |
| | (0.00942) | (0.00942) | (0.00940) | (0.00935) | (0.00940) | (0.00945) |
| Robot (Webb) | −0.10799*** | −0.11568*** | −0.10777*** | −0.11399*** | −0.10609*** | −0.10957*** |
| | (0.00909) | (0.00877) | (0.00982) | (0.00941) | (0.00834) | (0.00806) |
| AI (Webb) | 0.02370*** | 0.02667*** | 0.01062 | 0.01330 | 0.01996** | 0.02116** |
| | (0.00897) | (0.00871) | (0.00896) | (0.00875) | (0.00902) | (0.00878) |
| Suitability for Machine Learning | 0.05445*** | 0.04959*** | 0.05250*** | 0.04841*** | 0.03599*** | 0.03389*** |
| | (0.00831) | (0.00815) | (0.00812) | (0.00801) | (0.00736) | (0.00717) |
| Routine Cognitive | 0.05663*** | 0.05892*** | 0.04069*** | 0.04303*** | 0.03061*** | 0.03139*** |
| | (0.00760) | (0.00767) | (0.00749) | (0.00754) | (0.00576) | (0.00574) |
| Routine Manual | −0.11213*** | −0.11439*** | −0.09446*** | −0.09618*** | −0.11111*** | −0.11218*** |
| | (0.00790) | (0.00794) | (0.00821) | (0.00823) | (0.00745) | (0.00748) |
| AI Occupational Exposure Score | 0.00695 | 0.01726** | −0.01077 | −0.00186 | 0.00441 | 0.00877 |
| | (0.00775) | (0.00741) | (0.00812) | (0.00780) | (0.00643) | (0.00592) |
| Frey & Osborne Automation | −0.01143* | −0.01493** | −0.00138 | −0.00460 | −0.01471** | −0.01608** |
| | (0.00694) | (0.00696) | (0.00759) | (0.00746) | (0.00712) | (0.00702) |
| Log Avg. Salary | 0.05804*** | | 0.04863*** | | 0.02531 | |
| | (0.01870) | | (0.01860) | | (0.01727) | |
| Constant | −0.31072 | 0.32778*** | −0.27579 | 0.25903*** | 0.01717 | 0.29577*** |
| | (0.20590) | (0.00493) | (0.20502) | (0.00520) | (0.19008) | (0.00468) |
| N | 680 | 681 | 680 | 681 | 680 | 681 |
| $R^2$ | 0.68741 | 0.68212 | 0.60737 | 0.60198 | 0.71213 | 0.71126 |



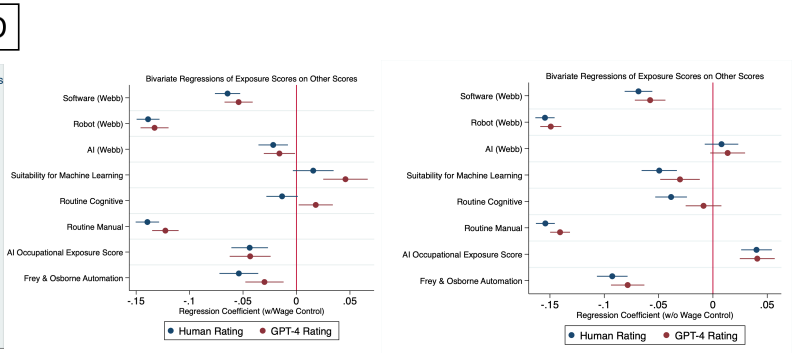




Figure 3: Validating the LLM Exposure Rubric

Panel A: Embeddings for all O*NET tasks plotted in 2-D space and colored according to the GPT-4 exposure score annotations for the given task. Black circles denote the 460 unique O*NET tasks whose text embeddings were most similar to the tasks that a sample of 954 ChatGPT users self-reported using ChatGPT for in their work. See the Supplementary Materials Section 9.4 for a table of the most common tasks users reported. Panel B: Left: Monthly forks in Github repositories with the terms "LLM" "language model" "code generation" "LaMDA" "GPT" or "BERT" (LLM Query) in the repository readme or code files plotted with monthly forks in Github repositories with the terms "statistics" "data science" "linear regression" "time series" or "statistical analysis" (Data Science Query). Right: Synthetic difference-in-difference plot of pre- and post-ChatGPT release trends in weekly forks across repository groups. Weights used to average the pre-treatment time-periods across groups for the synthetic difference-in-difference are presented along the x-axis. See supplementary materials for details on data collection and analysis in Section 9.3. Panel C: Regression of LLM exposure scores aggregated to occupations on prior measures of occupational exposure to AI and automation. Log Avg. Salary is based on annualized wages from the BLS-OES survey in May 2022. Columns 1 and 2 are based on our main $E1 + 0.5 * E2$ exposure measures given by GPT-4. Columns 3 and 4 are based on an alternative exposure rubric and rated by GPT-4 with the same $E1 + 0.5 * E2$ scoring. Columns 5 and 6 are based on exposure scores given by humans on the same rubric as columns 1 and 2. All standard errors are robust, with statistical significance at the 10%, 5%, and 1% respectively indicated by *,**,*** . All dependent variables and covariates except logged wages are standardized. We include software, robotics, and AI scores from (*8*), Suitability for Machine Learning ("SML") from (*6, 19*), Routine Cognitive and Manual scores from (*17*), AI Occupational Exposure scores from (*7*), and automation scores from (*18*). Panel D: Bar plot of bivariate correlations with exposure scores from previous studies described for Panel C. The GPT-4 and human ratings correspond to the identical measures in columns 1-4 in Panel C. These two graphs depict regression coefficients for isolated regressions of LLM exposure on each other score individually, with and without wage controls.

### 3.2 Limitations and Future Work

Alongside the limitations discussed above, we acknowledge certain specific constraints of this study warranting further exploration in the Supplementary Materials Section 9.1. Limitations discussed at length in the Supplementary Materials include the subjectivity of human judgements when rating task exposure, the brittleness of GPT-4 as a classifier in this context, and the validity of the task-based framework to assess exposure to technology at a worker or occupation level. Furthermore, our focus on the United States restricts the applicability of our findings to other nations with differing industrial organization, technological infrastructure, regulatory frameworks, linguistic diversity, and cultural contexts. We hope to address this limitation by extending the study's scope and by sharing our methods to enable further research.

## 4 Conclusion

Our study develops a new rubric of LLM capabilities and examines the potential impact of LLMs on various occupations and industries within the U.S. economy. We observe that most occupations exhibit some degree of exposure to LLMs. Higher-wage occupations generally exhibit more exposure, with approximately 19 percent of jobs having at least 50 percent of their tasks exposed to LLMs, considering current model capabilities and anticipated LLM-powered software.

The need for future research is apparent, with opportunities to delve into broader implications of LLM advancements—specifically their potential to augment or displace human labor, impact job quality, inequality, skill development, and other outcomes. Through a better understanding of LLM capabilities and their possible effects on the workforce, policymakers and stakeholders can make well-informed decisions, efficiently navigating the complex landscape of AI and its implications for the future of work.

# References


1. T. F. Bresnahan, M. Trajtenberg, *Journal of econometrics* **65**, 83 (1995).

2. J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, *ArXiv* **abs/1810.04805** (2019).

3. T. Bresnahan, Artificial intelligence technologies and aggregate growth prospects (2019).

4. A. K. Agrawal, J. S. Gans, A. Goldfarb, Ai adoption and system-wide change, *Tech. rep.*, National Bureau of Economic Research (2021).

5. A. Goldfarb, B. Taska, F. Teodoridis, *Research Policy* **52**, 104653 (2023).

6. E. Brynjolfsson, T. Mitchell, D. Rock, *AEA Papers and Proceedings* **108**, 43 (2018).

7. E. W. Felten, M. Raj, R. Seamans, *AEA Papers and Proceedings* **108**, 54 (2018).

8. M. Webb, The impact of artificial intelligence on the labor market (2020). Working paper, Stanford University.

9. J. J. Horton, *arXiv preprint arXiv:2301.07543* (2023).

10. R. G. Lipsey, K. I. Carlaw, C. T. Bekar, *Economic transformations: general purpose technologies and long-term economic growth* (Oup Oxford, 2005).

11. O*NET, O*net 27.2 database (2023).

12. BLS, Employment by detailed occupation (2022).

13. BLS, Occupational outlook handbook a-z index (2023).

14. BLS, Demographic characteristics (cps) (2023).

15. L. Ouyang, *et al.*, *arXiv preprint arXiv:2203.02155* (2022).

16. OpenAI, Gpt-4 technical report, *Tech. rep.*, OpenAI (2023).

17. D. Acemoglu, D. Autor, *Handbook of Labor Economics*, O. Ashenfelter, D. Card, eds. (Elsevier, 2011), vol. 4 of *Handbook of Labor Economics*, chap. 12, pp. 1043–1171.

18. C. B. Frey, M. A. Osborne, *Technological Forecasting and Social Change* **114**, 254 (2017).

19. E. Brynjolfsson, M. R. Frank, T. Mitchell, I. Rahwan, D. Rock, *Forthcoming* (2023).

20. T. Brown, *et al.*, *Advances in neural information processing systems* **33**, 1877 (2020).

21. A. Radford, *et al.*, *OpenAI blog* **1**, 9 (2019).

22. D. Hernandez, J. Kaplan, T. Henighan, S. McCandlish, *arXiv preprint arXiv:2102.01293* (2021).

23. J. Kaplan, *et al.*, *arXiv preprint arXiv:2001.08361* (2020).

24. R. Thoppilan, *et al.*, *arXiv preprint arXiv:2201.08239* (2022).

25. Y. Bai, *et al.*, Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback (2022). ArXiv:2204.05862 [cs].

26. T. Schick, *et al.*, *arXiv preprint arXiv:2302.04761* (2023).

27. G. Mialon, *et al.*, *arXiv preprint arXiv:2302.07842* (2023).

28. H. Chase, LangChain (2022).

29. A. Abid, M. Farooqi, J. Zou, *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21 (Association for Computing Machinery, New York, NY, USA, 2021), p. 298–306.

30. P. Schramowski, C. Turan, N. Andersen, C. A. Rothkopf, K. Kersting, *Nature Machine Intelligence* **4**, 258 (2022).

31. J. A. Goldstein, *et al.*, Generative language models and automated influence operations: Emerging threats and potential mitigations (2023).

32. OpenAI, Gpt-4 system card, *Tech. rep.*, OpenAI (2023).

33. M. Chen, *et al.*, *arXiv preprint arXiv:2107.03374* (2021).

34. S. Peng, E. Kalliamvakou, P. Cihon, M. Demirer, *arXiv preprint arXiv:2302.06590* (2023).

35. A. K. Singla, E. Horvitz, P. Kohli, A. Krause, *AAAI Conference on Human Computation & Crowdsourcing* (2015).

36. D. Shahaf, E. Horvitz, *Proceedings of the AAAI Conference on Artificial Intelligence* (2010).

37. H. Khlaaf, P. Mishkin, J. Achiam, G. Krueger, M. Brundage, A hazard analysis framework for code synthesis large language models (2022).

38. L. Weidinger, *et al.*, *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22 (Association for Computing Machinery, New York, NY, USA, 2022), p. 214–229.

39. I. Solaiman, *et al.*, Release strategies and the social impacts of language models (2019).

40. L. F. Katz, K. M. Murphy, *The quarterly journal of economics* **107**, 35 (1992).

41. D. H. Autor, F. Levy, R. J. Murnane, *The Quarterly journal of economics* **118**, 1279 (2003).

42. D. Acemoglu, P. Restrepo, *American economic review* **108**, 1488 (2018).

43. D. Acemoglu, P. Restrepo, *Journal of Economic Perspectives* **33**, 3 (2019).

44. D. H. Autor, L. F. Katz, M. S. Kearney, *American economic review* **96**, 189 (2006).

45. J. Van Reenen, *Labour economics* **18**, 730 (2011).

46. D. Acemoglu, P. Restrepo, *Econometrica* **90**, 1973 (2022).

47. B. Meindl, M. R. Frank, J. Mendonça, *arXiv preprint arXiv:2110.13317* (2021).

48. E. Felten, M. Raj, R. Seamans, *arXiv preprint arXiv:2303.01157* (2023).

49. S. Tolan, *et al.*, *Journal of Artificial Intelligence Research* **71**, 191 (2021).

50. M. Arntz, T. Gregory, U. Zierahn, *Economics Letters* **159**, 157 (2017).

51. K. Grace, J. Salvatier, A. Dafoe, B. Zhang, O. Evans, *Journal of Artificial Intelligence Research* **62**, 729 (2018).

52. D. Autor, C. Chin, A. M. Salomons, B. Seegmiller, New frontiers: The origins and content of new work, 1940–2018, *Tech. rep.*, National Bureau of Economic Research (2022).

53. M. R. Frank, *et al.*, *Proceedings of the National Academy of Sciences* **116**, 6531 (2019).

54. R. Bommasani, *et al.*, *arXiv preprint arXiv:2108.07258* (2021).

55. A. Korinek, Language models and cognitive automation for economic research, *Tech. rep.*, National Bureau of Economic Research (2023).

56. E. R. Mollick, L. Mollick, *Available at SSRN* (2022).

57. S. Noy, W. Zhang, *Available at SSRN 4375283* (2023).

58. J. Bessen, *The economics of artificial intelligence: an agenda* (University of Chicago Press, 2018), pp. 291–307.

59. A. Korinek, J. E. Stiglitz, *The economics of artificial intelligence: An agenda* (University of Chicago Press, 2018), pp. 349–390.

60. D. Acemoglu, D. Autor, J. Hazell, P. Restrepo, Ai and jobs: Evidence from online vacancies, *Tech. rep.*, National Bureau of Economic Research (2020).

61. S. G. Benzell, L. J. Kotlikoff, G. LaGarda, V. Y. Ye, Simulating endogenous global automation, *Working Paper 29220*, National Bureau of Economic Research (2021).

62. T. Bresnahan, S. Greenstein, D. Brownstone, K. Flamm, *Brookings Papers on Economic Activity. Microeconomics* **1996**, 1 (1996).

63. T. F. Bresnahan, E. Brynjolfsson, L. M. Hitt, *The quarterly journal of economics* **117**, 339 (2002).

64. J. Dixon, B. Hong, L. Wu, *Management Science* **67**, 5586 (2021).

65. P. A. David, *The American Economic Review* **80**, 355 (1990).

66. T. F. Bresnahan, *The economic journal* **109**, 390 (1999).

67. C. B. Frey, *The Technology Trap* (Princeton University Press, 2019).

68. E. Brynjolfsson, D. Rock, C. Syverson, *American Economic Journal: Macroeconomics* **13**, 333 (2021).

69. J. J. Feigenbaum, D. P. Gross, Organizational frictions and increasing returns to automation: Lessons from at&t in the twentieth century, *Tech. rep.*, National Bureau of Economic Research (2021).

70. D. Rock, *Available at SSRN 3427412* (2019).

71. T. Babina, A. Fedyk, A. He, J. Hodson, *Firm Growth, and Product Innovation (November 9, 2021)* (2021).

72. N. Zolas, *et al.*, Advanced technologies adoption and use by us firms: Evidence from the annual business survey, *Tech. rep.*, National Bureau of Economic Research (2021).

73. I. M. Cockburn, R. Henderson, S. Stern, *The economics of artificial intelligence: An agenda* (University of Chicago Press, 2018), pp. 115–146.

74. Z. Cheng, D. Lee, P. Tambe, *Available at SSRN* (2022).

75. T. Besiroglu, N. Emery-Xu, N. Thompson, *arXiv preprint arXiv:2212.08198* (2022).

76. M.-H. Huang, R. T. Rust, *Journal of service research* **21**, 155 (2018).

77. E. Brynjolfsson, T. Mitchell, *Science* **358**, 1530 (2017).

78. D. Acemoglu, D. Autor, *Handbook of labor economics* (Elsevier, 2011), vol. 4, pp. 1043–1171.

79. L. Kogan, D. Papanikolaou, L. D. W. Schmidt, B. Seegmiller, Technology, vintage-specific human capital, and labor displacement: Evidence from linking patents with occupations, *Working Paper 29552*, National Bureau of Economic Research (2021).

80. D. Arkhangelsky, S. Athey, D. A. Hirshberg, G. W. Imbens, S. Wager, *American Economic Review* **111**, 4088 (2021).

81. D. Clarke, D. Pailañir, S. Athey, G. Imbens, Synthetic difference-in-differences estimation, *Tech. rep.*, IZA Discussion Paper (2023).

# 1 Supplementary Materials

The following sections contain all supporting analysis and discussion for the main article above.

# 2 Literature Review

In this section, we explore the relevant literature in greater detail, highlighting technological and economic perspectives.

## 2.1 The Advancement of Large Language Models

The past few years have seen a surge in the development of generative AI models, drawing considerable interest from the AI research community and the public alike. Their abilities to handle a variety of complex language tasks have been amplified by factors such as an increase in model parameters, more training data, and advanced training configurations (*20–23*). Leading LLMs like LaMDA (*24*) and GPT-4 (*16*) demonstrate versatility in applications such as translation, classification, creative writing, and code generation—tasks that traditionally required specialized models crafted by expert engineers using domain-specific data.

Parallel advancements in enhancing the steerability and reliability of LLMs, through methods such as fine-tuning and reinforcement learning with human feedback, augment the models' ability to perceive user intent, thus improving their usability and functionality (*15, 25*). Furthermore, recent research uncovers LLMs' potential to direct other digital tools such as APIs, search engines and other generative AI systems - a capability that allows a more fluid integration for superior utility and performance (*26–28*). In an ultimate scenario, these developments could usher in an era where LLMs are capable of executing any task conventionally performed at a computer.

Generative AI models have been predominantly utilized as task-specific modules for functions such as image generation from captions or speech-to-text conversion. However, we posit that

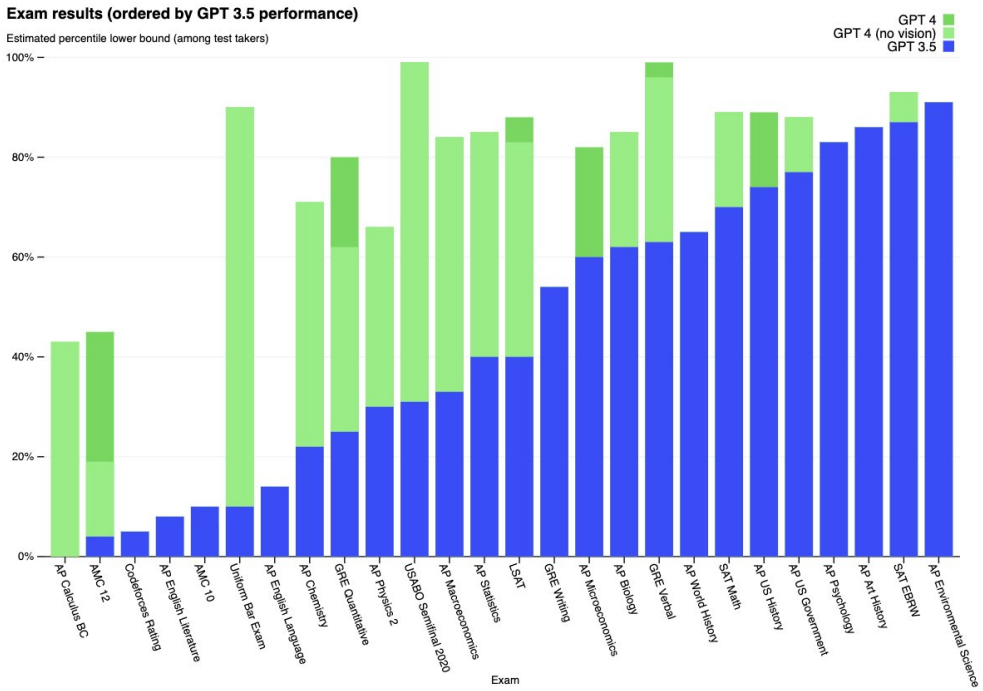



Figure 4: For a sense of how quickly model capabilities are progressing – consider the jump in exam performance between GPT-3.5 and GPT-4 (*16*).

the versatility of LLMs should be recognized as foundational elements for developing auxiliary tools. The cultivation of these tools and their integration into various systems may necessitate considerable reconfiguration of existing processes across industries and will inevitably take time.

Nonetheless, emerging trends demonstrate a growing adoption of LLMs, despite inherent limitations. They are being increasingly woven into specialized applications across areas such as writing assistance, coding, and legal research, enabling both businesses and individuals to incorporate LLMs into their workflows. As exemplified by GPT-4's substantial improvement over its predecessor GPT-3.5, these models are rapidly evolving, with their advancements readily apparent in our result analysis. Noticeable improvements are visible in Figure 4.

The relevance of accompanying technologies with large language models (LLMs) is underscored, as inherent issues of factual inaccuracies, biases, privacy threats, and disinformation risks may compromise the reliability of standalone, general-purpose LLMs (*29–32*). Nonetheless,

domain-specific workflows, encompassing specialized tooling, software, or human-supervised systems, can help to mitigate these limitations.

For instance, Casetext leverages LLMs to enhance legal research tools, aiding legal professionals in achieving more accurate research outcomes promptly, thereby offsetting potential inaccuracies of GPT-4 regarding legal cases. GitHub Copilot, a coding assistant employing LLMs, offers code generation and completion, with user-based accept/reject capabilities harnessing user expertise. Hence, while GPT-4 lacks innate contextual awareness, integrative solutions can provide it with metaphorical timekeeping abilities.

An emergent positive feedback loop may be realized as LLMs cross certain performance benchmarks, enabling them to assist in crafting the tools that enhance their functionality across diverse applications. This could reduce the financial and technical requirements for tool development, possibly expediting the uptake and integration of LLMs (*33, 34*).

LLMs can also be employed as valuable resources in the development of machine learning models, serving as coding assistants, data labeling services, or synthetic data generators. They could even influence economic decisions at the task level, for example, by optimizing task distribution between humans and machines (*35, 36*). As LLMs evolve and align more closely with user preferences, we can expect their performance to improve continually. However, it is imperative to acknowledge the serious risks accompanying these developments (*37–39*).

## 2.2 The Economic Impacts of Automation Technologies

An extensive literature addresses the labor market impacts of AI and automation technologies. The concept of skill-biased technological change and the task model of automation—often considered the standard framework for understanding technology's influence on labor—originated from research demonstrating that technological progress raises the demand for skilled workers over unskilled workers (*40*). This standard framework has seen expansions, exploring automation's

effects on task-based work (*17, 41, 42*). Such research reveals workers engaged in routine tasks to be more vulnerable to technology-induced displacement, termed routine-biased technological change. More recent research differentiates between task-displacement and task-reinstatement effects of technology, where new technology necessitates a broader spectrum of labor-intensive tasks (*42, 43*). he consequential wage disparity, attributed to wage reductions for routine-task workers, has been affirmed by several studies, underscoring the reshaping of the U.S. labor market by automation technologies (*44–46*).

Previous research methods to estimate the intersection of AI capabilities and worker tasks include patent-to-task mapping (*8, 47*), aligning AI capabilities with occupational abilities in O*NET (*7, 48*), comparing AI task benchmarks with worker tasks via cognitive abilities (*49*), labeling automation potential for US occupations and employing machine learning classifiers for estimation (*18*), task-level automation modeling and aggregation to occupational insights (*50*), and collecting expert forecasts (*51*). Pertinent to our paper, a new rubric to assess the suitability of worker activities for machine learning was introduced (*6, 19*). These approaches have noted diversified AI technology exposure at the task-level within occupations, suggesting it is rare for any occupation to be completely subsumed by AI tools. Positive correlation has been found between automation and augmentation exposures (*52*), and a growing body of research investigates the specific economic impacts and opportunities of LLMs (*34, 48, 53–57*). Our work contributes to this literature, offering insights into the potential applicability of language models within the labor market.

General-purpose technologies (e.g. printing, the steam engine, electrification, computing) are characterized by widespread proliferation, continuous improvement, and the generation of complementary innovations (*1, 10*). Their far-reaching consequences, which unfold over decades, are difficult to anticipate, particularly in relation to labor demand (*58–61*). The realization of general-purpose technologies' full potential requires extensive co-invention (*1, 10, 62–64*), a

costly and time-consuming process involving the discovery of new business procedures (*65–69*). Consequently, many studies of machine learning technologies focus on systems-level adoption, arguing that organizational systems may require redesign to effectively take advantage of novel machine learning advancements (*3–5*). Appropriately designed systems can yield considerable business value and improve firm performance (*70–72*), with AI tools facilitating the discovery process (*73–75*). (*5*) argue that machine learning as a broad category is likely a general-purpose technology. Our evidence supports a wider impact, as even subsets of machine learning software meet the criteria for general-purpose technology status independently. This paper's primary contributions are to provide a set of measurements of LLM impact potential and to demonstrate the use case of applying LLMs to develop such measurements efficiently and at scale. Additionally, we showcase the general-purpose potential of LLMs. By employing task-level information to assess whether LLMs fulfill the criteria of a general-purpose technology, we seek to merge the two perspectives for understanding the technology-labor relationship.

We attempt to build on these diverse literature streams in several ways. Echoing (*48*), we focus our analysis on the impact of LLMs, rather than addressing machine learning or automation technologies more broadly. Additionally, we propose a novel method that employs LLMs, specifically GPT-4, to assess tasks for exposure and automation potential, thereby bolstering human scoring efforts. Subsequently, we aggregate our findings to occupations and industries, capturing the overall potential exposure in the contemporary U.S. labor market.

While task impact potential is extensive, integration of LLMs and LLM-powered software into broader systems is crucial for maximizing potential. The diffusion of such general-purpose technologies into economic applications may initially encounter co-invention barriers since predicting the need for human oversight remains challenging, especially for tasks where model capabilities equal or exceed human proficiency. Although initial human supervision requirements may decelerate technology diffusion, users are likely to become increasingly familiar with the

LLMs and LLM-powered systems over time, particularly in discerning when and how to trust their outputs.

It's conceivable that for many tasks, the value of time savings and seamless integration will outweigh that of quality improvements. It's also possible that the initial focus of development will be on augmentation, followed by automation (*76*). This could potentially materialize as an augmentation phase where jobs initially become more uncertain (for instance, writers becoming freelancers) before tasks are fully automated.

General-purpose technologies are relatively rare and characterized by their pervasiveness, improvement over time, and the development of significant co-invention and spillovers (*10*). We provide some evidence of rapid growth in new LLM-related software development caused by the launch of ChatGPT in Section 9.3 of the Supplementary Materials, but for the most part take LLM improvement over time as a previously established phenomena (see review of literature on the advancement of LLMs in the supplementary materials). Our tests to determine if "GPTs are GPTs" are primarily tests of exposure pervasiveness and whether complementary innovations are needed to unlock otherwise feasible LLM applications. We find evidence that both of these conditions are likely.

If "GPTs are GPTs," the eventual trajectory of LLM development and application may be challenging for policymakers to predict and regulate. As with other general-purpose technologies, much of these algorithms' potential will emerge across a broad range of economically valuable use cases, including the creation of new types of work (*42, 52*), capital and labor augmentation, and automation. Our research serves to measure what is technically feasible now, but necessarily will miss the evolving impact potential of the LLMs over time.

# 3 Rubrics

In our discussion, we interchangeably use "LLMs" and "GPTs", aligning these terminologies with the GPT-family models accessible via platforms such as ChatGPT or the OpenAI Playground, inclusive of GPT-3.5 but not GPT-4 at the time of assessment. Our focus is primarily on LLMs with the ability to generate text and code. The broader term "generative AI" extends this sphere to include other formats such as images or audio. "LLM-powered software" refers to applications built upon LLMs or those that synergize LLMs with other generative AI models.

## 3.1 Overall Exposure

### 3.1.1 Overall Exposure Rubric Summary

**Summary of exposure rubric**

No exposure (E0) if:

- using the described LLM results in no or minimal reduction in the time required to complete the activity or task while maintaining equivalent quality[a] or
- using the described LLM results in a decrease in the quality of the activity/task output.

Direct exposure (E1) if:

- using the described LLM via ChatGPT or the OpenAI playground can decrease the time required to complete the DWA or task by at least half (50%).

LLM+ Exposed (E2) if:

- access to the described LLM alone would not reduce the time required to complete the activity/task by at least half, but
- additional software could be developed on top of the LLM that could reduce the time it takes to complete the specific activity/task with quality by at least half. Among these systems, we count access to image generation systems.[b]

[a] Equivalent quality means that a third party, typically the recipient of the output, would not notice or care about LLM assistance.

[b] In practice, we categorize access to image capabilities separately (E3) to facilitate annotation, though we combine E2 and E3 for all analyses.

### 3.1.2 Full Exposure Rubric

E Exposure Taxonomy

Consider the most powerful OpenAI large language model (LLM) This model can complete many tasks that can be formulated as having text input and text output where the context for the input can be captured in 2000 words. The model also cannot draw up-to-date facts (those from <1 year ago) unless they are captured in the input.

Assume you are a worker with an average level of expertise in your role trying to complete the given task. You have access to the LLM as well as any other existing software or computer hardware tools mentioned in the task. You also have access to any commonly available technical tools accessible via a laptop (e.g. a microphone, speakers, etc.). You do not have access to any other physical tools or materials.

Please label the given task according to the taxonomy below.

## E0 – No exposure

Label tasks E0 if direct access to the LLM through an interface like ChatGPT or the OpenAI playground cannot reduce the time it takes to complete this task with equivalent quality by half or more.

If a task requires a high degree of human interaction (for example, in person demonstrations) then it should be classified as E0.

## E1 – Direct exposure

Label tasks E1 if direct access to the LLM through an interface like ChatGPT or the OpenAI playground alone can reduce the time it takes to complete the task with equivalent quality by at least half. This includes tasks that can be reduced to: - Writing and transforming text and code according to complex instructions, - Providing edits to existing text or code following specifications, - Writing code that can help perform a task that used to be done by hand, - Translating text between languages, - Summarizing medium-length documents, - Providing

feedback on documents, - Answering questions about a document, or - Generating questions a user might want to ask about a document.

## E2 – Exposure by LLM-powered applications

Label tasks E2 if having access to the LLM alone may not reduce the time it takes to complete the task by at least half, but it is easy to imagine additional software that could be developed on top of the LLM that would reduce the time it takes to complete the task by half. This software may include capabilities such as: - Summarizing documents longer than 2000 words and answering questions about those documents - Retrieving up-to-date facts from the Internet and using those facts in combination with the LLM capabilities - Searching over an organization's existing knowledge, data, or documents and retreiving information

Examples of software built on top of the LLM that may help complete worker activities include: - Software built for a home goods company that quickly processes and summarizes their up-to-date internal data in customized ways to inform product or marketing decisions - Software that is able to suggest live responses for customer service agents speaking to customers in their company's customer service interface - Software built for legal purposes that can quickly aggregate and summarize all previous cases in a particular legal area and write legal research memos tailored to the law firm's needs - Software specifically designed for teachers that allows them to input a grading rubric and upload the text files of all student essays and have the software output a letter grade for each essay - Software that retrieves up-to-date facts from the internet and uses the capabilities of the LLM to output news summaries in different languages

## E3 – Exposure given image capabilities

Suppose you had access to both the LLM and a system that could view, caption, and create images. This system cannot take video media as inputs. This system cannot accurately retrieve very detailed information from image inputs, such as measurements of dimensions within an image. Label tasks as E3 if there is a significant reduction in the time it takes to complete the

task given access to a LLM and these image capabilities: - Reading text from PDFs, - Scanning images, or - Creating or editing digital images according to instructions.

## Annotation examples:

Occupation: Inspectors, Testers, Sorters, Samplers, and Weighers Task: Adjust, clean, or repair products or processing equipment to correct defects found during inspections. Label (E0/E1/E2/E3): E0 Explanation: The model does not have access to any kind of physicality, and more than half of the task (adjusting, cleaning and repairing equipment) described requires hands or other embodiment.

Occupation: Computer and Information Research Scientists Task: Apply theoretical expertise and innovation to create or apply new technology, such as adapting principles for applying computers to new uses. Label (E0/E1/E2/E3): E1 Explanation: The model can learn theoretical expertise during training as part of its general knowledge base, and the principles to adapt can be captured in the text input to the model.

Activity: Schedule dining reservations. Label (E0/E1/E2/E3): E2 Explanation: Automation technology already exists for this (e.g. Resy) and it's unclear what an LLM offers on top of using that technology (no-diff). That said, you could build something that allows you to ask the LLM to make a reservation on Resy for you. (E3)

Activity: Negotiate purchases or contracts. Label (E0/E1/E2/E3): E2 Explanation: You could have each party transcribe their point of view and then feed this to an LLM to resolve any disputes (E3). That said, many people would need to buy into using new technological tools to accomplish this (system).

Occupation: Allergists and Immunologists Task: Prescribe medication such as antihistamines, antibiotics, and nasal, oral, topical, or inhaled glucocorticosteroids. Label (E0/E1/E2/E3): E2 Explanation: The model can provide guesses for different diagnoses and write prescriptions and case notes. However, it still requires a human in the loop using their judgment and knowledge to

make the final decision.

—

## 3.2 Automation Exposure

We implemented an automation-focused rubric as well, though we only retrieved labels from GPT-4 in this case.

### 3.2.1 Automation Rubric Prompt

# T Automation Rubric

1. Determine if the occupation/task pair meets the definition of T0 No-Automation Exposure. If it does, label it as T0 and stop. 2. If the occupation/task pair does not meet the definition of T0 No-Automation Exposure, determine if the occupation/task pair meets one of the other definitions above and select the label that applies:

- T4: Full automation exposure - T3: High automation exposure - T2: Moderate automation exposure - T1: No automation exposure

## Rubric

LLM++ is a software tool that can perform a wide range of digital tasks but cannot program physical machines or objects.

Consider the most powerful OpenAI large language model (LLM). This model can complete many tasks that can be formulated as having text, image, or audio input and text, image, or audio output. The model also cannot draw up-to-date facts (those from <1 year ago) unless they are captured in the input. This system cannot accurately retrieve very detailed information from image inputs, such as measurements of dimensions within an image.

On its own, the LLM can perform tasks that can be reduced to: - Writing and transforming text and code according to complex instructions, - Providing edits to existing text or code following

specifications, - Writing code that can help perform simple, repetitive digital tasks, - Debugging code or software - Programming in computer languages like Python, C++, Java, Stata - Assisting with a data analysis - Translating text between languages, - Summarizing short documents, - Providing feedback on documents, - Answering questions about a document, - Generating questions a user might want to ask about a document, - Writing questions for an interview or assessment, - Writing and responding to emails, including ones that involve refuting information or engaging in a negotiation (but only if the negotiation is via written correspondence), - Maintain records of written data, - Prepare training materials based on general knowledge, - Searching for a relevant document over a small number of documents, - Reading text from PDFs, - Scanning images, - Creating or editing digital images according to instructions, - Captioning an image, - Transcribing text, or - Informing anyone of any information via any written or spoken medium.

Assume you are a worker with an average level of expertise in your role trying to complete the given task. You have access to the LLM as well as any other existing software or computer hardware tools mentioned in the task. You also have access to any commonly available technical tools accessible via a laptop (e.g. a microphone, speakers, etc.). You do not have access to any other physical tools or materials.

You also have access to LLM++ which is additional software developed on top of the LLM. LLM++ includes capabilities such as: - Summarizing documents longer than 2000 words and answering questions about those documents, - Retrieving up-to-date facts from the Internet and using those facts in combination with the LLM capabilities, - Searching over an organization's existing knowledge, data, or documents and retrieving information, - Retrieving highly specialized domain knowledge, - Making recommendations given data or written input, - Analyzing written information to inform decisions, - Preparing training materials based on highly specialized knowledge, - Performing legal research, - Providing counsel on issues, - Writing lesson plans, and - Maintaining databases.

We refer to this suite of software tools as LLM++. LLM++ can NOT program physical machines or objects.

Please label the given task according to the rubric below.

**T0 No-Automation Exposure** A class of tasks for which LLM++ can not conceivably perform any aspect of the task in any manner.

**T4 Full Automation Exposure** A class of tasks where, in most contexts in which this task is currently performed by a human, LLM++ can complete all aspects of this task with high quality when prompted by a human. The output does not normally require oversight by a human. Oversight is not normally required for tasks labeled T4 because the consequences for failure or inaccuracy are small for this task, human judgment is not necessary to complete this task, and generative models can consistently perform this task with very high quality.

**T3 High Automation Exposure** A class of tasks where, in most contexts in which this task is currently performed by a human, LLM++ could complete 90-100% of the components of the task when prompted, but the output requires oversight from a human. Oversight is normally required because the consequences for failure or inaccuracy are significant for this task, human judgment is necessary to complete this task, and/or generative models cannot perform all aspects of this task with high quality consistently. These tasks rely almost exclusively on the processing of digital information, but human judgment is needed to ensure that any digital outputs from LLM++ are high enough quality to be acceptable for the particular context.

** T2 Moderate Automation Exposure** A class of tasks where, in most contexts in which this task is performed by a human, LLM++ could complete between 50%-90% of the components of the task at high quality. These tasks normally rely heavily on the processing of digital information, but a significant portion of the task also involves actions that LLM++ cannot perform with high quality. These tasks require at least some human action beyond just double-checking generative model outputs (such as interpretation, judgment, human-to-human communication, or physical

actions).

** T1 Low Automation Exposure** A class of tasks where, in most contexts in which this task is performed by a human, LLM++ could complete between 0%-50% of the components of the task at high quality. These tasks normally rely only partially on the processing of digital information, while the majority of the task involves actions that LLM++ cannot perform with high quality. A majority of the actions that need to be taken to complete this task require a human to perform the action.

## Definitions

**High quality** means someone receiving or reviewing the output would not be able to tell the difference between whether it came from LLM++ or a human. For tasks that require a lot of interaction during the completion of the task (e.g. meetings, negotiations), high quality means the people you were interacting with either would not know or would not care that they were interacting with with LLM++.

**Digital information** or **information that can easily be expressed digitally** includes but is not limited to text, audio, images, PDFs, books, and code.

## Annotation Examples

Occupation: Special Education Teachers, Preschool Task: Develop individual educational plans (IEPs) designed to promote students' educational, physical, or social development. Automation score (T0/T1/T2/T3/T4): T2 Explanation: While LLM++ can assist in the creation of educational plans, this task requires significant interpretation and judgment by a human in order to be effective. Special Education Teachers must consider the individual needs and abilities of each student, as well as communicate with parents and other educators to ensure the plan is appropriate and effective. While LLM++ can provide suggestions and potential plans, it cannot make the final decision or communicate with stakeholders.

Occupation: Construction and Building Inspectors Task: Inspect and monitor construction

sites to ensure adherence to safety standards, building codes, or specifications. Automation score (T0/T1/T2/T3/T4): T1 Explanation: While LLM++ can provide some support in the inspection and monitoring of construction sites, this task involves a significant amount of physical inspection and communication with workers and supervisors on-site. Additionally, the inspector may need to make on-the-spot judgments based on what they observe and may need to refer to local building codes and regulations which may not be available in digital format. As such, this task requires a significant amount of human input and oversight.

Occupation: Food and Tobacco Roasting, Baking, and Drying Machine Operators and Tenders Task: Dump sugar dust from collectors into melting tanks and add water to reclaim sugar lost during processing. Automation score (T0/T1/T2/T3/T4): T0 Explanation: This task requires physical actions that cannot be completed by LLM++.

Occupation: Historians Task: Translate or request translation of reference materials. Automation score (T0/T1/T2/T3/T4): T4 Explanation: LLM++ can easily complete this task by translating text from one language to another. Human intervention is not usually necessary.

Occupation: Electronics Engineer Task: Prepare documentation containing information such as confidential descriptions or specifications of proprietary hardware or software, product development or introduction schedules, product costs, or information about product performance weaknesses. Automation score (T0/T1/T2/T3/T4): T2 Explanation: LLM++ can assist with documentation but human judgment is needed for understanding specifications, costs, and performance weaknesses. Thus, this task falls under T2 Moderate Automation Exposure.

Occupation: Advanced Practice Psychiatric Nurse Task: Write prescriptions for psychotropic medications as allowed by state regulations and collaborative practice agreements. Automation score (T0/T1/T2/T3/T4): T1 Explanation: LLM++ could be used to help create prescriptions given digital input about the patient's context, but a human would need to double-check the information to ensure that it meets regulations and is appropriate for the patient.

## Now, apply the above rubric to the example below:

### 3.2.2 Automation Rubric Results

Our automation exposure measures are a proof-of-concept that GPT-4 may be able to evaluate its own potential to automate specific tasks, but we have not validated the GPT-4 evaluations against the judgement of human raters. Of course this rubric has all of the same weaknesses as the exposure rubric, but additionally was not validated with human evaluation. The results we find, however, suggest some intuitive findings and broadly have high correlation with the exposure measures. Still, these are still early results. To convert our automation labels from the rubric into numerical scores, we consider T0 (no exposure to automation) as a score of 0, T4 as full exposure to have a score of 1, then T1, T2, and T3 to have scores of 0.25, 0.5, and 0.75 respectively. This maps our T rubric scores into numerical scores between 0 and 1. Preliminary results suggest exposure as defined in the E rubric scores from GPT-4 and automation as defined in the T rubric have a strong positive association. Table 1 contains summary statistics on our GPT-4 evaluations of the automation rubric. Table 3 shows the regression results from a regression of GPT-4 $E1 + 0.5 * E2$ exposure scores on GPT-4 automation ratings. Figure 5 shows the scatterplot of the same exposure scores against the automation scores we generate at the task level, and Figure 6 shows the binscatter of automation score exposure against log wage rated at the occupation-level. We see similar trends in these early GPT-4 measures of automation as we do in overall exposure.

Overall exposure and automation exposure are highly correlated, with an $R^2$ of 55.6 percent in a univariate regression of exposure scores on automation scores at the task-level. This means that where automation is high, it is likely the case that exposure, which includes augmentation potential, is also relatively high. Automation and augmentation potential will often co-occur, as in (*52*).

Automation exposure ratings follow a similar trend to general exposure ratings (see Figure 6).

Higher wage occupations tend to be more likely to have automatable tasks. No occupation has 100 percent of tasks exposed to automation by GPT under this rubric, though many occupations have high levels of exposure that suggest they might be reconfigured entirely. Occupations with the highest automation exposure include telephone operators, travel agents, word processors and typists, credit authorizers, checkers, and clerks, desktop publishers, and telemarketers. Overall, only 1.86 percent of tasks are assessed to be exposed to "full automation" suggesting that there may be significant last-mile barriers to automating many tasks.

In light of this, emphasizing augmentation instead of automation may not be useful in charting a path forward for designing jobs. Even in cases where it is possible to prioritize augmentation, market-level factors like price and income elasticities of supply and demand (among others) dictate the extent to which individual workers are aided or displaced by new configurations of work (*77*). Whether workers are displaced or complemented on balance will depend on the particulars of the application and emergent use cases as new systems and complements emerge. Based on this evidence, it is also premature to suggest that high broad-based exposure implies the wholesale elimination of jobs for similar reasons.

Table 1: Distribution of GPT-4 Automation Annotations for LLM++ Across Tasks

| **Automation** | **Task Count** | **Task Share** |
|---|---|---|
| No automation exposure | 5,528 | 28.69% |
| Low automation exposure | 2,937 | 15.25% |
| Moderate automation exposure | 7,796 | 40.47% |
| High automation exposure | 2,646 | 13.73% |
| Full automation exposure | 358 | 1.86% |
| Total | 19,265 | 100% |

Table 2: Column 1 lists the five possible GPT-4-generated annotations on the automation taxonomy for LLM++. It is important to remember that these annotations are generated based on the definition given of LLM++ in the taxonomy, which includes not only LLM capabilities, but the potential for software built on top of an LLM to automate the task. The number of tasks assigned to each label is presented in column 2, and the share of all tasks assigned the label is presented in column 3.

| | GPT-4 Exposure Rating 1 | GPT-4 Exposure Rating 2 |
|---|---|---|
| | (1) | (2) |
| GPT-4 Automation Score | $0.936^{***}$ | $0.847^{***}$ |
| | (0.00609) | (0.00737) |
| Constant | $0.045^{***}$ | $-0.005^{**}$ |
| | (0.00201) | (0.00209) |
| N | 19265 | 19265 |
| $R^2$ | 0.556 | 0.457 |

Table 3: Regression of GPT-4 Exposure Rubric Scores on GPT-4 Automation Rubric Scores

Table Notes: The table shows the regression of our GPT-4 Exposure Rubric scores following the $E1 + 0.5 * E2$ convention described above on the GPT-4 Automation Score as defined by the Automation ("T") rubric. The unit of observation is the task. T0, T1, T2, T3, and T4 are mapped to scores of 0, 0.25, 0.5, 0.75, and 1 respectively. In both cases, the automation score explains about half of the variance of the exposure scores.

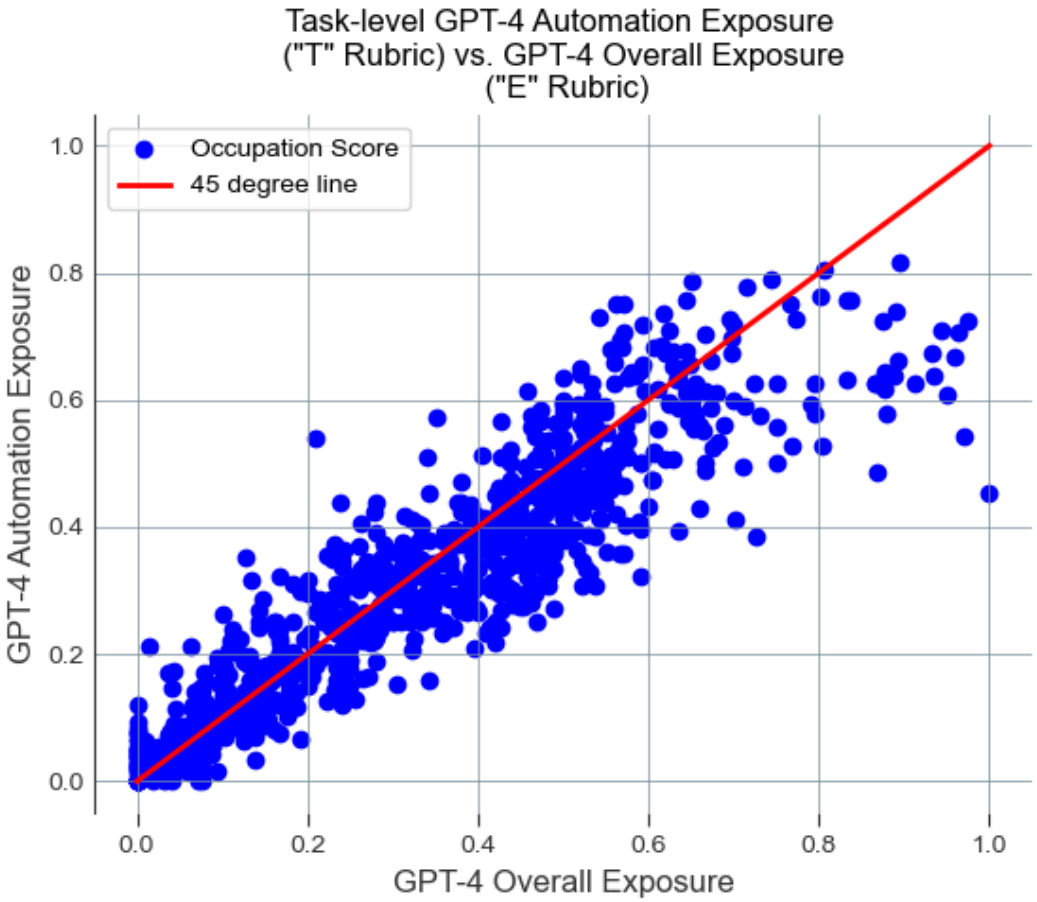



Figure 5: GPT-4 Ratings of Automation Exposure ("T") vs. Overall Exposure ("E")

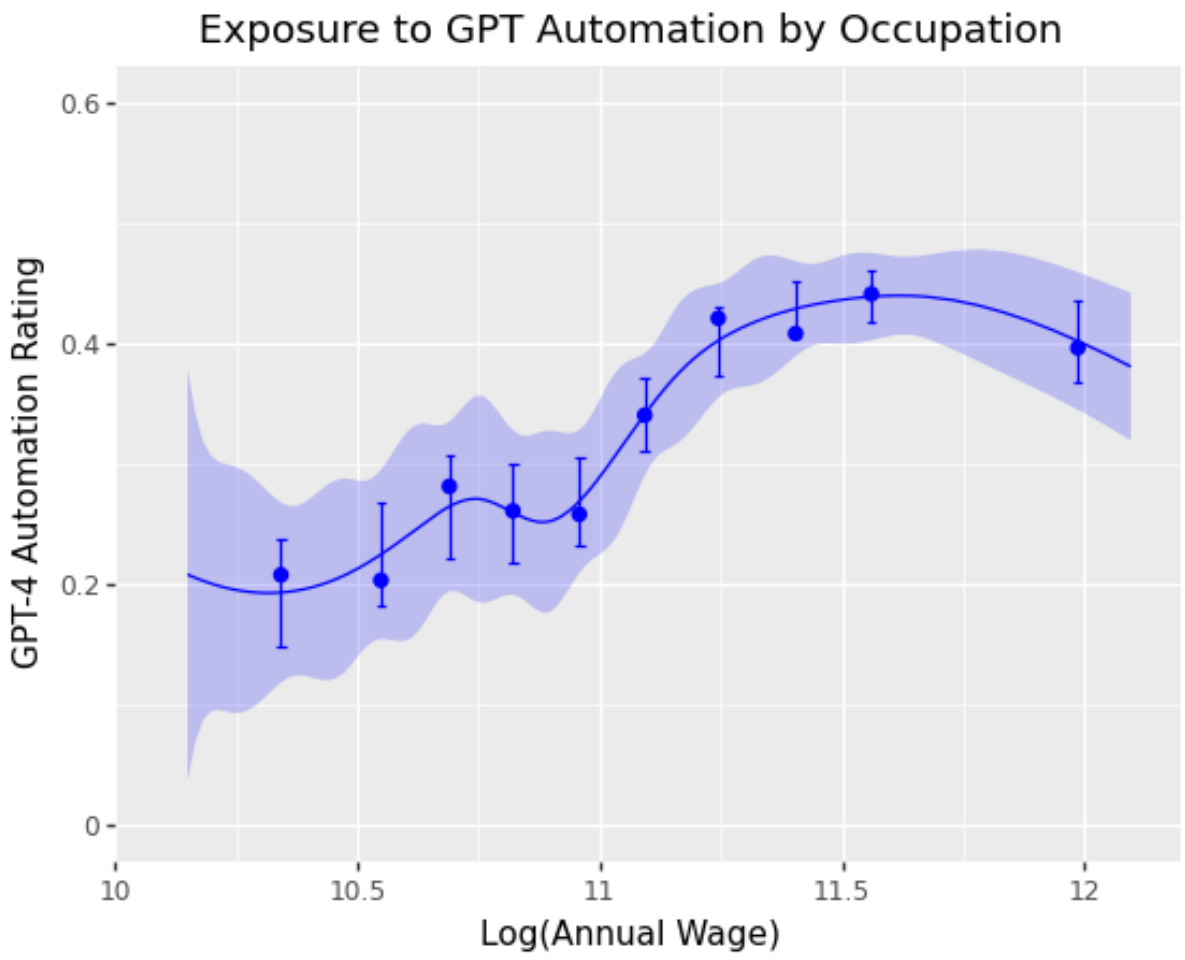



Figure 6: Binscatter of GPT-4 Automation Exposure Ratings by Occupation vs. Log Wages

To measure exposure, we keep the 50% constant and vary the tools the subject has access to. To measure automation and augmentation we keep the tools constant and vary the amount of time access to the tools saves a human.

# 4 Data Collection

## 4.1 Data on Activities and Tasks Performed by Occupation in the U.S. (also in main text)

We use the O*NET 27.2 database (*11*), which contains information on 1,016 occupations, including their respective Detailed Work Activities (DWAs) and tasks. A DWA is a comprehensive action that is part of completing task, such as "Study scripts to determine project requirements." A task, on the other hand, is an occupation-specific unit of work that may be associated with zero, one, or multiple DWAs. We offer a sample of tasks and DWAs in Table 4. The two datasets we use consist of:

- 19,265 tasks, consisting of a "task description" and a corresponding occupation, and
- 2,087 DWAs, where most DWAs are connected to one or more tasks, and tasks may be associated with one or more DWAs, though some tasks lack any associated DWAs.

| Task ID | Occupation Title | DWAs | Task Description |
|---|---|---|---|
| 14675 | Computer Systems Engineers/Architects | Monitor computer system performance to ensure proper operation. | Monitor system operation to detect potential problems. |
| 18310 | Acute Care Nurses | Operate diagnostic or therapeutic medical instruments or equipment. Prepare medical supplies or equipment for use. | Set up, operate, or monitor invasive equipment and devices, such as colostomy or tracheotomy equipment, mechanical ventilators, catheters, gastrointestinal tubes, and central lines. |
| 4668.0 | Gambling Cage Workers | Execute sales or other financial transactions. | Cash checks and process credit card advances for patrons. |
| 15709 | Online Merchants | Execute sales or other financial transactions. | Deliver e-mail confirmation of completed transactions and shipment. |
| 6529 | Kindergarten Teachers, Except Special Education | – | Involve parent volunteers and older students in children's activities to facilitate involvement in focused, complex play. |
| 6568 | Elementary School Teachers, Except Special Education | – | Involve parent volunteers and older students in children's activities to facilitate involvement in focused, complex play. |

Table 4: Sample of occupations, tasks, and Detailed Work Activities from the O*NET database. We see that aggregating over activities alone is imprecise, as evidenced by the fact that we'd expect Gambling Cage Workers to complete the given DWA in person, using some physicality while we'd expect Online Merchants to complete the same activity solely with a computer.

## 4.2 Data on Wages, Employment, and Demographics

We obtain employment and wage data from the 2020 and 2021 Occupational Employment series provided by the Bureau of Labor Statistics. This dataset encompasses occupational titles, the number of workers in each occupation, and occupation-level employment projections for 2031, typical education required for entry in an occupation and on-the-job training required to attain competency in an occupation (*12*). We use the BLS-recommended crosswalk to O*NET (*13*) to link the O*NET task and DWA dataset and the BLS Labor Force Demographics (*14*), which is derived from the Current Population Survey (CPS). Both of these data sources are collected by the U.S. government and primarily capture workers who are not self-employed, are documented, and are working in the so-called formal economy.

# 5 Skills

## 5.1 Skill Importance

In this section, we explore the relationship between skill importance within an occupation, as annotated in the O*NET dataset, and our exposure measures. Using O*NET's Basic Skills (skill definitions can be found in 5.2) and normalizing skill importance for each occupation, we enhance the interpretability of the results. We then perform regression analysis on our exposure measures ($E1$, $E1 + 0.5 * E2$, $E1 + E2$) to examine the associations between skill importance and exposure.

Our findings indicate a strong negative correlation between the importance of **science** and **critical thinking** skills and exposure, implying less likelihood of current LLMs impacting occupations requiring these skills. Conversely, **programming** and **writing** skills exhibit a strong positive association with exposure, indicating a higher exposure risk of occupations involving these skills to LLM influence (see Table 5 for detailed results).

| Basic Skill | $E1$<br>(std err) | $E1 + 0.5 * E2$<br>(std err) | $E1 + E2$<br>(std err) |
|---|---|---|---|
| | *All skill importance scores are normalized to be between 0 and 1.* | | |
| **Constant** | 0.082***<br>(0.011) | -0.112***<br>(0.011) | 0.300***<br>(0.057) |
| **Active Listening** | 0.128**<br>(0.047) | 0.214***<br>(0.043) | 0.449***<br>(0.027) |
| **Mathematics** | -0.127***<br>(0.026) | 0.161***<br>(0.021) | 0.787***<br>(0.049) |
| **Reading Comprehension** | 0.153***<br>(0.041) | 0.470***<br>(0.037) | -0.346***<br>(0.017) |
| **Science** | -0.114***<br>(0.014) | -0.230***<br>(0.012) | -0.346***<br>(0.017) |
| **Speaking** | -0.028<br>(0.039) | 0.133***<br>(0.033) | 0.294***<br>(0.042) |
| **Writing** | 0.368***<br>(0.042) | 0.467***<br>(0.037) | 0.566***<br>(0.047) |
| **Active Learning** | -0.157***<br>(0.027) | -0.065**<br>(0.024) | 0.028<br>(0.032) |
| **Critical Thinking** | -0.264***<br>(0.036) | -0.196***<br>(0.033) | -0.129**<br>(0.042) |
| **Learning Strategies** | -0.072*<br>(0.028) | -0.209***<br>(0.025) | -0.346***<br>(0.034) |
| **Monitoring** | -0.067**<br>(0.023) | -0.149***<br>0.020) | -0.232***<br>(0.026) |
| **Programming** | 0.637***<br>(0.030) | 0.623***<br>(0.022) | 0.609***<br>(0.024) |

Table 5: Regression of occupation-level, human-annotated exposure to GPTs on skill importance for each skill in the O*NET Basic skills category, plus the programming skill.

## 5.2 O*NET Basic Skills Definitions

## Basic Skills

Developed capacities that facilitate learning or the more rapid acquisition of knowledge.

## Content

Background structures needed to work with and acquire more specific skills in a variety of different domains.

- **Reading Comprehension** — Understanding written sentences and paragraphs in work-

related documents.

- **Active Listening** — Giving full attention to what other people are saying, taking time to understand the points being made, asking questions as appropriate, and not interrupting at inappropriate times.
- **Writing** — Communicating effectively in writing as appropriate for the needs of the audience.
- **Speaking** — Talking to others to convey information effectively.
- **Mathematics** — Using mathematics to solve problems.
- **Science** — Using scientific rules and methods to solve problems.

## Process

Procedures that contribute to the more rapid acquisition of knowledge and skill across a variety of domains

- **Critical Thinking** — Using logic and reasoning to identify the strengths and weaknesses of alternative solutions, conclusions or approaches to problems.
- **Active Learning** — Understanding the implications of new information for both current and future problem-solving and decision-making.
- **Learning Strategies** — Selecting and using training/instructional methods and procedures appropriate for the situation when learning or teaching new things.
- **Monitoring** — Monitoring/Assessing performance of yourself, other individuals, or organizations to make improvements or take corrective action.

### Cross-Functional Skills

Note: We selected only Programming from the list of cross-functional skills because of our prior knowledge about the models' ability to code.

- **Programming** - Writing computer programs for various purposes.

# 6 Education

## 6.1 Barriers to Entry

Next, we examine barriers to entry to better understand if there is differentiation in exposure due to types of jobs. One such proxy is an O*NET occupation-level descriptor called the "Job Zone." A Job Zone groups occupations that are similar in (a) the level of education needed to get a job in the occupation, (b) the amount of related experience required to do the work, and (c) the extent of on-the-job training needed to do the work. In the O*NET database, there are 5 Job Zones, with Job Zone 1 requiring the least amount of preparation (3 months) and Job Zone 5 requiring the most extensive amount of preparation, 4 or more years. We observe that median income increases monotonically across Job Zones as the level of preparation needed also increases, with the median worker in Job Zone 1 earning $\$30,230$ and the median worker in Job Zone 5 earning $\$80,980$.

All of our measures ($E1$, $E1 + 0.5 * E2$, and $E1 + E2$) show an identical pattern, that is, exposure increases from Job Zone 1 to Job Zone 4, and either remains similar or decreases at Job Zone 5. In Figure 7, we plot the percentage of workers at every threshold of exposure. We find that, on average, the percentage of workers in occupations with greater than 50% $E1 + 0.5 * E2$ exposure in Job Zones 1 through 5 have $E1 + 0.5 * E2$ at 0.00% (Job Zone 1), 6.11% (Job Zone 2), 10.57% (Job Zone 3), 34.5% (Job Zone 4), and 26.45% (Job Zone 5), respectively.

| | Job Zone 1 | Job Zone 2 | Job Zone 3 | Job Zone 4 | Job Zone 5 |
|---|---|---|---|---|---|
| **Preparation** | None or little (0-3 months) | Some (3-12 months) | Medium (1-2 years) | Considerable (2-4 years) | Extensive (4+ years) |
| **Education** | High school diploma or GED (optional) | High school diploma | Vocational school, on-the-job training, or associate's degree | Bachelor's degree | Master's degree or higher |
| **Example Occupations** | Food prep workers, dishwashers, floor sanders | Orderlies, customer service reps, tellers | Electricians, barbers, medical assistants | Database admins, graphic designers, cost estimators | Pharmacists, lawyers, astronomers |
| **Median Income** | $30,230 | $38,215 | $54,815 | $77,345 | $81,980 |
| **Tot Emp (000s)** | 13,100 | 73,962 | 37,881 | 56,833 | 21,221 |
| **H $E1$** | 0.03 | 0.07 | 0.11 | 0.23 | 0.23 |
| **M $E1$** | 0.04 | 0.12 | 0.14 | 0.18 | 0.13 |
| **H $E1 + 0.5 * E2$** | 0.06 | 0.16 | 0.26 | 0.47 | 0.43 |
| **M $E1 + 0.5 * E2$** | 0.06 | 0.20 | 0.32 | 0.51 | 0.45 |
| **H $E1 + E2$** | 0.09 | 0.24 | 0.41 | 0.71 | 0.63 |
| **M $E1 + E2$** | 0.08 | 0.27 | 0.51 | 0.85 | 0.76 |

Table 6: Mean exposure to GPTs by job zone. For each job zone, we also present the median of median annual income for each constituting occupation in USD, and the total number of workers in all occupations for that job zone, in the thousands.

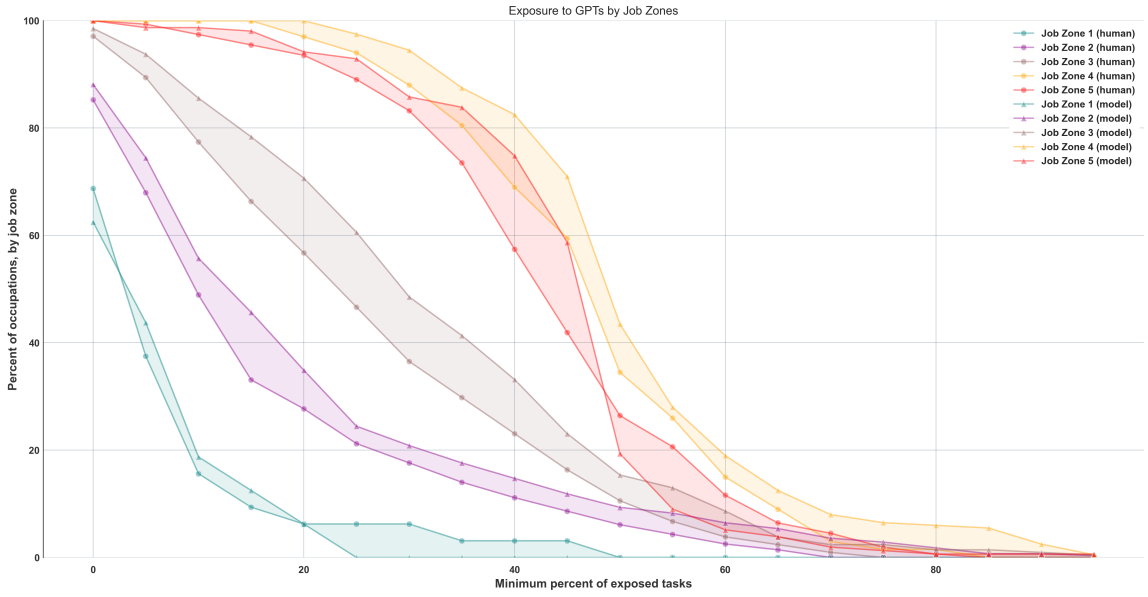



Figure 7: $E1 + 0.5 * E2$ exposure ratings of occupations in the five Job Zones, which are groups of similar occupations that are classified according to the level of education, experience, and on-the-job training needed to perform them.

### 6.1.1 Typical Education Needed for Entry

Since inclusion in a Job Zone accounts for both the education required—which itself is a proxy for skill acquisition—and the preparation required, we seek data to disentangle these variables. We use two variables from the Bureau of Labor Statistics' Occupational data: "Typical Education Needed for Entry" and "On-the-job Training Required to Attain Competency" in an occupation. By examining these factors, we aim to uncover trends with potential implications for the workforce. There are 3,504,000 workers for whom we lack data on education and on-the-job training requirements, and they are therefore excluded from the summary tables.

Our analysis suggests that individuals holding Bachelor's, Master's, and professional degrees are more exposed to LLMs and LLM-powered software than those without formal educational credentials (see Table 8). Interestingly, we also find that individuals with some college education but no degree exhibit a high level of exposure to LLMs and LLM-powered software. Upon examining the table displaying barriers to entry, we observe that the jobs with the least exposure require the most training, potentially offering a lower payoff (in terms of median income) once competency is achieved. Conversely, jobs with no on-the-job training required or only internship/residency required appear to yield higher income but are more exposed to LLMs.

| | Median Income | Emp (000s) | H $E1$ | M $E1$ | H $E1 + 0.5 * E2$ | M $E1 + 0.5 * E2$ | H $E1 + E2$ | M $E1 + E2$ |
|---|---|---|---|---|---|---|---|---|
| No formal educational credential | $31,900 | 36,187 | 0.05 | 0.06 | 0.10 | 0.10 | 0.15 | 0.15 |
| High school diploma or equivalent | $45,470 | 67,033 | 0.09 | 0.13 | 0.20 | 0.25 | 0.31 | 0.37 |
| Postsecondary nondegree award | $48,315 | 9,636 | 0.07 | 0.15 | 0.19 | 0.28 | 0.31 | 0.41 |
| Some college, no degree | $40,970 | 2,898 | 0.23 | 0.34 | 0.39 | 0.53 | 0.55 | 0.72 |
| Associate's degree | $60,360 | 3,537 | 0.12 | 0.14 | 0.31 | 0.36 | 0.49 | 0.59 |
| Bachelor's degree | $78,375 | 71,698 | 0.23 | 0.17 | 0.47 | 0.51 | 0.70 | 0.84 |
| Master's degree | $79,605 | 3,216 | 0.26 | 0.14 | 0.46 | 0.44 | 0.66 | 0.74 |
| Doctoral or professional degree | $82,420 | 5,290 | 0.21 | 0.13 | 0.41 | 0.43 | 0.60 | 0.74 |

Table 7: Mean exposure scores for occupations, grouped by typical education needed for entry into the occupation. Alongside exposure scores, we display the median of median annual income for each occupation, as well as the total number of workers in each group, in thousands.

| On The Job Training Required | None | Apprenticeship | Internship or Residency | Short-term | Moderate-term | Long-term |
|---|---|---|---|---|---|---|
| **Median Income** | $77,440 | $55,995 | $77,110 | $33,370 | $46,880 | $48,925 |
| **Tot Emp (000s)** | 90,776 | 3,066 | 3,063 | 66,234 | 31,285 | 5,070 |
| **H $E1$** | 0.20 | 0.01 | 0.16 | 0.11 | 0.09 | 0.08 |
| **M $E1$** | 0.16 | 0.02 | 0.06 | 0.15 | 0.12 | 0.10 |
| **H $E1 + 0.5 * E2$** | 0.42 | 0.04 | 0.36 | 0.21 | 0.21 | 0.18 |
| **M $E1 + 0.5 * E2$** | 0.46 | 0.06 | 0.38 | 0.25 | 0.25 | 0.22 |
| **H $E1 + E2$** | 0.63 | 0.07 | 0.55 | 0.32 | 0.32 | 0.28 |
| **M $E1 + E2$** | 0.76 | 0.10 | 0.71 | 0.34 | 0.38 | 0.33 |

Table 8: Mean exposure scores for occupations, grouped by level of on-the-job training required to attain competency in the job. Alongside exposure scores, we display the median of median annual income for each occupation, as well as the total number of workers in each group, in thousands.

# 7 Industrial and Productivity Exposure

Figures 8 and 9 show the overall employment-weighted relative exposure of 3-digit NAICS industries according to human raters and GPT-4 respectively (based on our exposure rubric). The impact potential is present across nearly all industries, with wide heterogeneity. Both methods agree generally on relative exposures: data processing, information processing, and hospitals all have high exposure.

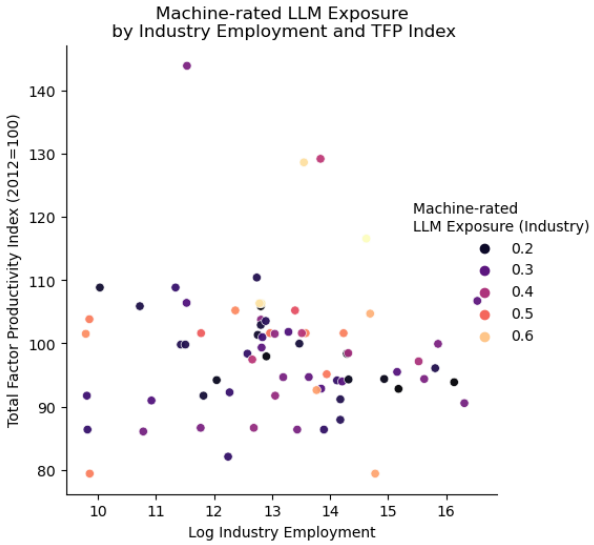



Figure 10: Machine-rated LLM Exposure Industry Total Factor Productivity Scatter

Figure Notes: Total Factor Productivity level as of 2020 is on the y-axis and logged industry employment is on the x-axis. Points are industries (3-digit NAICS) colored by their GPT4-rated exposure, weighing occupational exposures by total employed in each 6-digit SOC Code. There is little correlation between change in productivity level since 2012 and exposure.

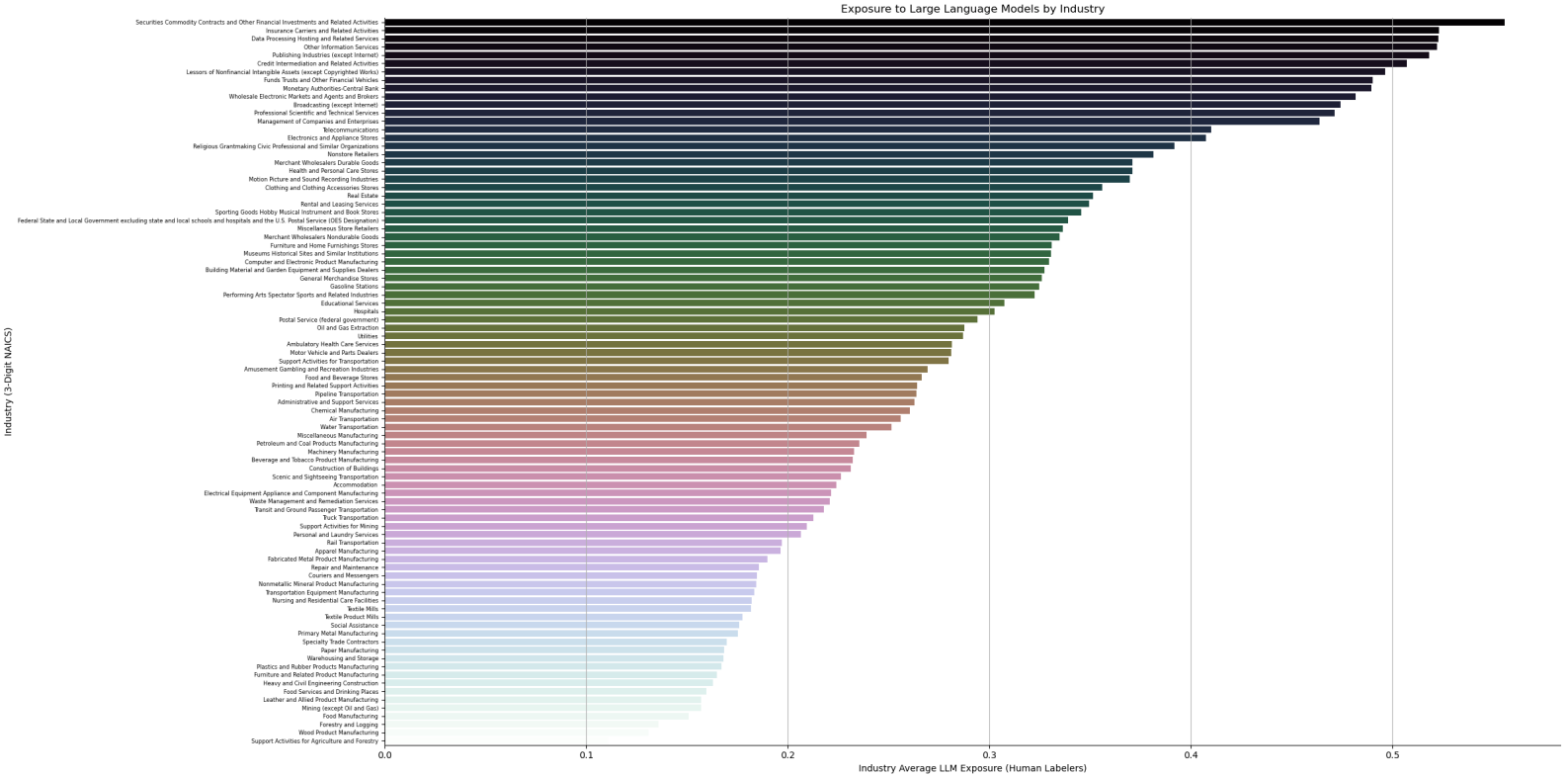



Figure 8: Human-labeled exposure aggregated up to the 3-digit NAICS industry. Exposure-levels are derived from the $E1 + 0.5 * E2$ scoreset at the task-level, then aggregated to occupations. Industries measures are built from employment-weights at the SOC6-level within 3-digit NAICS.

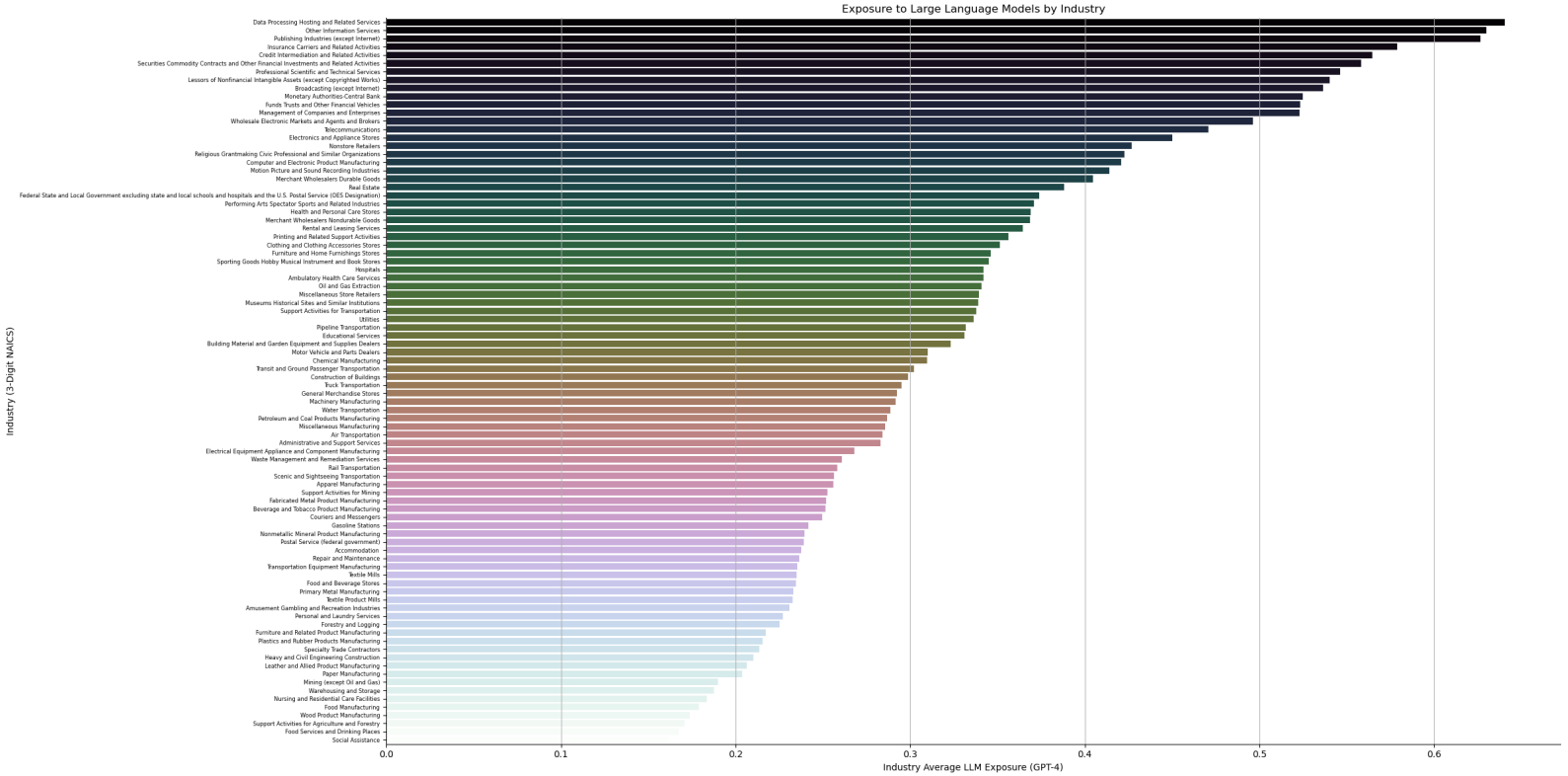



Figure 9: GPT4-labeled exposure aggregated up to the 3-digit NAICS industry. Exposure-levels are derived from the $E1 + 0.5 * E2$ scoreset at the task-level, then aggregated to occupations. Industries measures are built from employment-weights at the SOC6-level within 3-digit NAICS.

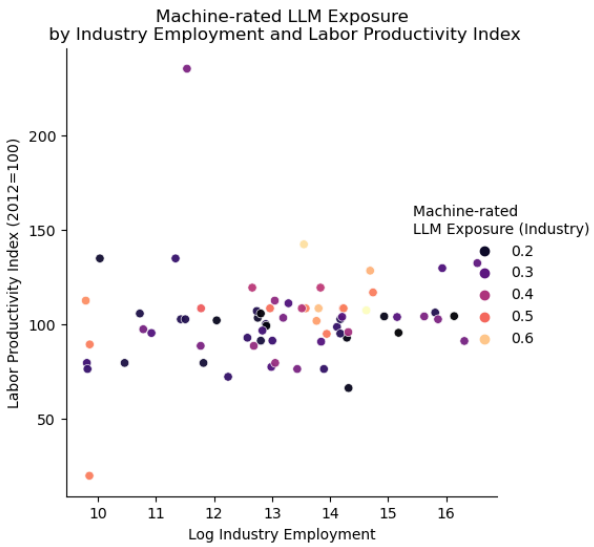



Figure 11: Machine-rated LLM Exposure Industry Labor Productivity Scatter

Labor Productivity level as of 2020 is on the y-axis and logged industry employment is on the x-axis. Points are industries (3-digit NAICS) colored by their GPT4-rated exposure, weighing occupational exposures by total employed in each 6-digit SOC Code. There is little correlation between change in productivity level since 2012 and exposure.

Recent productivity growth (both total factor and labor) appears uncorrelated with exposure as well. Figures 10 and 11 show little relationship between productivity growth as of 2020 since 2012 and current exposure to LLMs as rated by the model. A high correlation between already fast-growing productive industries and exposure might mean an exacerbation of Baumol's cost disease. In other words, if LLMs are likely to increase productivity differentially across industries, one concern is that the most productive would become even more productive. With inelastic demand for the production of those industries, the most productive sectors would shrink as a proportion of inputs in the economy. We see little to suggest this will be the case. Productivity growth since 2012 and exposure to LLM technologies appear unrelated. Productivity data are available from https://www.bea.gov/data/special-topics/integrated-industry-level-production-account-klems.

# 8 Exposed Occupations

## 8.1 Occupations Without Any Exposed Tasks

| Occupations with no labeled exposed tasks |
|---|
| Agricultural Equipment Operators |
| Athletes and Sports Competitors |
| Automotive Glass Installers and Repairers |
| Bus and Truck Mechanics and Diesel Engine Specialists |
| Cement Masons and Concrete Finishers |
| Cooks, Short Order |
| Cutters and Trimmers, Hand |
| Derrick Operators, Oil and Gas |
| Dining Room and Cafeteria Attendants and Bartender Helpers |
| Dishwashers |
| Dredge Operators |
| Electrical Power-Line Installers and Repairers |
| Excavating and Loading Machine and Dragline Operators, Surface Mining |
| Floor Layers, Except Carpet, Wood, and Hard Tiles |
| Foundry Mold and Coremakers |
| Helpers–Brickmasons, Blockmasons, Stonemasons, and Tile and Marble Setters |
| Helpers–Carpenters |
| Helpers–Painters, Paperhangers, Plasterers, and Stucco Masons |
| Helpers–Pipelayers, Plumbers, Pipefitters, and Steamfitters |
| Helpers–Roofers |
| Meat, Poultry, and Fish Cutters and Trimmers |
| Motorcycle Mechanics |
| Paving, Surfacing, and Tamping Equipment Operators |
| Pile Driver Operators |
| Pourers and Casters, Metal |
| Rail-Track Laying and Maintenance Equipment Operators |
| Refractory Materials Repairers, Except Brickmasons |
| Roof Bolters, Mining |
| Roustabouts, Oil and Gas |
| Slaughterers and Meat Packers |
| Stonemasons |
| Tapers |
| Tire Repairers and Changers |
| Wellhead Pumpers |

Table 9: All 34 occupations for which none of our measures labeled any tasks as exposed.

## 8.2 Occupations With Greatest Exposure

| Group | Occupations with highest exposure | % Exposure |
|---|---|---|
| **Human $E1$** | Interpreters and Translators | 76.5 |
| | Survey Researchers | 75.0 |
| | Poets, Lyricists and Creative Writers | 68.8 |
| | Animal Scientists | 66.7 |
| | Public Relations Specialists | 66.7 |
| **Human $E1 + 0.5 * E2$** | Survey Researchers | 84.4 |
| | Writers and Authors | 82.5 |
| | Interpreters and Translators | 82.4 |
| | Public Relations Specialists | 80.6 |
| | Animal Scientists | 77.8 |
| **Human $E1 + E2$** | Mathematicians | 100.0 |
| | Tax Preparers | 100.0 |
| | Financial Quantitative Analysts | 100.0 |
| | Writers and Authors | 100.0 |
| | Web and Digital Interface Designers | 100.0 |
| | *Humans labeled 15 occupations as "fully exposed."* | |
| **Model $E1$** | Mathematicians | 100.0 |
| | Correspondence Clerks | 95.2 |
| | Blockchain Engineers | 94.1 |
| | Court Reporters and Simultaneous Captioners | 92.9 |
| | Proofreaders and Copy Markers | 90.9 |
| **Model $E1 + 0.5 * E2$** | Mathematicians | 100.0 |
| | Blockchain Engineers | 97.1 |
| | Court Reporters and Simultaneous Captioners | 96.4 |
| | Proofreaders and Copy Markers | 95.5 |
| | Correspondence Clerks | 95.2 |
| **Model $E1 + E2$** | Accountants and Auditors | 100.0 |
| | News Analysts, Reporters, and Journalists | 100.0 |
| | Legal Secretaries and Administrative Assistants | 100.0 |
| | Clinical Data Managers | 100.0 |
| | Climate Change Policy Analysts | 100.0 |
| | *The model labeled 86 occupations as "fully exposed."* | |
| **Highest variance** | Search Marketing Strategists | 14.5 |
| | Graphic Designers | 13.4 |
| | Investment Fund Managers | 13.0 |
| | Financial Managers | 13.0 |
| | Insurance Appraisers, Auto Damage | 12.6 |

Table 10: Occupations with the highest exposure according to each measurement

The bottom section lists the occupations with the highest $\sigma^2$ value, indicating that they had the most variability in exposure scores. Exposure percentages indicate the share of an occupation's task that are exposed to GPTs ($E1$) or GPT-powered software ($E1 + 0.5 * E2$ and $E1 + E2$), where exposure is defined as driving a reduction in time it takes to complete the task by at least 50%. As such, occupations listed in this table are those where we estimate that GPTs and GPT-powered software are able to save workers a significant amount of time completing a large share of their tasks, but it does not necessarily suggest that their tasks can be fully automated by these technologies.

# 9 Validation of Measures

## 9.1 Limitations of our methodology

At this stage, some general-purpose technology criteria are easier to evaluate than others. Our primary focus at this early stage is to test the hypothesis that LLMs have a pervasive influence on the economy, similar to the approach taken by (*5*), who analyzed machine learning diffusion through job postings to assess its status as a general-purpose technology. Instead of using job postings or studying machine learning in general, we employ the task evaluation approach with both human and GPT-4 annotations. This analysis may reveal whether the impacts are limited to a specific set of similar tasks or occupations or if they will be more widespread.

### 9.1.1 Subjective human judgments

A fundamental limitation of our approach lies in the subjectivity of the labeling. In our study, we employ annotators who are familiar with LLM capabilities. However, this group is not occupationally diverse, potentially leading to biased judgments regarding LLMs' reliability and effectiveness in performing tasks within unfamiliar occupations. We acknowledge that obtaining high-quality labels for each task in an occupation requires workers engaged in those occupations or, at a minimum, possessing in-depth knowledge of the diverse tasks within those occupations. This represents an important area for future work in validating these results.

### 9.1.2 Measuring LLMs with GPT-4

Recent research indicates that GPT-4 serves as an effective discriminator, capable of applying intricate taxonomies and responding to changes in wording and emphasis (*16*). The outcomes of GPT-4 task classification are sensitive to alterations in the rubric's wording, the prompt's order and composition, the presence or absence of specific examples in the rubric, the level of detail provided, and the definitions given for key terms. Iterating on the prompt, based on observed

outcomes in a small validation set, can enhance the agreement between model outputs and the rubric's intent. Consequently, there are slight differences between the rubric presented to humans and the one used for GPT-4. This decision was made deliberately to guide the model towards reasonable labels without excessively influencing human annotators. As a result, we use multiple annotation sources, but none should be considered the definitive ground truth relative to the others. In this analysis, we present results from human annotators as our primary results. Further improvement and innovation in crafting effective rubrics for LLM classification remains possible. Still, we observe a high degree of agreement between human ratings and GPT-4 ratings at the occupation level concerning overall exposure to LLM systems (see Figure 12).

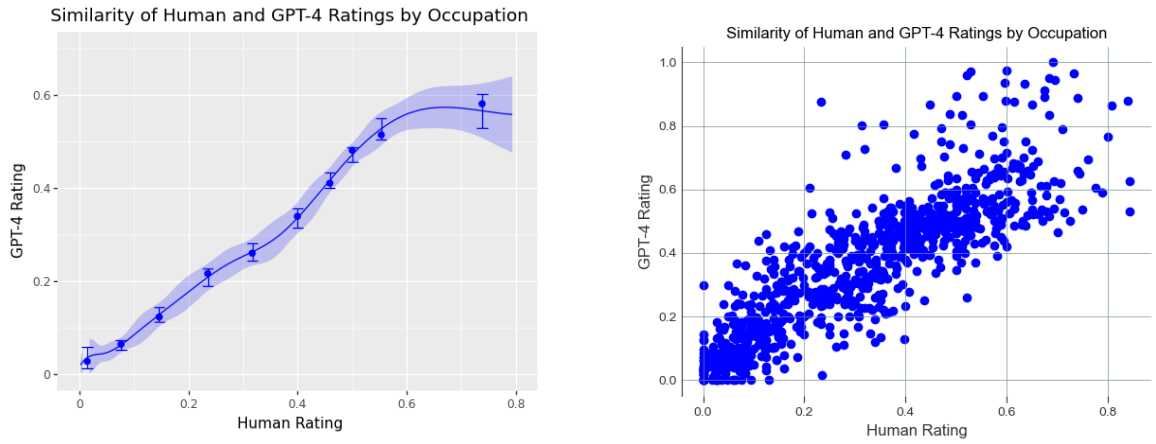



Figure 12: Human raters (x-axis) and GPT-4 ratings (y-axis) show a high degree of agreement about LLM exposure by occupation. Near the highest levels of exposure following the $E1+0.5*E2$ method of aggregating exposure scores to occupations, GPT-4 ratings tend to be lower than Human ratings. We present the raw scatter plot and the binscatter. Near the top end of exposure ratings, humans are on average more likely to rate an occupation as exposed.

### 9.1.3 Additional Weaknesses

- **Validity of task-based framework.** It is unclear to what extent occupations can be entirely broken down into tasks, and whether this approach systematically omits certain categories of skills or tasks that are tacitly required for competent performance of a job. Additionally,

tasks can be composed of sub-tasks, some of which are more automatable than others. Some tasks may function as pre-cursor to other tasks, such that the completion of downstream tasks is dependent on precursor tasks. If indeed, the task-based breakdown is not a valid representation of how most work in an occupation is performed, our occupation-level exposure analysis would be significantly less informative.

- **Lack of expertise and task interpretation.** Human annotators were mostly unaware of the specific occupations mapped to each DWA during the labeling process. This led to unclear logic for aggregating tasks and occupations, as well as some evident discrepancies in labels, demonstrated in Table 4. We experimented with various aggregation methods and discovered that even with a maximum-matching approach (taking the matching human<>model label if one existed), the agreement remained relatively consistent. Ultimately, we collected additional labels for task/occupation pairs where there was significant disagreement.

- **Forward-looking and subject to change, with some early evidence.** Accurately predicting future LLM applications remains a significant challenge, even for experts (*16*). The discovery of new emergent capabilities, changes in human perception biases, and shifts in technological development can all affect the accuracy and reliability of predictions regarding the potential impact of LLMs on worker tasks and the development of LLM-powered software. Our projections are inherently forward-looking and based on current trends, evidence, and perceptions of technological possibilities. As a result, they may change as new advancements arise in the field. For example, some tasks that seem unlikely for LLMs or LLM-powered software to impact today might change with the introduction of new model capabilities. Conversely, tasks that appear exposed might face unforeseen challenges limiting language model applications.

Future research should consider two additional studies: one exploring LLM adoption patterns across sectors and occupations, and another scrutinizing the actual capabilities and limitations of state-of-the-art models in relation to worker activities beyond the scope of our exposure scores. For example, despite GPT-4's recent advances in multimodal capabilities, our $E1$ ratings on direct LLM exposure did not account for vision capabilities (*16*). Future work should consider the impact of such advancing capabilities. Moreover, we acknowledge potential discrepancies between theoretical and practical performance, particularly in complex, open-ended, and domain-specific tasks.

- **Sources of disagreement.** While we did not rigorously examine sources of disagreement, we found a few places where humans and the model tended to get "stuck" in their assessments:
  - Tasks or activities where while an LLM could theoretically help or accomplish the task, adopting it to do so would require multiple people to change their habits or expectations (e.g. meetings, negotiations),
  - Tasks or activities where there is currently some regulation or norm that requires or suggests human oversight, judgment or empathy (e.g. making decisions, counseling), and
  - Tasks or activities where there already exists a technology that can reasonably automate the task (e.g. making reservations).
- **Other barriers to deployment.** The functionality of large language models (LLMs) is significantly influenced by the degree of trust users accord them and the users' adaptability. For example, in the legal realm, the utility of models is contingent upon the professionals' trust in model outputs without recourse to corroborate original documents or conduct independent research. Other factors such as technology affordability, adaptability, worker

and firm predilections, and incentives critically shape the assimilation of LLM-based tools. Consequently, mitigating ethical and safety risks associated with LLMs, including bias, misinformation, and misalignment, could boost their acceptance (*32*). The adoption of LLMs, however, will exhibit variance across economic sectors due to aspects like data accessibility, regulatory landscape, and power-interest distribution. Hence, gaining comprehensive insights into LLM adoption and usage by workers and firms necessitates a deeper dissection of these subtleties, which our rubrics currently overlook.

- **Possible assumption of augmentation given rubric design.** Our principal findings derive from a prompt assessing potential for time reduction in task completion without compromise on quality, which might inadvertently assume task augmentation as the primary exposure vector. While we refrain from arbitrating between augmentation, automation, or other technological impacts, we introduce an automation rubric with a distinct prompt aimed at measuring the replacement potential of human labor by machines. Our results generally align with the variation in exposure scores, partially attributed to automation exposure. However, it's noteworthy that our automation scores have been annotated exclusively by GPT-4, without human evaluator involvement, suggesting a need for caution in their interpretation.

## 9.2 Comparison to Earlier Efforts

We compare our measurements to previous efforts documenting the distribution of automation exposure in the economy and find broadly consistent results. Most other technology exposure measures we examine are statistically significantly correlated with our preferred exposure measure, while measures of manual routineness and robotics exposure show negative correlations. The variance explained by these earlier efforts (*6–8, 18, 19, 78*), along with wage controls, ranges from 60 to 72%, indicating that 28 to 40% of the variation in our AI exposure measure remains

unaccounted for by previous technology exposure measurements.

This paper aims to build on a number of previous empirical studies examining the occupational exposure to advances in AI and/or automation. Previous studies have used a variety of methods, including:

- Using occupational taxonomies like O*NET to characterize which occupations have routine vs. non-routine and manual vs. cognitive task content (*41, 78*).
- Mapping text descriptions of tasks to descriptions of technological advances in patents. (*8, 79*)
- Linking capabilities of AI systems to occupational abilities and aggregating exposure estimates to the occupations where those abilities are required. (*7, 48*)
- Mapping the results of AI task benchmark evaluations (ImageNet, Robocup, etc.) to 59 worker tasks through a set of 14 cognitive abilities drawn from the cognitive science literature. (*49*)
- Expert labeling of automation potential for a set of O*NET occupations where experts had high confidence, combined with a probabilistic classifier to estimate automation potential for the remainder of O*NET occupations. (*18*)
- Developing a rubric for evaluating the "suitability for machine learning" (SML) of activities that workers are completing in the economy (*6, 19, 77*).

We provide a set of summary statistics on many of these prior efforts in Table 11.

This paper's methodology primarily builds upon the SML approach by developing a rubric to evaluate the overlap between LLM capabilities and worker tasks as reported in the O*NET database. Table 12 presents the results of OLS regressions of our new LLM exposure measurements on

occupation-level exposure measures from (*7*) ("AI Occupational Exposure Score" in the table), (*18*) (Frey & Osborne Automation), scores from all three technologies in (*8*), normalized routine manual and cognitive scores from (*78*), and (*6, 19*) (SML). We also use annualized occupational salaries from the most recent BLS Occupational Employment Survey as a control. There are four separate output variables representing new scores in this paper that are predicted by earlier efforts.

GPT-4 Exposure Rating 1 corresponds to our overall exposure rubric as evaluated by GPT-4, where full exposure potential is coded as 1, no exposure potential is coded as 0, and partial exposure (E2 in our labeling scheme) is coded as 0.5. GPT-4 Exposure Rating 2 is scored similarly for overall exposure, but with a slightly different prompt. The results are very similar across the two prompts. Human Exposure Rating represents the same rubric as in GPT-4 Exposure Rating 1 but is scored by humans, as discussed in an earlier section of the paper. These results correspond to the $E1 + 0.5 * E2$ set of statistics presented above.

The results across each type of measurement are consistent. We find generally positive and statistically significant correlations between our LLM exposure measures and previous measurements targeting software and AI. Encouragingly, the SML exposure scores by occupation show significant and positive associations with the exposure scores we develop in this paper, demonstrating a level of cohesion between the two studies with similar approaches. The Webb software and AI patent-based measures, SML, and normalized (demeaned and divided by standard deviation) routine cognitive scores all exhibit positive associations with some of our measures. AI Occupational Exposure from (*7*) is the most correlated in a bivariate regression with our scores but not statistically significant in the multivariate case. Overall these prior efforts explain 60 to 70 percent of the variation in LLM exposure.

Software, SML, and routine cognitive scores all show positive and statistically significant associations with LLM exposure scores at a 1% level. Coefficients on AI scores from (*8*) are also

positive and statistically significant at a 5% level, but our secondary prompt on overall exposure to LLMs in columns 3 and 4 does not exhibit a statistically significant relationship. For the most part, the AI Occupational Exposure Score is not correlated with our exposure measures. Webb's Robot exposure scores, routine manual task content, and the overall Automation metric from (*18*) are all negatively correlated with our primary GPT-4 and human-assessed overall exposure ratings, conditional on the other measurements. This negative correlation reflects the limited exposure of physical tasks to LLMs. Manual work is not exposed to LLMs or even LLMs with additional systems integration for the time being.

Low correlations with (*7*) and (*18*) could potentially be explained by differences in approaches. Linking AI capabilities to worker abilities or scoring exposure directly based on the occupation's characteristics, rather than aggregating up to the occupation from DWA or task-level scoring (as in the SML paper and our own), offer a slightly different perspective on the content of occupations.

In all regressions, the $R^2$ ranges between 60.7% (column 3) and 72.8% (column 5). This suggests that our measure, which explicitly focuses on LLM capabilities, has between 28 and 40% unexplained variance compared to other measurements. Particularly in the case of AI-related exposure scores, we anticipate that a combination of other measurements would have a strong correlation with our scores. However, earlier efforts had limited information about the future progress of LLMs or LLM-powered software. We expect that our understanding of future machine learning technologies is similarly imperfectly captured by our rubric today.

| | Min | 25th Perc. | Median | 75th Perc | Max | Mean | Std. Dev. | Count |
|---|---|---|---|---|---|---|---|---|
| GPT-4 Exposure Rating 1 | 0.00 | 0.13 | 0.34 | 0.50 | 1.00 | 0.33 | 0.22 | 750 |
| GPT-4 Exposure Rating 2 | 0.00 | 0.09 | 0.24 | 0.40 | 0.98 | 0.26 | 0.20 | 750 |
| Human Exposure Rating | 0.00 | 0.09 | 0.29 | 0.47 | 0.84 | 0.29 | 0.21 | 750 |
| Software (Webb) | 1.00 | 25.00 | 50.00 | 75.00 | 100.00 | 50.69 | 30.05 | 750 |
| Robot (Webb) | 1.00 | 22.00 | 52.00 | 69.00 | 100.00 | 48.61 | 28.61 | 750 |
| AI (Webb) | 1.00 | 28.00 | 55.00 | 82.00 | 100.00 | 54.53 | 29.65 | 750 |
| Suitability for Machine Learning | 2.60 | 2.84 | 2.95 | 3.12 | 3.55 | 2.99 | 0.18 | 750 |
| Normalized Routine Cognitive | -3.05 | -0.46 | 0.10 | 0.63 | 3.42 | 0.07 | 0.86 | 750 |
| Normalized Routine Manual | -1.81 | -0.81 | -0.11 | 0.73 | 2.96 | 0.05 | 1.01 | 750 |
| AI Occupational Exposure Score | 1.42 | 3.09 | 3.56 | 4.04 | 6.54 | 3.56 | 0.70 | 750 |
| Frey & Osborne Automation | 0.00 | 0.07 | 0.59 | 0.88 | 0.99 | 0.50 | 0.38 | 681 |
| Log Avg. Salary | 10.13 | 10.67 | 11.00 | 11.34 | 12.65 | 11.02 | 0.45 | 749 |

Table 11: Summary statistics for a suite of prior efforts to measure occupational exposure to AI and automation. We have also included summary statistics for measurements newly presented in this work. We include all measures from (*8*), normalized routine cognitive and manual scores from (*78*) (means may deviate slightly from 0 due to imperfect matching of occupational groups), Suitability for Machine Learning from (*6, 19, 77*), AI Occupational Exposure from (*7*), and Automation exposure from (*18*). We include as many occupations as we can match, but since O*NET taxonomies have changed as these measures have been developed, some of the roles may be missing from the most recent version of O*NET 6-digit occupations.

| | GPT-4 Exposure Rating 1 | | GPT-4 Exposure Rating 2 | | Human Exposure Rating | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Software (Webb) | 0.03406*** | 0.03705*** | 0.03347*** | 0.03580*** | 0.02893*** | 0.03032*** |
| | (0.00942) | (0.00942) | (0.00940) | (0.00935) | (0.00940) | (0.00945) |
| Robot (Webb) | −0.10799*** | −0.11568*** | −0.10777*** | −0.11399*** | −0.10609*** | −0.10957*** |
| | (0.00909) | (0.00877) | (0.00982) | (0.00941) | (0.00834) | (0.00806) |
| AI (Webb) | 0.02370*** | 0.02667*** | 0.01062 | 0.01330 | 0.01996** | 0.02116** |
| | (0.00897) | (0.00871) | (0.00896) | (0.00875) | (0.00902) | (0.00878) |
| Suitability for Machine Learning | 0.05445*** | 0.04959*** | 0.05250*** | 0.04841*** | 0.03599*** | 0.03389*** |
| | (0.00831) | (0.00815) | (0.00812) | (0.00801) | (0.00736) | (0.00717) |
| Routine Cognitive | 0.05663*** | 0.05892*** | 0.04069*** | 0.04303*** | 0.03061*** | 0.03139*** |
| | (0.00760) | (0.00767) | (0.00749) | (0.00754) | (0.00576) | (0.00574) |
| Routine Manual | −0.11213*** | −0.11439*** | −0.09446*** | −0.09618*** | −0.11111*** | −0.11218*** |
| | (0.00790) | (0.00794) | (0.00821) | (0.00823) | (0.00745) | (0.00748) |
| AI Occupational Exposure Score | 0.00695 | 0.01726** | −0.01077 | −0.00186 | 0.00441 | 0.00877 |
| | (0.00775) | (0.00741) | (0.00812) | (0.00780) | (0.00643) | (0.00592) |
| Frey & Osborne Automation | −0.01143* | −0.01493** | −0.00138 | −0.00460 | −0.01471** | −0.01608** |
| | (0.00694) | (0.00696) | (0.00759) | (0.00746) | (0.00712) | (0.00702) |
| Log Avg. Salary | 0.05804*** | | 0.04863*** | | 0.02531 | |
| | (0.01870) | | (0.01860) | | (0.01727) | |
| Constant | −0.31072 | 0.32778*** | −0.27579 | 0.25903*** | 0.01717 | 0.29577*** |
| | (0.20590) | (0.00493) | (0.20502) | (0.00520) | (0.19008) | (0.00468) |
| N | 680 | 681 | 680 | 681 | 680 | 681 |
| $R^2$ | 0.68741 | 0.68212 | 0.60737 | 0.60198 | 0.71213 | 0.71126 |

Table 12: Regression of LLM-exposure scores on prior measures of occupational exposure to AI and automation. We also include annualized wages from the BLS-OES survey in May 2021. Each measure is standardized to mean zero and variance 1. Generally we find strong positive associations with previous efforts, though large residual variance to still be explained by our new measures. Columns 1 and 2 are based on our main $E1 + 0.5 * E2$ exposure measure from GPT-4 ratings. Columns 3 and 4 are based on a similar slightly different exposure rubric also rated by GPT-4 for robustness. Columns 5 and 6 reflect human ratings on the same rubric as columns 1 and 2.

| | GPT-4 Exposure Rating 1 | | GPT-4 Exposure Rating 2 | | Human Exposure Rating | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Software (Webb) | 0.00113*** | 0.00123*** | 0.00111*** | 0.00119*** | 0.00096*** | 0.00101*** |
| | (0.00031) | (0.00031) | (0.00031) | (0.00031) | (0.00031) | (0.00031) |
| Robot (Webb) | −0.00378*** | −0.00405*** | −0.00377*** | −0.00399*** | −0.00371*** | −0.00383*** |
| | (0.00032) | (0.00031) | (0.00034) | (0.00033) | (0.00029) | (0.00028) |
| AI (Webb) | 0.00080*** | 0.00090*** | 0.00036 | 0.00045 | 0.00067** | 0.00071** |
| | (0.00030) | (0.00029) | (0.00030) | (0.00030) | (0.00030) | (0.00030) |
| Suitability for Machine Learning | 0.29522*** | 0.26888*** | 0.28468*** | 0.26245*** | 0.19514*** | 0.18373*** |
| | (0.04503) | (0.04418) | (0.04404) | (0.04342) | (0.03990) | (0.03886) |
| Normalized Routine Cognitive | 0.06601*** | 0.06868*** | 0.04743*** | 0.05015*** | 0.03568*** | 0.03659*** |
| | (0.00886) | (0.00894) | (0.00872) | (0.00879) | (0.00671) | (0.00669) |
| Normalized Routine Manual | −0.11147*** | −0.11371*** | −0.09390*** | −0.09561*** | −0.11045*** | −0.11152*** |
| | (0.00785) | (0.00789) | (0.00817) | (0.00818) | (0.00741) | (0.00744) |
| AI Occupational Exposure Score | 0.00993 | 0.02465** | −0.01537 | −0.00265 | 0.00630 | 0.01252 |
| | (0.01107) | (0.01059) | (0.01160) | (0.01114) | (0.00918) | (0.00845) |
| Frey & Osborne Automation | −0.03024* | −0.03950** | −0.00364 | −0.01217 | −0.03890** | −0.04253** |
| | (0.01835) | (0.01841) | (0.02007) | (0.01972) | (0.01883) | (0.01858) |
| Log Avg. Salary | 0.05804*** | | 0.04863*** | | 0.02531 | |
| | (0.01870) | | (0.01860) | | (0.01727) | |
| Constant | −1.12937*** | −0.45743*** | −0.96117*** | −0.39935*** | −0.47078* | −0.17706 |
| | (0.26859) | (0.15327) | (0.26365) | (0.15017) | (0.24684) | (0.13256) |
| N | 680.00000 | 681.00000 | 680.00000 | 681.00000 | 680.00000 | 681.00000 |
| $R^2$ | 0.68741 | 0.68212 | 0.60737 | 0.60198 | 0.71213 | 0.71126 |

Table 13: Regression of LLM-exposure scores on prior measures of occupational exposure to AI and automation. We also include annualized wages from the BLS-OES survey in May 2021. Each measure is kept in its original scale, with the exception of routine cognitive and routine manual scores from (*78*). Those two scores are standardized to mean zero and variance 1. Generally we find strong positive associations with previous efforts, though large residual variance to still be explained by our new measures. Columns 1 and 2 are based on our main $E1 + 0.5 * E2$ exposure measure from GPT-4 ratings. Columns 3 and 4 are based on a similar slightly different exposure rubric also rated by GPT-4 for robustness. Columns 5 and 6 reflect human ratings on the same rubric as columns 1 and 2.

## 9.3 Analysis of LLM Impact on Developer Activity Using Github Forks Data

Understanding the impact of LLM deployment on software development is critical to assessing their potential as GPTs. To test the hypothesis that the release of ChatGPT caused an increase in development activity, we obtained data on forks over time from 1,167 public Github repositories that, based on the content of their Readme files, descriptions, and code, can be reasonable assumed to be differentially impacted by the release of a LLM like ChatGPT. We then use a synthetic difference-in-difference approach (as in (*80*)) to estimate the causal impact of ChatGPT release on weekly Github forks in impacted repositories. Forks are a feature of the Github platform that allows users to create a personal copy of someone else's repository. In practice, this is often useful when a user wants to use the existing code as a starting point for their own work without directly modifying the original repository. We hypothesize that repositories focused on LLM-related coding projects should see a large spike in forks after the release of ChatGPT relative to those developing code for statistical analysis and data science purposes.

In order to identify a set of repositories that we hypothesize would be heavily impacted by the release of ChatGPT, we first attempted to scrape forks data from the 600 most forked repositories with the query terms "LLM" or "language model" or "code generation" or "LaMDA" or "GPT" or "BERT" in either the repository readme file, description, or in code saved in the repository. This group represents our treated group in the synthetic difference-in-difference set up. To identify a comparable group of repositories that we hypothesize would experience a limited change in activity after ChatGPT release (and could therefore act as a control group), we attempted to scrape forks data from the 600 most forked public repositories with the query terms "statistics" or "data science" or "linear regression" or "time series" or "statistical analysis" in either the repository readme file, description, or in code saved in the repository. The first set of query terms is meant to identify repositories focused on building with LLMs (LLM repos) while the second is

meant to identify repositories having the do with data science and statistics (DataSci repos). 24 repositories matched both sets of query terms and were conservatively assigned to the LLM repo group prior to analysis. A small subset of the queried repositories did not allow scraping of the readme files and/or forks data, yielding a final sample of 591 LLM repos and 576 DataSci repos. In total, these repositories were forked 1,645,190 times between March 7, 2008 and April 28, 2023. In Figure 6, monthly forks are plotted over time for all repositories in the sample between 2016 and 2023. Figure 7 shows monthly forks over time split by query group.

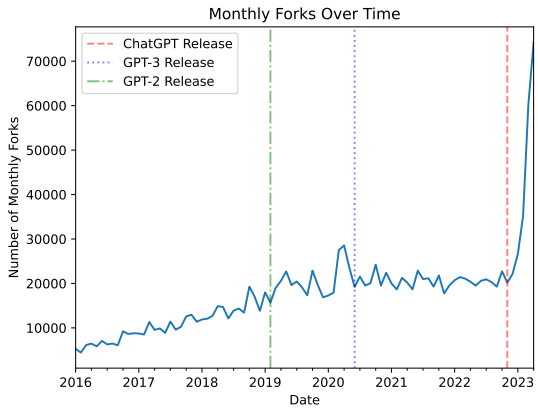



Figure 13: Number of forks per month between January 2016 and April 2023, all queried Github repositories

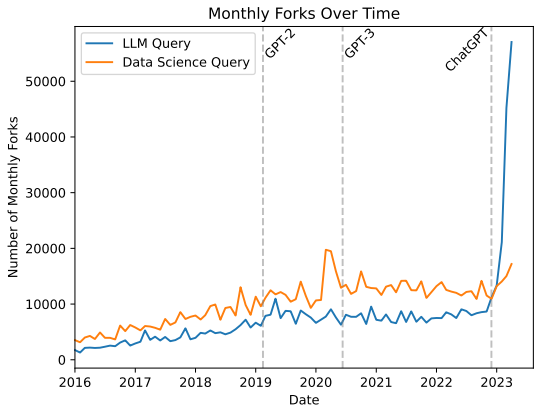



Figure 14: Number of forks per month between January 2016 and April 2023, Github repositories split by query group

These figures paint a clear picture of increased developer activity in the LLM repositories coinciding with the release of ChatGPT. To estimate the causal impact that ChatGPT release had on developer activity in LLM repos, we present results in Table 14 of three regression specifications. Results for our main specification presented in Column 1 are that of a synthetic difference-in-difference estimator. As discussed in (*80*), the synthetic difference-in-difference combines aspects of the synthetic control and traditional difference-in-difference methodologies. This method is particularly useful for causal inference under quasi-experimental settings where treatment and control groups do not have parallel pre-treatment trends in aggregate (overcoming

a common challenge to causal identification with traditional difference-in-difference methods). Additionally, the synthetic difference-in-difference estimator does not require that the treated unit be within a convex hull of control units (a requirement when using a traditional synthetic control estimator for causal identification).

Equation 1 represents our main synthetic difference-in-difference specification. The dataset we are using is a balanced panel with $N$=1,167 repositories and $T$=122 time periods (weeks), where the number of weekly forks for repository $i$ in period $t$ is denoted by $Y_{it}$, and exposure to the ChatGPT treatment is denoted by $W_{it} \in \{0, 1\}$. $N_{DS}$=576 (DataSci) repositories are the control units, [8] while the remaining $N_{LLM}$=591 (LLM) repositories that (we contend) are differentially exposed to ChatGPT impacts after time $T_{pre}$.

As with a synthetic control estimator, we compute weights $\hat{\omega}_{\text{sdid}}$ that match pre-treatment trends in the number of forks of DataSci repos with those for LLM repos, e.g., $\sum_{i=1}^{N_{DS}} \hat{\omega}_{\text{sdid}} Y_{it} \approx N_{LLM}^{-1} \sum_{i=N_{DS}+1}^{N} Y_{it}$ for all $t = 1, \ldots, T_{pre}$. We also use time weights $\hat{\lambda}_{\text{sdid}}$ that balance pre-ChatGPT weeks with post-ChatGPT weeks. The synthetic difference-in-difference estimator then incorporates these weights in a two-way fixed effects regression to estimate the average causal effect of ChatGPT on weekly forks in the LLM repos (denoted by $\tau$):

$$(\hat{\tau}_{\text{sdid}}, \hat{\mu}, \hat{\alpha}, \hat{\beta}) = \arg \min_{\tau,\mu,\alpha,E1+0.5*E2} \left[ \sum_{i=1}^{N} \sum_{t=1}^{T} \left(Y_{it} - \mu - \alpha_i - E1 + 0.5 * E2_t - W_{it}\tau\right)^2 \hat{\omega}_{\text{sdid},i} \hat{\lambda}_{\text{sdid},t} \right] \tag{1}$$

We use the *sdid* stata package developed by (*81*) to estimate this average treatment effect on treated units (ATT) using the two-way fixed effects regression in equation 1 above (presented in Column 1 of table 14) as well as a traditional OLS difference-in-difference estimator (Column 2),

[8] Causal identification here relies on the assumption that these repositories were differentially exposed to ChatGPT due to their focus on code development for data science purposes rather than a focus on LLM-related coding projects. In practice, due to ChatGPT's relevance to any kind of coding, the DataSci repos can plausibly be expected to see an increase in forking activity as well, meaning that the ATT coefficient estimates should be interpreted as a conservative estimate relative to the true ATT.

and a synthetic control (Column 3). Graphical representations of each are presented in Figures 15 and 16.

Table 14: Impact of ChatGPT Release on Developer Activity (Weekly Forks) on Github

| | (1) Synth. DiD Estimate on Weekly Forks | (2) DiD Estimate on Weekly Forks | (3) Synth. Control Estimate on Weekly Forks |
|---|---|---|---|
| LLM Repo x PostChatGPT Treatment | 7.385*** | 7.795*** | 8.416*** |
| | (1.351) | (1.336) | (1.378) |
| N | 142,374 | 142,374 | 142,374 |

Standard errors in parentheses
*p<0.10, **p<0.05, ***p<0.01

Table 14 shows the results of the synthetic difference in difference, estimating an average treatment effect of ChatGPT release on LLM repos of 7.385 additional forks per week, with a standard deviation of 1.351 forks per week. Since this estimator allows us to relax the parallel trends assumption, we treat it as our main specification. However, the similar point estimate in column 2 and the plotted pre-treatment trends in Figure 15 suggest that parallel trends is a reasonable identifying assumption in this instance. The synthetic control estimates in column (3) further validate the direction and magnitude of impact. We interpret these results as a strong sign the the release of ChatGPT has already driven a significant increase in developer activity, adding evidence that today's most advanced LLMs are sparking complementary development on top of the base-models.

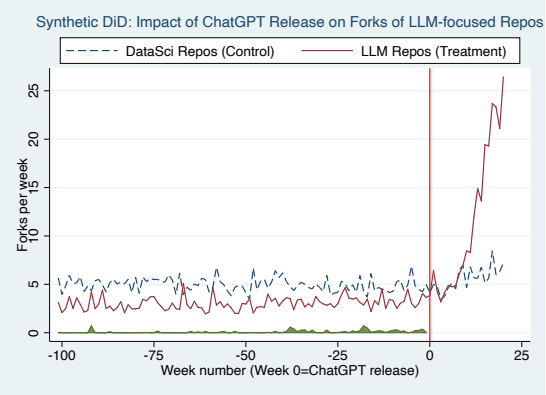



Figure 15: Pre- and post- ChatGPT release trends in weekly forks across repo groups. Weights used to average the pre-treatment time-periods across groups for the synthetic difference-in-difference are presented at the bottom of the graph. ATT estimates are presented in Columns 1 and 2 of Table 14.

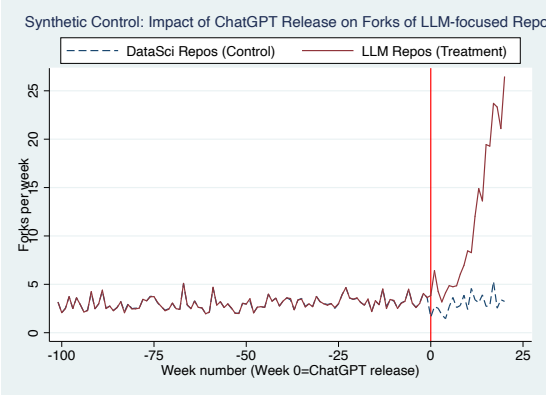



Figure 16: Synthetic Control trend line. ATT estimate is presented in Column 3 of Table 14.

## 9.4 Survey Results with ChatGPT Users

In February 2023 OpenAI invited members of the OpenAI Discord Channel to participate in a voluntary survey to gather information about how they are using ChatGPT in their work. Respondents opted in to participating in the survey, are from a range of countries around the globe, and have a variety of occupations (see Table 15). The survey asked members whether they used ChatGPT in their current jobs and specifically asked them, "What are some ways that you use ChatGPT in your work? Please list and describe up to three."

Table 15: Job Categories for Discord Survey Respondents

| **Job Category** | **Count** |
|---|---|
| Software Eng/Dev/Game Developer/Programmer | 259 |
| Other | 225 |
| Manager/Executive/Entrepreneur/Business Owner | 120 |
| Student | 84 |
| Educator | 69 |
| Writer | 56 |
| Marketing/Advertising | 40 |
| Designer | 29 |
| Researcher | 27 |
| Visual Arts | 21 |
| Data Scientist or Analyst | 20 |
| Healthcare | 17 |
| Sales | 13 |
| Finance | 9 |
| Customer Service | 4 |
| Real Estate | 3 |
| **Total Responses** | **877** |

To assess the overlap between the tasks marked as exposed by our rubric and those that respondents report actually using the model to help with in their jobs, we first get the text embeddings for the tasks that survey respondents report using ChatGPT for in their work. We then compare the cosine similarity of those embeddings matrices to the embeddings for all O*NET tasks, identifying the single O*NET task that is most similar to each user-reported task. The table below presents the most-reported O*NET tasks that respondents reported using ChatGPT for, as well as the GPT-4 and human annotations made using our exposure rubric. Human annotations marked two of the top 30 tasks as E0 (i.e. no exposure) and the GPT-4 labels marked one task as such. Overall, humans scored 13 out of the top 30 tasks as E1 (exposed to an LLM like ChatGPT directly) and 15 of the tasks as E2 (exposed to an LLM + additional software). GPT-4 was much likely to rate these top tasks as E1 than E2, rating 24 of the 30 tasks as E1 and only 5 as E2.

| Task | GPT-4 Label | Human Label | count |
|---|---|---|---|
| Research new software or design concepts. | E2 | E2 | 45 |
| Compose and edit scripts or provide screenwriters with story outlines from which scripts can be written. | E1 | E1 | 20 |
| Help children with homework and school work. | E1 | E1 | 20 |
| Compose explanatory summaries of contents of books and other reference materials. | E1 | E1 | 17 |
| Write and execute scripts to automate tasks, such as parsing large data files. | E1 | E2 | 14 |
| Answer students' questions. | E1 | E0 | 14 |
| Develop or implement data analysis algorithms. | E1 | E2 | 13 |
| Answer user inquiries regarding computer software or hardware operation to resolve problems. | E2 | E1 | 12 |
| Train robots, using artificial intelligence software or interactive training techniques, to perform simple or complex tasks, such as designing and carrying out a series of iterative tests of chemical samples. | E1 | E2 | 12 |
| Write online blog entries that address news developments or offer additional information, opinions, or commentary on news events. | E1 | E1 | 12 |
| Write algorithms or programming code for ad hoc robotic applications. | E1 | E2 | 11 |
| Write new functions or applications in programming languages to conduct analyses. | E1 | E1 | 11 |
| Write computer programs or scripts to be used in querying databases. | E1 | E1 | 11 |
| Compile and write documentation of program development and subsequent revisions, inserting comments in the coded instructions so others can understand the program. | E1 | E2 | 10 |
| Maintain understanding of current Web technologies or programming practices through continuing education, reading, or participation in professional conferences, workshops, or groups. | E2 | E2 | 10 |
| Write supporting code for Web applications or Web sites. | E1 | E2 | 9 |
| Conduct market research analysis to identify search query trends, real-time search and news media activity, popular social media topics, electronic commerce trends, market opportunities, or competitor performance. | E2 | E2 | 8 |
| Write articles and books. | E1 | E1 | 8 |
| Prepare, rewrite and edit copy to improve readability, or supervise others who do this work. | E1 | E1 | 8 |
| Optimize Web site exposure by analyzing search engine patterns to direct online placement of keywords or other content. | E2 | E2 | 8 |
| Post and update content on the company's Web site and social media outlets. | E1 | E1 | 7 |
| Write articles, bulletins, sales letters, speeches, and other related informative, marketing and promotional material. | E1 | E1 | 7 |
| Read copy or proof to detect and correct errors in spelling, punctuation, and syntax. | E1 | E1 | 7 |
| Conduct searches to find needed information, using such sources as the Internet. | E1 | E2 | 6 |
| Design or develop automated testing tools. | E1 | E1 | 6 |
| Design and create special programs for tasks such as statistical analysis and data entry and cleaning. | E1 | E2 | 5 |
| Present drafts and ideas to clients. | E1 | E2 | 5 |
| Discuss the design with the client. | E0 | E2 | 5 |
| Compose, type, and distribute meeting notes, routine correspondence, or reports, such as presentations or expense, statistical, or monthly reports. | E1 | E0 | 5 |
| Collaborate with computer manufacturers and other users to develop new programming methods. | E1 | E2 | 5 |

Table 16: Top 30 most reported O*NET tasks that OpenAI discord survey respondents report using ChatGPT for in their work. There were a total of 954 survey responses.

# Acknowledgments

Thank you to the group of annotators who helped us annotate task exposure, including Muhammad Ahmed Saeed, Bongane Zitha, Merve Özen Şenen, J.J., and Peter Hoeschele. We also thank

Lauryn Fuld, Ashley Glat, Michael Lampe, and Julia Susser for excellent research assistance. We thank Miles Brundage for significant feedback on this paper.

We thank Todor Markov and Vik Goel for setting up the infrastructure we use to run our rubrics with GPT-4. We thank Lama Ahmad, Donald Bakong, Matt Beane, Seth Benzell, Erik Brynjolfsson, Parfait Eloundou-Enyegue, Morgan Frank, Carl Frey, Sarah Giroux, Gillian Hadfield, Johannes Heidecke, Jonathan Hersh, Alan Hickey, Lorin Hitt, Eric Horvitz, Kartik Hosanagar, Shengli Hu, Ashyana Kachra, Christina Kim, Katya Klinova, Daniel Kokotajlo, Gretchen Krueger, Michael Lampe, Aalok Mehta, Larissa Schiavo, Daniel Selsam, Sarah Shoker, Prasanna Tambe, and Jeff Wu for feedback and edits at various stages of the project.

# LLM assistance statement

GPT-4 and ChatGPT were used for writing, coding, and formatting assistance in this project.