The Impact of ChatGPT on Task Structures

William Ratnavale*

May 16, 2025

Abstract

On November 30, 2022, the announcement of ChatGPT's advanced capabilities introduced a potential shock to the U.S. labor market. This paper evaluates the announcement's impact on establishment-level task demand, contributing to the growing literature that seeks to identify the causal effects of AI-related shocks on firm-level labor demand. The empirical analysis is disciplined by a neural network-based model and implements a difference-in-difference strategy that exploits a dataset of 38 million job ads posted by a balanced panel of establishments. Within this framework, changes in task demand for establishments with high and medium exposure to Artificial Intelligence (AI) following the ChatGPT announcement are compared to changes in task demand by establishments with lower exposure. The results suggest a deviation from the trends observed during computerization. While advancements in AI, such as LLMs, continue to reduce demand for routine cognitive tasks (down 4.1pp), they also appear to decrease demand for nonroutine analytical tasks (down 3.5–4.3pp) and nonroutine interactive tasks (down 2.0–2.6pp). At the same time, routine manual tasks see a 3.5pp increase in demand, and nonroutine manual tasks remain largely unaffected. These findings suggest that, unlike computerization, which complemented non-routine analytical tasks, generative AI is instead reducing demand for them, marking a shift in technological impact.

*Email: wratnava@umd.edu, University of Maryland, College Park. I would like to thank Professor Katharine Abraham and Professor Sergio Urzua for their advice and guidance throughout this process.

I. Introduction

Recent advancements in natural language processing (NLP) and deep learning (DL) have led to remarkable progress in artificial intelligence (AI).[1] These breakthroughs paved the way for large language models (LLMs), with ChatGPT, released in November 2022, as the first LLM to successfully perform tasks like writing and customer support that previously only humans could do. ChatGPT was a shock to the market, reaching one million users within five days. Later research found that it reduces task completion time by 40% for college-educated workers performing professional writing tasks. Separately, estimates from the Business Trends and Outlook Survey show that AI usage in U.S. companies increased by 46% between September 2023 and February 2024 (Noy & Zhang, 2023; Bonney et al., 2024). This rapid adoption of AI in the U.S. raises important questions about AI's broader effects on human labor and whether it will ultimately complement or substitute human workers. Specifically, which tasks will AI substitute or complement, and to what extent?

To understand the implications of this question, it is helpful to consider how past technological innovations, such as computerization, have reshaped labor markets. Computerization enabled the automation of routine, codifiable tasks—particularly those found in middle-skill occupations like clerical, administrative, and manufacturing jobs—while complementing non-routine work requiring problem-solving and technical expertise. As a result, beginning in the 1990s, computerization contributed to a sharp polarization in the labor market, with job growth concentrated at the high- and low-wage ends of the wage distribution, while middle-wage jobs experienced significant declines (Autor, Katz, & Kearney, 2006). At the same time, non-routine tasks involving adaptability and face-to-face interaction—often found in low-skill service sector jobs catering to the wealthy—were harder to routinize and automate, helping explain the growth in these roles (Autor, Levy, & Murnane, 2003). Autor and Dorn (2013) argue that routine-intensive occupations are particularly vulnerable to displacement by automation, as they were during the computerization wave. Much like computerization displaced middle-skill, routine-intensive jobs, advancements in AI, which can automate a wider range of tasks, including some non-routine work, may extend this trend and put pressure not just on the middle of the wage distribution but on higher- and lower-wage jobs as well.

Current research suggests that AI technologies' advancing automation ca-

[1] Generally, AI refers to any software that analyzes large datasets utilizing highly complex algorithmic techniques to generate predictions about future outcomes. Machine learning (ML), NLP, and DL are some of the most popular forms of AI, each serving a specific purpose. ML processes extensive datasets to form highly accurate predictive models. One example would be feeding pollution data into an ML algorithm to better understand each pollutant's impact on the environment. Looking next at NLP, this technique leverages ML to interpret and process human language, whether through speech or text. The most advanced AI technique, DL, employs artificial neural networks to analyze data in sophisticated ways similar to how the human brain functions. This AI technique, in particular, is behind most of the recent advances in self-driving cars and, when combined with NLP techniques, has given rise to chatbots like ChatGPT.

pabilities to contribute to a declining labor share of national income, with ambiguous effects on inequality (Acemoglu & Restrepo, 2022). These distributional outcomes appear to stem partly from heterogeneous productivity effects both across and within occupations. Some studies find that within occupations such as lawyers, software engineers, management consultants, and customer service agents, lower-skilled or less-experienced workers benefit more from AI-driven productivity gains than their higher-skilled counterparts (Choi et al., 2023; Dell'Acqua et al., 2023; Peng et al., 2023; Brynjolfsson et al., 2023). However, other research suggests that higher-income or more highly-skilled workers are more likely to experience productivity boosts from AI (Felten et al., 2023; Eloundou et al., 2024). While some studies find that AI adoption slightly increases wages in occupations with high software use (Felten et al., 2019), others report significant declines in non-AI job postings and shifts toward AI-intensive roles (Acemoglu et al., 2022). These mixed findings suggest that AI's distributional effects depend not only on broad occupational categories but also on the specific tasks workers perform, as tasks can differ widely in their potential for automation or complementation—even within the same job title. These concerns make it critical to understand which types of tasks are most vulnerable to substitution by AI and which are likely to be complemented.

This paper investigates the causal impact of ChatGPT's release on establishment-level task demand, focusing on identifying the types of tasks most vulnerable to substitution or complementarity by generative AI. It employs a difference-in-differences framework comparing establishments with higher and lower pre-existing exposure to AI technologies to achieve this. The paper adopts the framework of traditional economic literature for task classification, specifically the approach of Spitz-Oener (2006), which categorizes tasks into five types: non-routine analytical tasks such as research, planning, and evaluation; non-routine interactive tasks such as selling, coordinating, and delegating work; routine cognitive tasks such as bookkeeping and data entry; routine manual tasks such as assembly line work; and non-routine manual tasks such as equipment repair and housekeeping. This classification scheme has been widely used in economics to quantify task content and analyze how the demand for various types of work evolves with technological changes, as demonstrated in prior research by Atalay et al. (2018, 2020).

The paper extends the Spitz-Oener framework by employing a Continuous Bag of Words (CBOW) neural network-based model to identify similar tasks not explicitly listed in the taxonomy.[2] Since the original task categories are not exhaustive, the CBOW model projects words into a high-dimensional vector space to capture their semantic relationships based on contextual usage within

[2]The Continuous Bag of Words (CBOW) model is a type of neural network used in natural language processing to predict words given their surrounding context. It works by learning distributed representations of words, capturing their meanings in a vector space. This paper will follow the approach used by **?**, who utilized this model to classify task mentions and compute similarities between them, offering a framework for understanding how different tasks are related in terms of their semantic meaning. For the full list of seed task words, as well as additional words identified through the CBOW model, see Appendix A1.

the dataset. For example, the word "analyze" represents a non-routine analytical task in the Spitz-Oener framework. The model finds similar words, like "summarize," which also fall into this category. Instead of relying only on exact task words from the literature, the CBOW model leverages context to detect abbreviations and newly emerging language, measuring task demand more flexibly and accounting for how language evolves. Using this extended methodology, I apply the CBOW model to a novel dataset of approximately 38 million quarterly job descriptions from 2020 to 2024 posted by a balanced panel of establishments to proxy establishment-level task demand over time.

To identify tasks compatible with ChatGPT and other LLMs, I employ Webb's (2020) methodology on the newly available Artificial Intelligence Patent Dataset (AIPD 2023) from the United States Patent and Trademark Office (USPTO). This dataset identifies AI-related patents from 1976 to 2023, providing detailed information about the AI patents, including the specific facets of AI they pertain to, such as natural language processing, computer vision, machine learning algorithms, robotics, and other subfields of AI (Giczy et al., 2024). Using a dependency parsing algorithm, I sift through the AI-related patents to identify the tasks most exposed to AI technologies. These tasks are then aggregated at the occupation level to measure occupational AI exposure. To compute AI exposure scores at the establishment level, I calculate a weighted average of occupational AI exposure, where the weights correspond to the share of job postings for each occupation within the establishment prior to the release of ChatGPT. In my main specification, high-exposure establishments are those in the 60th percentile or higher of AI exposure, while the medium-exposure group consists of establishments in the 40th-60th percentile, and the low-exposure establishments are those in the 40th percentile or lower. My paper focuses primarily on these thresholds, but Appendix A2 includes robustness checks using alternative AI exposure cut-offs.

My results suggest a deviation from the trends observed during computerization. Although advances in AI, such as LLMs, continue to reduce demand for routine cognitive tasks (down 4.1pp), they also appear to decrease demand for nonroutine analytical tasks (down 3.5 - 4.3pp) and nonroutine interactive tasks (down 2.0-2.6 pp) - a departure from earlier trends where both increased. At the same time, routine manual tasks experience a 3.5 percentage point increase in demand, while non-routine manual tasks remain largely unaffected. These preliminary findings indicate a nuanced shift in task demand, with some parallels to computerization and notable reversals. Estimates are based on a balanced panel of 18,000 establishments, which likely skews toward larger firms that consistently post job openings, limiting generalizability to smaller or irregular employers.

II. Theory

In the spirit of Acemoglu et al. (2022), I consider a setting in which establishments allocate tasks between human labor and AI-powered algorithms in response to advances in AI productivity. Each establishment e produces output y_e by combining tasks x from a set T, where T is the set of all feasible tasks. Each task $x \in T$ belongs to exactly one of five disjoint subsets $T_k \subseteq T$, where $k \in \{1, 2, 3, 4, 5\}$ indexes non-routine analytical, non-routine interactive, non-routine manual, routine cognitive, and routine manual tasks.

At the task level, establishments allocate production between human labor $\ell_e(x)$ and AI-powered algorithms $a_e(x)$, subject to a constant elasticity of substitution (CES) production function whose elasticity depends on the task type k. Let $k(x)$ denote the task type associated with task x, and let $\sigma_{k(x)}$ be the corresponding elasticity of substitution. Then task-level output is given by:

$$y_e(x) = \left(\gamma_\ell(x) \ell_e(x)^{\frac{\sigma_{k(x)}-1}{\sigma_{k(x)}}} + \gamma_a(x) a_e(x)^{\frac{\sigma_{k(x)}-1}{\sigma_{k(x)}}} \right)^{\frac{\sigma_{k(x)}}{\sigma_{k(x)}-1}},$$

where $\gamma_\ell(x)$ and $\gamma_a(x)$ represent the productivity of human and AI labor for task x, respectively. This formulation allows for heterogeneous substitution patterns between AI and human labor by permitting the elasticity of substitution to vary across task types. As a result, increases in AI productivity can have differential effects on the demand for tasks, depending on how easily AI can substitute for human input in each task type.

Assuming $\sigma = \infty$ for all task types, consider a small increase in the productivity of AI algorithms $y_a(x)$ within the task set $T^A \subseteq T$, where the set T^A denotes the tasks that AI can perform more cost-effectively than human labor. Under these conditions, the resulting change in human labor demand ℓ_e in establishment e is given by:

$$d \ln \ell_e = (-1 + (\varepsilon_e \cdot \rho_e - 1) \cdot \pi_e) \cdot \text{exposure to AI}_e,$$

where $\varepsilon_e > 1$ is the output demand elasticity, $\rho_e > 0$ is the pass-through rate (i.e., the extent to which cost reductions are passed through to prices), and $\pi_e \geq 0$ is the percentage reduction in task-level production costs due to AI. [4]

This equation highlights an important trade-off: if $(\varepsilon_e \cdot \rho_e - 1) \cdot \pi_e > 1$, then the productivity effect dominates, and AI exposure can increase human labor demand even if there's some degree of substitution between AI and labor. In this case, AI raises overall output, leading to increased demand for human labor despite the substitution effect. However, if the productivity effect is weaker, AI will displace human labor, particularly in the tasks that can be automated.

This framework provides a basis for understanding how increases in AI productivity affect human labor demand. The release of ChatGPT represents a salient example, as it expanded the range of tasks AI can perform within T^A. This development enables an empirical test of how establishments with varying AI exposure reallocate tasks and adjust labor demand in response.

[4]The proof for this assertion is provided in Appendix A of Acemoglu et al. (2022), along with the expressions for ρ_e and π_e.

III. Data

To estimate the impact of ChatGPT on establishment-level task demand, I construct a quarterly panel dataset where the unit of observation is each establishment-quarter. The two main challenges in this analysis are (1) identifying task demand for different task types and (2) determining establishment-level exposure to ChatGPT. To measure task demand, I use Revelio Labs' job postings data, which provides establishment-level information on posted job descriptions. To quantify AI exposure, I use USPTO data on AI-related patents, linking them to establishments based on the extent to which AI can perform the tasks required by the occupations within the establishment. Below, I describe both data sources in detail and outline the construction of the panel.

A) Revelio Labs Data

Revelio Labs collects job posting data from 250k+ employer websites, major job boards, and staffing firm job boards, compiling over 2 billion job postings from 5.25 million companies. Revelio's proprietary algorithms deduplicate job postings. Data coverage begins in 2007, with more comprehensive postings captured from 2021 onward. The dataset includes a wide range of variables for each job posting, such as job title, Standard Occupational Classification (SOC) code, company name, job text, North American Industry Classification System (NAICS) industry code, salary information, metro area, posting removal date, and posting date. The data is collected at a monthly frequency; however, to have more observations in my balanced panel, I aggregate it to the quarterly level. I focus on the period from 2020 to 2024, as this range includes significantly more postings with non-missing NAICS codes, allowing for better industry-level comparisons and reweighting. Establishments are typically defined as distinct physical locations of a firm. However, due to data limitations, I define an establishment as a unique company within a particular metro area. The central goal of using this dataset is to quantify establishment-level demand for different types of tasks. To measure task demand, I count the number of task-related keywords that appear in each job description.

A.1 Measuring Task Demand from Job Descriptions

To illustrate this approach, I present a sample job posting from the Revelio Labs dataset, shown in Figure 1. The example shows (1) the original job text collected and (2) the cleaned and lemmatized version used in the analysis after removing stop words and normalizing word forms. Task mentions are identified using a set of keywords expanded via a CBOW model trained on the cleaned job postings. Below, I highlight matched task-relevant terms in the cleaned version to show how task classification is performed in practice.

Original Job Description

FULL TIME 8A-4PCHARACTERISTIC DUTIES AND RESPONSIBILITIESESSENTIAL FUNCTIONS1. Assist in the planning developing, organizing, implementing, evaluating, and directing of the Maintenance Department. (Includes department policies, procedures, position descriptions. etc.)2. Assist in the development and implementation of departmental policies and procedures to ensure that the maintenance of the premises, facility and equipment is current at all times.3. Develop and maintain a good work rapport with intra-department personnel, and other departments within the facility, to ensure that maintenance programs can be properly planned and maintained to meet the needs of the facility.4. Ensure that the plant and equipment are properly maintained for patient/resident comfort and convenience.5. Supervise/monitor work of department supervisors/personnel to ensure compliance of directives and established procedures.6. Interpret policies and procedures to department personnel as necessary.7. Ensure that the Maintenance Department's administrative procedures are followed.Job Type: Full-time

---

Cleaned Job Description

essential function assist planning develop organize implement evaluate directing maintenance department include department policy procedure position description etc assist development implementation departmental policy procedure ensure maintenance premise facility equipment current time develop maintain good work rapport intra department personnel department within facility ensure maintenance program properly plan maintain meet need facility ensure plant equipment properly maintain patient resident comfort convenience supervise monitor work department supervisor personnel ensure compliance directive establish procedure interpret policy procedure department personnel necessary ensure maintenance department administrative procedure follow

Figure 1: **Example Job Description and Extracted Task Mentions**

Notes: Task-relevant keywords are highlighted in color. Dark Blue = Non-Routine Analytical, Teal = Non-Routine Interactive, Dark Red = Routine Manual. "*equipment*" was included as a Routine Manual task word based on its semantic similarity to known Routine Manual task verbs. To identify semantically similar terms, I trained a CBOW word embedding model on the full corpus of job postings in my balanced panel and computed cosine similarity scores between each candidate word and all words in manually curated seed sets for each task type. Cosine similarity represents how similar two words are based on the contexts in which they appear. It is defined as:

$$\text{cosine\_similarity}(u, v) = \frac{u \cdot v}{\|u\|\|v\|}$$

Following the methodology of Atalay et al. (2020), I included a word in the expanded task list if its cosine similarity to any word in the seed set exceeded a threshold of 0.55. In this case, "*equipment*" was added to the Routine Manual category because it appeared in highly similar contexts to verbs such as *operate*, *equip*, and *control*, despite being a noun. Because phrases like *operate equipment* and *maintain equipment* frequently appear in job postings, including "*equipment*" helps capture mentions of routine manual work even when the associated verb is not retained during preprocessing.

In this example, the cleaned job description contains a total of 9 task mentions: 4 classified as Non-Routine Analytical (*planning, evaluate, plan, interpret*), 3 as Non-Routine Interactive (*organize, directing, supervise*), and 2 as Routine Manual (*equipment*). I define task demand as the frequency with which employers reference different task types in job postings. This reflects the relative emphasis placed on different kinds of work in hiring language — a proxy for the capabilities firms seek to acquire through new hires. While this approach does not measure how much time workers spend on each task once hired, it captures task demand expressed by employers during the hiring process. In other words, when a firm posts a job, it reveals the types of tasks it intends to expand its capacity for by bringing in new workers. Job postings, therefore, provide a public signal of which tasks firms value and prioritize at a given point in time. Because task demand is not only about hours worked but also about the kinds of work firms seek more at the margin, shifts in task-related language likely reflect meaningful changes in the composition of labor demand. In the full dataset, I aggregate task counts across all job postings within each establishment-quarter to construct quarterly measures of task demand by task type.

A.2 Representativeness of the Job Postings Data

A key question about this dataset is its representativeness, as it primarily draws from online job postings. To address this, I compare it to the Job Openings and Labor Turnover Survey (JOLTS) job openings estimates, but the two datasets measure different aspects of labor demand. Revelio postings data track the flow of new job openings, whereas JOLTS vacancy estimates measure the stock of job openings at the end of each month. This fundamental difference means that a single job opening could either appear multiple times in JOLTS (if it remains unfilled across months) or not appear at all (if it is posted and filled within the same month).

To make these measures more comparable, there are two possible approaches. First, one could estimate end-of-month vacancy stocks in Revelio by leveraging both the posting date and removal date recorded in the data. By identifying whether a job posting was still active at the end of each month, one could construct a stock measure of open positions that better aligns with JOLTS vacancy estimates. Second, one could use new hires from JOLTS as a proxy for job postings, comparing them to the number of postings in Revelio. However, both approaches have limitations. Not all posted jobs get filled, and the timing mismatch between a job posting and its eventual hire complicates direct comparisons. A crude adjustment would be to compare Revelio postings in month t to JOLTS new hires in month t+1, which implicitly assumes that each job posting is filled within a month. Since my data is quarterly, I apply this approach at the quarterly level. In this case, the Revelio data correspond to calendar quarters (January–March, April–June, July–September, October–December), while the JOLTS data use three-month periods shifted by one month (February–April, May–July, August–October, November–January).

To assess how well the Revelio data align with JOLTS, I examine whether

Revelio vacancy stock estimates track JOLTS vacancy stock estimates and whether Revelio posting estimates track JOLTS hires estimates for the following month. Figure 2 compares Revelio vacancy stock estimates with JOLTS openings, while Figure 3 compares Revelio posting estimates with JOLTS hires in the subsequent month.

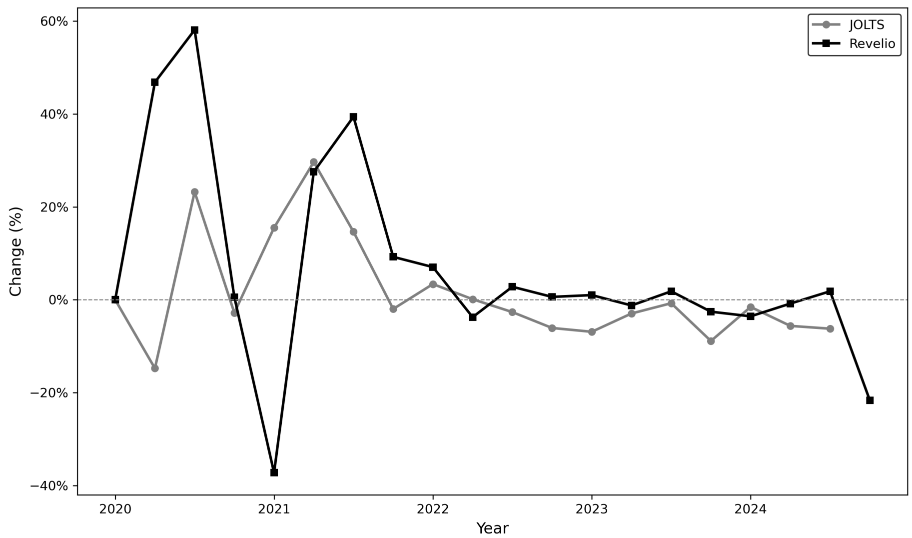


Figure 2: **Comparison of Revelio Vacancy Stocks and JOLTS Openings**

Note: Data are quarterly. JOLTS series are seasonally unadjusted to align with Revelio job postings, which are not seasonally adjusted.

As can be seen in Figure 2, there is a large increase in Revelio vacancy stocks in 2020. While this does not closely track JOLTS openings, this discrepancy is expected since 2020 marks a period when Revelio significantly expanded its data coverage, adding a much larger volume of postings. Apart from the sharp dip in mid-2021, the Revelio vacancy stock estimates appear to track JOLTS relatively well from the second half of 2021 onward.

When we turn to hires in Figure 3, we observe a similar pattern: there is a massive increase in Revelio postings at the beginning due to the expansion in data coverage. However, beyond this initial surge, the Revelio new posting estimates appear to align more closely with JOLTS hires than the Revelio vacancy stock estimates align with the JOLTS vacancy estimates. From 2021 onward, the trends in Revelio postings and JOLTS hires track each other relatively well, suggesting that while Revelio postings may not perfectly capture vacancy stocks, they provide a reasonable proxy for hires.

A potential concern with using the Revelio data is the over- or underrepresentation of certain industries, which could affect the comparability of these measures to JOLTS. To address this, in Appendix A3, I present industry-level

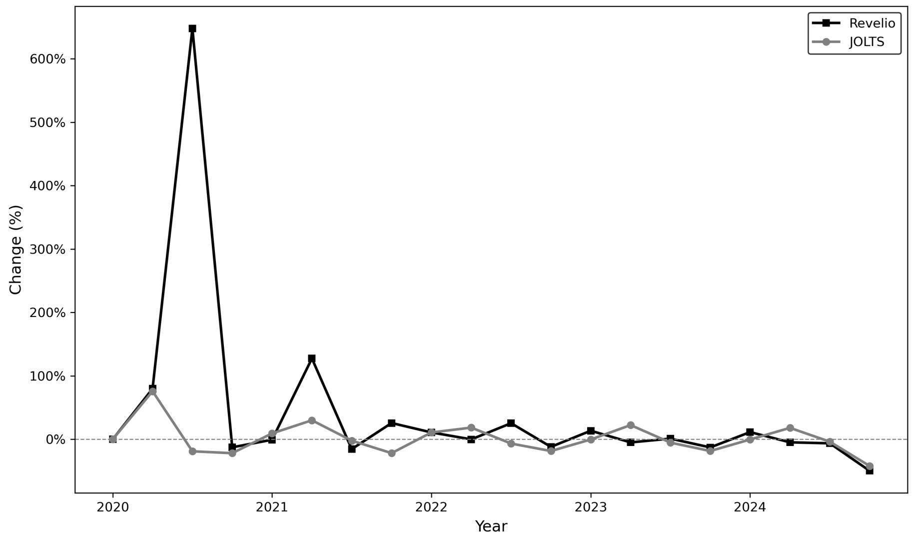


Figure 3: **Comparison of Revelio Postings and JOLTS Hires**

Note: Data are quarterly. JOLTS hires are seasonally unadjusted to maintain comparability with Revelio postings, which are not seasonally adjusted.

comparisons between Revelio and JOLTS to assess how well the distributions align. Additionally, to ensure that aggregate trends are not driven by shifts in industry composition over time, I reweight the task demand measures according to industry shares in JOLTS hires each quarter in Figure 4. This choice is motivated by the fact that JOLTS hires data track Revelio postings more closely than JOLTS job openings, making it a more reliable benchmark for industry-level adjustments.

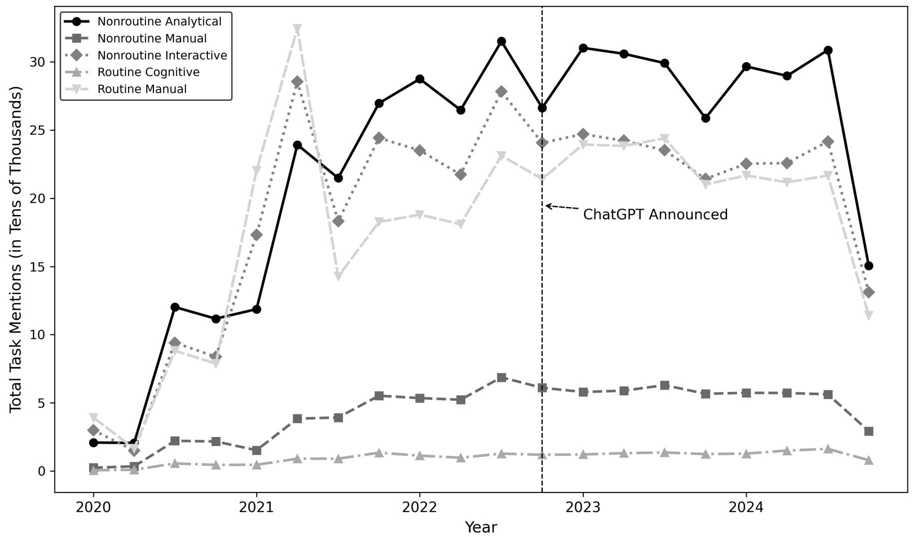


Figure 4: **Quarterly Trends in Task Demand**

Note: Task demand is reweighted by industry shares each quarter to account for potential over- or under-representation of industries in the Revelio data.

B) AI Exposure Index

A key challenge in defining AI exposure lies in accurately identifying which tasks are exposed to AI and determining the degree of that exposure, a process that is inherently complex given the evolving and multifaceted nature of AI technologies. Currently, three prominent measures of AI exposure address this challenge: Felten et al.'s AI Occupational Impact (AIOI) measure (2018, 2019), Brynjolfsson, Mitchell, and Rock's Suitability for Machine Learning (SML) index (2018, 2019), and Webb's AI exposure score (2020).

The AIOI measure, developed by Felten et al. (2019), links AI advancements to activities associated with occupations using data from the Electronic Frontier Foundation (EFF) AI Progress Measurement dataset and the O*NET database from the U.S. Department of Labor. The EFF dataset tracks the progress of nine AI applications—such as computer vision, natural language processing, and robotics—from 2010 to 2015. O*NET provides detailed occupational data, including abilities rated by importance and prevalence within each occupation. Felten et al. combined these ability ratings with the progress in AI technologies, using crowdsourced responses from Amazon Mechanical Turk (mTurk) to evaluate how AI advancements influence AI exposure across occupations.

Similarly, the Suitability for Machine Learning (SML) index, developed by Brynjolfsson, Mitchell, and Rock (2019), uses a 23-item rubric to assess the automation potential of tasks through machine learning. This rubric is applied to Direct Work Activities (DWAs) from O*NET, and the scoring is facilitated through the CrowdFlower crowdsourcing platform with high SML scores indicating tasks that are highly susceptible to automation via machine learning.

Lastly, the AI exposure score, developed by Webb (2020), identifies AI-related keywords manually (e.g., "neural network") and uses Google Patents Public Data, provided by IFI CLAIMS Patent Services, to collect patents containing these keywords. Webb employs a dependency parsing algorithm to extract key verb-noun combinations (e.g., "diagnose disease") from patent titles to represent tasks that AI can perform. The frequency of these verb-noun pairs is used to quantify the AI exposure of a task. Webb then multiplies the AI exposure for each task by the importance of that task within an occupation to calculate the overall AI exposure score for each occupation based on O*NET 2020 data.

A key consideration when evaluating AI exposure measures is their respective limitations. A weakness of the AIOI and SML measures is their reliance on crowd-sourced evaluations of the potential of task automation. While this approach provides insight, it introduces variability and subjectivity, making results less replicable and generalizable. Additionally, the AIOI measure only considers AI advancements from 2010 to 2015, which limits its relevance for understanding more recent AI impacts. Webb's AI exposure index avoids these pitfalls by using patent data to identify AI-related innovations, offering a more objective foundation. However, Webb's measure also has limitations, as it relies on keyword-based patent searches to identify AI-related patents and only incorporates data up to 2020, predating major advancements like GPT-3 and subsequent AI developments. This time lag means it does not fully capture the impacts of more recent generative AI technologies.

This paper builds on Webb's (2020) methodology by leveraging the newly available Artificial Intelligence Patent Dataset (AIPD 2023). Released by the Office of the Chief Economist (OCE) for USPTO, the AIPD identifies which of the 15.4 million U.S. patent documents—comprising patents and pre-grant publications from 1976 through early 2023—are related to AI. This updated dataset leverages highly advanced natural language processing techniques, specifically BERT for Patents, to accurately identify which patents are most exposed to AI technologies. By utilizing BERT's ability to understand context and semantics within text, the dataset improves the classification of patents, including those near the decision "boundary" between AI and non-AI technologies.

The AIPD 2023 dataset provides information on AI patents, including patent abstracts, filing dates, and model-assigned confidence scores that indicate the likelihood of a patent belonging to a specific AI category. These confidence scores range in certainty, with thresholds such as 93% signifying a high degree of confidence in classification. To ensure that the selected patents are highly relevant to LLMs, I restrict my analysis to patents where the classification model is at least 93% confident in its assignment. The AIPD 2023 categorizes AI patents into eight groups: machine learning, evolutionary computation, natural language processing, vision, speech, knowledge processing, planning and control, and AI hardware. This categorization is essential for identifying the patents most relevant to LLMs like ChatGPT in late 2022. Specifically, I focus on natural language processing, knowledge processing, and machine learning patents, as they encompass key functions that LLMs are designed to perform, such as text generation, language understanding, and knowledge synthesis.

This paper focuses on generative AI, as opposed to prior literature that primarily examines first-generation AI, such as rule-based systems and narrow automation. This shift represents a significant improvement, as it allows me to isolate the capabilities of LLMs specifically. Categories like evolutionary computation, planning, control, and AI hardware are excluded, as they pertain to tasks and technologies outside the scope of what LLMs primarily address. For example, evolutionary computation involves optimization algorithms inspired by natural selection, often used in specialized fields like robotics or algorithm design, which are far removed from the natural language processing and conversational capabilities of LLMs. While LLMs can assist in describing or documenting such processes, they do not directly perform or implement evolutionary computation methods themselves. By narrowing the focus to patents directly tied to LLM functionalities, this approach helps make the exposure scores developed in this paper more relevant for analyzing how ChatGPT's release and subsequent LLM advancements may impact the labor market.

Building on Webb's methodology (2020), I improve the identification of tasks in patents by incorporating a BERT-based Longformer model fine-tuned on a Stanford Question Answering Dataset (SQuAD), a dataset for question-answering tasks. Webb's title-based method is limited because many patent titles lack verbs, making it difficult to extract verb-noun pairs. For example, the title "Automated provisioning of relational information for a summary data visualization" contains no verb, preventing meaningful extraction. Using only titles, this method identifies verb-noun pairs in only 14% of patents. To improve coverage, I extract tasks from patent abstracts. However, abstracts are lengthy and contain numerous verb-noun pairs, making it challenging to isolate the most relevant task description. To address this, I use BERT to extract the key function of each patent by posing the question: "What is the purpose of the described apparatus or method?" For example, consider the patent abstract shown in Figure 5.

A related information provision system may be used to $\boxed{\text{identify constituent data for a selected portion of summary data}}$ on a summary data visualization such as a chart or graph. The portion of the summary data may be selected by a user, a system event, and/or another process. The constituent data may include values and/or metadata for one or more data sets summarized by the summary data, formulas and/or other information, which may indicate how the summary data are obtained from the constituent data. Information regarding the constituent data ("related information") may be displayed for a user on a constituent visualization or the like. If the information is already displayed, the system may draw the user's attention to the information. If the information is displayed in a different view, the system may display a navigational element leading to it, or direct the user's attention to an existing navigational element.

Figure 5: **Example patent abstract with the extracted key phrase highlighted.**

From this, BERT identifies the key functional phrase: "identify constituent data for a selected portion of summary data." Using dependency parsing, I extract "identify data" as the key verb-noun combination of interest. This approach allows me to recover verb-noun pairs in 80% of patents. This approach substantially improves task identification compared to the 14% coverage achieved using titles alone.

With all verb-noun pairs extracted from AI-related patents, the next step is to assess how exposed each pair is to AI and match them to tasks in the O*NET database. However, a direct comparison is challenging because different words can describe the same underlying task (e.g., analyze data vs examine data). Without standardization, such cases would be treated as distinct, potentially understating the true exposure of a task to AI.

To address this, I begin by decomposing each O*NET task into verb-noun pairs. For example, the task "Collect and analyze data to detect deficient controls, duplicated effort, extravagance, fraud, or non-compliance," listed under Accountants and Auditors, yields two verb-noun pairs: (Collect, data) and (analyze, data). I then standardize the vocabulary of these pairs using WordNet. For nouns, I replace each with its second-level WordNet hypernym, aggregating similar terms two steps up the semantic hierarchy. The tradeoff in selecting the second level is that lower levels preserve fine-grained distinctions, while higher levels risk overgeneralization. For example, the nouns software developer and programmer remain distinct at the lowest level but are grouped under software engineer at the second level. Similarly, mechanical engineers and civil engineers are grouped under engineers. At higher levels, such as level six, almost all occupations collapse into broad terms like entity, losing meaningful distinctions. The second level thus provides a balance between specificity and generalization. For verbs, I apply a similar strategy based on WordNet's synonym synsets, but add a constraint: bidirectionality. Two verbs are considered equivalent only if each appears in the other's synset. Once bidirectional synonym verb groups are formed, I replace each verb with the most frequently occurring synonym verb within the O*NET data. For example, if analyze and examine are bidirectional synonyms but analyze appears more frequently in O*NET, then both examine data and analyze data are standardized as analyze data. Likewise, predict and forecast are grouped, and if predict is more common, then forecast market trends is standardized as predict market trends. I apply this standardization process to both the O*NET and patent-derived verb-noun pairs.

Having standardized task identification, I next calculate the AI exposure of each verb-noun pair based on its frequency in AI-related patents. Specifically, the exposure score for a given verb-noun pair c is computed as:

$$r_c = \frac{f_c}{\sum_{c \in C} f_c}$$

where f_c denotes the frequency of verb-noun pair c in AI patents, and C is the set of all standardized verb-noun pairs that appear in those patents.

To compute the AI exposure of a given task k from occupation o, I sum the exposure scores r_c for all verb-noun pairs c within the set S_k, where S_k denotes the set of standardized verb-noun pairs extracted from task k in the O*NET occupation data:

$$\text{TaskExposure}_k = \sum_{c \in S_k} r_c$$

Each occupation consists of multiple tasks, each of which may vary in relevance to that occupation.[3] To compute occupation-level AI exposure, I aggregate the task-level exposure scores using O*NET-provided task importance weights $w_{k,o}$, which reflect how important task k is to occupation o, and normalize by the total weighted count of verb-noun pairs:

$$\text{Exposure}_o = \frac{\sum_{k \in K_o} [w_{k,o} \cdot \text{TaskExposure}_k]}{\sum_{k \in K_o} [w_{k,o} \cdot |S_k|]}$$

where K_o is the set of tasks associated with occupation o, and $|S_k|$ is the number of verb-noun pairs extracted from task k. This yields a weighted average of AI exposure across tasks, accounting for both the intensity of AI relevance and the importance of each task within the occupation. Without the denominator, occupations with more tasks (k) or tasks containing more verb-noun pairs ($c \in S_k$) would appear more exposed simply due to volume, even if their actual

[3]If a verb-noun pair from an O*NET task does not appear in the AI patent data, it is assigned an exposure score of 0.

per-task exposure is low. The denominator adjusts for this by dividing by the weighted total number of verb-noun pairs—effectively answering: *Of all the task content relevant to this occupation (weighted by importance), what fraction is exposed to AI?*

Using this methodology, Figure 6 presents the estimated AI exposure across major occupational groups, categorized by 2-digit SOC codes according to O*NET. The rankings indicate the occupations most exposed to generative AI under this framework. The second figure presents the same analysis using Webb's (2020) AI exposure data.

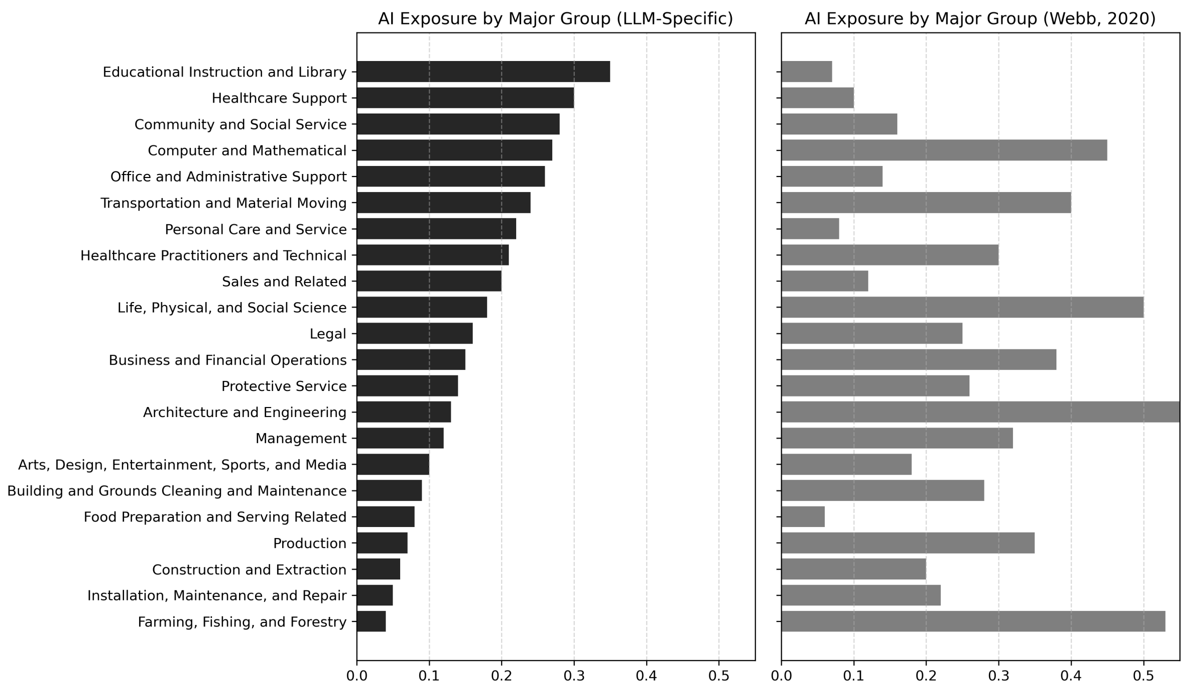


Figure 6: **Comparison of AI Exposure Scores Across Major Occupational Groups. The left panel shows AI exposure based on LLM-specific patents, while the right panel replicates the analysis using Webb's (2020) AI exposure scores.**

A key difference between these two measures is that Webb's exposure scores were developed at a time when AI patents were more focused on robotics, automation, and physical-task augmentation. As a result, some physically intensive occupations—such as architecture, engineering, and farming—appear more exposed to AI in his framework. In contrast, my AI exposure measure, which incorporates more recent patents emphasizing LLMs, finds that exposure is higher for occupations involving educational, administrative, and professional services work-fields that are more directly impacted by advancements in generative AI.

This shift aligns with expectations, given that LLMs primarily affect knowledge-based and language-intensive tasks rather than physical or manual work. For example, educational instruction and library occupations rank among the most exposed in my estimates, whereas they were among the least exposed under Webb's framework. Similarly, office and administrative support now appear highly exposed, reflecting LLMs' capabilities in document processing and summarization.

Lastly, to determine which establishments are exposed to AI, I compute an establishment-level AI exposure score by taking a weighted average of occupational AI exposure scores, where the weights correspond to the share of job postings for each occupation within the establishment. Exposure is calculated based on job postings through the third quarter of 2022, immediately preceding the release of ChatGPT.[4] Establishments are then classified into exposure tiers: low exposure (0–40th percentile), medium exposure (40th–60th percentile), and high exposure (60th–100th percentile).

IV. Methodology

Since task demand across different types within the same establishment is likely serially correlated, and the short-run effects of generative AI may be shaped by frictions in adjustment and hiring, I include a lagged dependent variable to capture persistence in task demand. Given that my post-treatment window spans only two years of quarterly data, this focus on the short run aligns with theories of sticky adjustment, where labor reallocation does not happen instantaneously. However, including a lagged outcome variable raises endogeneity concerns. To address these concerns, I follow the approach of Arellano and Bond (1991) and estimate the following equation using difference Generalized Method of Moments (GMM).[5]

$$y_{i,t}^h = \beta_1^h y_{i,t-1}^h + \beta_2^h (\text{ExposureGroup}_i \times \text{Post}_t) + \beta_3^h X_{i,t} + \lambda_i + \gamma_t + \delta_{s(i),t} + \epsilon_{i,t}^h \quad (1)$$

In this specification, $y_{i,t}^h$ denotes either the log number of task mentions or the task share for task type h at establishment i in quarter t, both of which serve as proxies for task-specific labor demand.

The term $\text{ExposureGroup}_i \times \text{Post}_t$ captures the interaction of medium- and high-exposure group dummies with a post-treatment indicator (equal to 1 after

[4] Robustness checks using alternative cutoff thresholds for defining exposure levels are presented in Appendix A2. The results are relatively stable across these specifications.

[5] Serial correlation may exist both within the same task type over time and across different task types. To assess this, I examine the residual correlation structure (see Appendix A4). While errors across different task types within an establishment are not strongly correlated, there is clear autocorrelation in the residuals within the same task type over time. This supports the inclusion of the lagged dependent variable, consistent with the idea that task demand evolves gradually due to hiring frictions, learning curves, and adjustment costs. Arellano and Bond (1991) propose a dynamic panel estimator that addresses this concern by differencing out establishment fixed effects and using deeper lags of the dependent variable as instruments, thus mitigating endogeneity while capturing persistence in behavior.

the release of ChatGPT), with the low-exposure group serving as the omitted category. The vector of control variables $X_{i,t}$ includes the log of establishment-level vacancy counts, the log of total word counts at the establishment from their job postings, the log average salary, the percent of jobs requiring a bachelor's degree or some college, the percent requiring postgraduate education, and the average O*NET-based computer importance score (ranging from 1 to 5).

The model includes establishment fixed effects λ_i, time fixed effects γ_t, and industry-by-time fixed effects $\delta_{s(i),t}$, where $s(i)$ indexes an establishment's broad 2-digit NAICS industry. These components help account for unobserved heterogeneity, common time shocks, and industry-specific trends over time. Including industry-by-time fixed effects is particularly important in this setting, as it controls for industry-specific shocks that vary over time—such as the differential impact of COVID-19 across sectors during the pre-treatment period. However, this also means that these fixed effects absorb any differences in the effect of ChatGPT across industries. As a result, identification comes from within-industry variation in AI exposure over time, rather than from between-industry comparisons. Incorporating these fixed effects improves the plausibility of parallel trends and enhances comparability across exposure groups, especially in specifications using task shares as outcomes.

Vacancies capture the number of job ads posted by an establishment, which influences the breadth of tasks mentioned (e.g., more vacancies likely imply more roles and thus more tasks). They also partially proxy for hiring behavior and firm size. Word count reflects the level of detail in a posting—longer descriptions naturally include more task mentions. Additionally, systematic differences in word count across establishments may reflect variation in how tasks are communicated or the nature of work itself, making it an important control. The average salary is included because firms with higher or lower wages may systematically demand more or fewer tasks of a given type. Controlling for salary improves the comparability of task demand across exposure groups.

For the coefficient on $\text{ExposureGroup}_i \times \text{Post}_t$ to be interpreted causally, two key assumptions must hold. First, the release of ChatGPT must be exogenous, meaning that it was largely unanticipated by establishments. Second, it must be the case that, in the absence of the ChatGPT release, high exposure, medium exposure and low exposure establishments would have followed similar trends in task demand across all five task categories. If these conditions hold, then the estimated coefficients can be interpreted as causal effects of ChatGPT exposure on task demand.

While $\boldsymbol{\beta}_2$ captures the causal effect of ChatGPT's release on multiple dimensions of task demand, the interpretation varies depending on the specific outcome variable. One of these outcomes is the *log of total task mentions* for a given task type h, while another is the *share of task mentions*—i.e., the proportion of all task mentions at an establishment that belong to task type h.

In the case of log total mentions, the interpretation of β_2^h is straightforward: it reflects the percentage change in the volume of demand for task type h at high-exposure establishments relative to low-exposure ones, conditional on pre-treatment trends and establishment-level covariates.

However, when the outcome is *task share*, the interpretation is more nuanced. Because shares are compositional, β_2^h captures *relative changes* in task demand, how the share of task type h changed compared to other task types, rather than the absolute number of task mentions. A positive coefficient means that task h became more prominent in the overall task mix *relative to expectations*, while a negative coefficient means it became less prominent.

To illustrate, consider a scenario where all nonroutine tasks are fully substituted away following ChatGPT's release—perhaps because AI tools now perform them autonomously. Suppose routine manual tasks also decline, but to a lesser extent. Even though absolute demand for routine manual tasks has fallen, they may now represent nearly *100%* of the remaining task mentions at some establishments. In this case, the *share* of routine manual work increases, and β_2^{RM} may appear strongly positive—not because routine manual work is expanding, but because it declined *less than expected*, relative to other tasks and relative to the control group.

Conversely, imagine that high-exposure establishments dramatically expand their hiring for nonroutine analytical roles, while reducing postings for other tasks. Even if the *absolute demand* for nonroutine analytical tasks rises more than any other task type, the estimated coefficient β_2^{NRA} could still be negative if, *conditional on prior demand patterns, exposure level, and other covariates*, the model would have predicted an even larger increase. In other words, the coefficient is not identifying whether demand rose or fell in a vacuum—it is identifying whether it rose *more or less than expected*, based on the counterfactual trend inferred from the control group.

This highlights why interpretation must focus on β_2^h as a *causal effect on the composition of task demand*, conditional on prior trends and adjustment frictions—not as a literal measure of whether a task became more or less prevalent in absolute terms. Moreover, because the post-treatment window covers only the first two years after ChatGPT's release, the coefficient should be interpreted as a *short-run effect*. It does not represent the eventual *long-run equilibrium* that may emerge as establishments more fully adapt their organizational structures and task allocation in response to generative AI.

This distinction is especially important when comparing these results to earlier studies on computerization, such as Autor, Levy, and Murnane (2003), who examine how the distribution of task types changed across occupations as computer use became more widespread. Their analysis relates changes in the task composition of occupations to the increasing importance of computers over several decades, using 1960 as a baseline. Because their setting reflects long-run adjustments—where labor reallocation has had time to unfold—their estimates are not constrained by the short-run adjustment costs that characterize my setting.

Moreover, while their analysis focuses on how occupational task content evolves in response to technological change, my framework uses establishment-level data to estimate how firms, treated as relevant decision-making agents, respond in the short run to a specific and well-defined technological shock. My estimates are causal, relative to a control group, and conditional on a rich set of covariates, and thus capture a different quantity of interest. Rather than reflecting broad secular trends, they isolate the short-run effects of generative AI exposure.

V. Results

I begin by presenting descriptive evidence on the establishments in my sample, starting with their industry composition. Figure 7 displays the share of establishments by industry across the low, medium, and high AI exposure groups. High-exposure establishments are disproportionately concentrated in sectors such as Private Education and Health Services, Professional and Business Services, Trade, Transportation, and Utilities, and Manufacturing. Low-exposure establishments are similarly concentrated in Trade, Transportation, and Utilities, as well as in Manufacturing and Leisure and Hospitality. Interestingly, industries like Information and Financial Activities comprise only a small share of any group—a somewhat unexpected pattern. The relatively similar sectoral composition of the high and low groups may enhance the plausibility of comparing them in subsequent regressions.

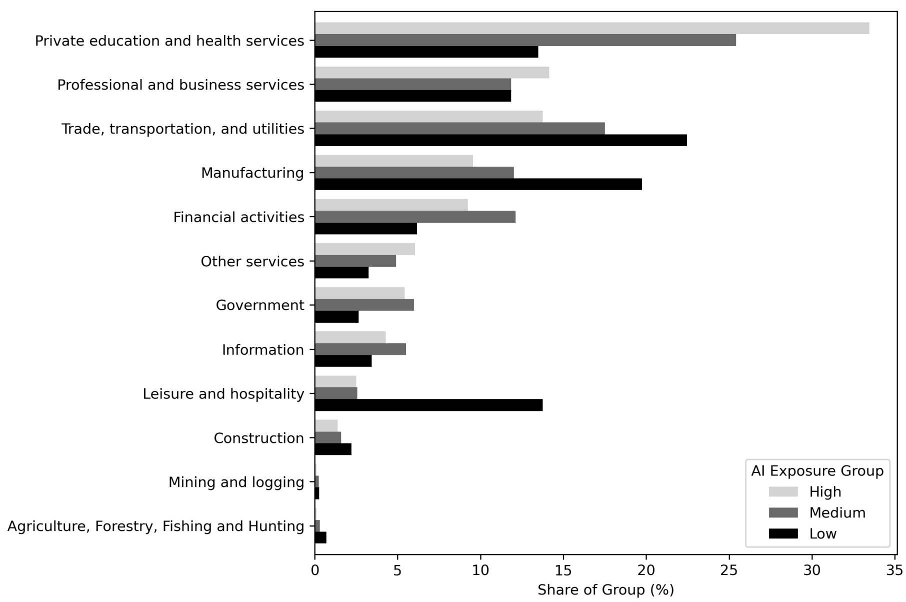


Figure 7: **Industry Distribution Within AI Exposure Groups**

Note: Industries are broad 2-digit NAICS codes. "Low exposure" refers to establishments below the 40th percentile of AI exposure, "Medium exposure" refers to those between the 40th and 50th percentiles, and "High exposure" includes those above the 60th percentile.

Turning to the summary statistics reported in Table 1, establishments in the high- and low-exposure groups each account for roughly 143,500 observations,

Table 1: Summary Statistics by AI Exposure Group, 2020–2024

	NRA	NRI	NRM	RM	RC	Vacancies	Words	Salary	Computer Importance
A. High Exposure to ChatGPT									
Mean	126.56	87.18	18.06	73.87	3.74	112.70	24664.20	60399.10	3.88
Std	531.63	401.04	100.46	291.68	21.30	359.33	90125.89	27650.10	0.60
Observations	143520	143520	143520	143520	143520	143520	143520	143520	143004
B. Medium Exposure to ChatGPT									
Mean	113.65	86.75	16.06	67.46	3.62	116.87	24874.84	59561.95	3.94
Std	418.76	303.20	76.84	174.16	19.97	434.15	115977.58	28133.08	0.52
Observations	71779	71779	71779	71779	71779	71779	71779	71779	71550
C. Low Exposure to ChatGPT									
Mean	53.17	51.13	12.93	53.76	3.35	78.19	14790.98	47958.64	3.59
Std	185.62	169.85	50.44	173.90	32.01	224.44	50126.09	20807.98	0.60
Observations	143518	143518	143518	143518	143518	143518	143518	143518	143224

Notes: Each row represents an establishment–quarter observation. The summary statistics are calculated by pooling all establishment–quarter observations from 2020 to 2024 and averaging within each exposure group. NRA = Nonroutine Analytical, NRI = Nonroutine Interactive, NRM = Nonroutine Manual, RM = Routine Manual, RC = Routine Cognitive. The five task variables represent the total number of mentions of each task type across all job postings within the establishment in that quarter. "Words" refers to the total word count across all job postings. "Computer Importance" is a 1–5 score based on O*NET worker surveys; I compute a weighted average for each establishment–quarter based on the occupation mix of job postings.

while the medium-exposure group includes 71,779. Despite its smaller sample size, the medium-exposure group records the highest averages in vacancy counts (116.87) and job posting word counts (24,874.84), which may reflect differences in establishment size. It also has the highest average computer importance score (3.94), likely due to its greater share of establishments in industries such as Information and Financial Activities, where digital skills are especially salient. The high-exposure group shows the highest levels of nonroutine analytical (126.56) and interactive (87.18) tasks, along with higher salaries and longer postings, consistent with their concentration in complex, skill-intensive sectors. By contrast, low-exposure establishments exhibit the lowest averages across nearly all dimensions, including wages, task complexity, and digital task content, aligning with their presence in lower-skilled, less technologically intensive industries. Given these differences, controlling for establishment characteristics helps make parallel trends more plausible and strengthens the credibility of the difference-in-differences estimates.

I next assess the plausibility of the parallel trends assumption. To do this, I conduct F-tests for pre-treatment trends in each task type, comparing treated (high exposure) and partially treated (medium exposure) groups to the control (low exposure) group. I perform these tests using both the log of total task mentions and task shares as outcomes. The pretrend balance generally fails when using total task counts, but substantially improves when using shares, particularly for routine and interactive tasks. Based on this, I prioritize task shares as the outcome in my main difference-in-differences specifications and focus my interpretation on the 40–60 threshold, where balance is most credible (see Appendix VI., Table 14).

Having established that the 40–60 specification exhibits the most plausible pre-trends—particularly when task shares are used as outcomes—I now turn to estimating the effect of AI exposure on task demand. Specifically, I examine the coefficient β_2^h from Equation 1, which captures the post-ChatGPT difference in task demand between high- and medium-exposure establishments relative to the low-exposure group. Below are five separate tables—one for each task type—reporting estimates of β_2^h under varying sets of controls.

When examining task shares, I find a clear decline in nonroutine analytical (NRA) tasks among more AI-exposed establishments. Relative to the low-exposure group, and conditional on pre-treatment trends and establishment-level covariates, NRA shares fell by 4.3 percentage points in the high-exposure group and 3.5 points in the medium-exposure group. Only the high-exposure group satisfies the parallel trends assumption, making that estimate more plausibly causal. This reflects a short-run, within-establishment shift in the relative composition of tasks following the release of ChatGPT.

Nonroutine interactive (NRI) tasks exhibit a similar pattern: relative to the control group and conditional on covariates, task shares declined by 2.6 percentage points in the high-exposure group and 2.0 points in the medium-exposure group. As with NRA, only the high-exposure group passes the pre-trend test, supporting a causal interpretation.

In contrast, nonroutine manual (NRM) task shares show no statistically sig-

Table 2: Effects of ChatGPT on Nonroutine Analytical Task Demand

	Nonroutine Analytical Task Demand, 2020–24					
	(1)	(2)	(3)	(4)	(5)	(6)
	A. High Exposure to ChatGPT					
Task Mentions	0.043	-0.191	-0.072	-0.119	-0.074	-0.081
	(0.009)	(0.014)	(0.009)	(0.005)	(0.009)	(0.010)
Establishment share	0.000	-0.041	-0.041	-0.009	-0.041	-0.043
	(0.001)	(0.002)	(0.002)	(0.001)	(0.002)	(0.003)
	B. Medium Exposure to ChatGPT					
Task Mentions	0.007	-0.258	-0.093	-0.130	-0.094	-0.099
	(0.016)	(0.019)	(0.011)	(0.008)	(0.011)	(0.012)
Establishment share	-0.004	-0.034	-0.034	-0.014	-0.034	-0.035
	(0.002)	(0.003)	(0.003)	(0.002)	(0.003)	(0.003)
Observations	322,934	321,998	321,967	321,967	321,967	321,967
Covariates:						
Vacancies				✓	✓	✓
Word Count			✓	✓	✓	✓
Salary		✓	✓	✓	✓	✓
% Jobs with bachelor's or some college		✓	✓	✓	✓	✓
% Jobs with postgrad education		✓	✓	✓	✓	✓
Time_FE		✓	✓		✓	✓
Industry x Time_FE						✓

Note: Task mentions, vacancies, word count, and salary are all log-transformed. "High exposure" refers to establishments above the 60th percentile of AI exposure, and "Medium exposure" refers to those between the 40th and 50th percentiles. Estimates are based on difference GMM (Arellano-Bond), instrumenting for lagged dependent variables. Establishment and industry fixed effects are differenced out and already controlled for in this framework. Industry refers to broad 2-digit NAICS codes.

Table 3: Effects of ChatGPT on Nonroutine Interactive Task Demand

	Nonroutine Interactive Task Demand, 2020–24					
	(1)	(2)	(3)	(4)	(5)	(6)
	A. High Exposure to ChatGPT					
Task Mentions	0.019	-0.095	0.002	-0.147	-0.001	-0.015
	(0.010)	(0.015)	(0.010)	(0.005)	(0.010)	(0.010)
Establishment share	-0.017	-0.204	-0.014	-0.154	-0.015	-0.026
	(0.017)	(0.019)	(0.012)	(0.009)	(0.012)	(0.013)
	B. Medium Exposure to ChatGPT					
Task Mentions	-0.027	-0.009	-0.009	-0.024	-0.009	-0.008
	(0.001)	(0.003)	(0.003)	(0.001)	(0.003)	(0.003)
Establishment share	-0.030	-0.022	-0.022	-0.026	-0.022	-0.019
	(0.002)	(0.003)	(0.003)	(0.002)	(0.003)	(0.003)
Observations	322,934	321,998	321,967	321,967	321,967	321,967
Covariates:						
Vacancies				✓	✓	✓
Word Count			✓	✓	✓	✓
Salary		✓	✓	✓	✓	✓
% Jobs with bachelor's or some college		✓	✓	✓	✓	✓
% Jobs with postgrad education		✓	✓	✓	✓	✓
Time_FE		✓	✓		✓	✓
Industry x Time_FE						✓

Note: Task mentions, vacancies, word count, and salary are all log-transformed. "High exposure" refers to establishments above the 60th percentile of AI exposure, and "Medium exposure" refers to those between the 40th and 50th percentiles. Estimates are based on difference GMM (Arellano-Bond), instrumenting for lagged dependent variables. Establishment and industry fixed effects are differenced out and already controlled for in this framework. Industry refers to broad 2-digit NAICS codes.

Table 4: Effects of ChatGPT on Nonroutine Manual Task Demand

	Nonroutine Manual Task Demand, 2020–24					
	(1)	(2)	(3)	(4)	(5)	(6)
	A. High Exposure to ChatGPT					
Task Mentions	-0.056	0.024	-0.021	-0.157	-0.021	-0.033
	(0.009)	(0.015)	(0.013)	(0.007)	(0.013)	(0.014)
Establishment share	0.000	0.003	0.012	-0.094	0.012	0.001
	(0.014)	(0.018)	(0.016)	(0.011)	(0.016)	(0.017)
	B. Medium Exposure to ChatGPT					
Task Mentions	-0.012	0.007	0.007	-0.010	0.007	0.006
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Establishment share	-0.010	0.007	0.007	-0.007	0.007	0.007
	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Observations	322,934	321,998	321,967	321,967	321,967	321,967
Covariates:						
Vacancies				✓	✓	✓
Word Count			✓	✓	✓	✓
Salary		✓	✓	✓	✓	✓
% Jobs with bachelor's or some college		✓	✓	✓	✓	✓
% Jobs with postgrad education		✓	✓	✓	✓	✓
Time_FE		✓	✓		✓	✓
Industry x Time_FE						✓

Note: Task mentions, vacancies, word count, and salary are all log-transformed. "High exposure" refers to establishments above the 60th percentile of AI exposure, and "Medium exposure" refers to those between the 40th and 50th percentiles. Estimates are based on difference GMM (Arellano-Bond), instrumenting for lagged dependent variables. Establishment and industry fixed effects are differenced out and already controlled for in this framework. Industry refers to broad 2-digit NAICS codes.

Table 5: Effects of ChatGPT on Routine Cognitive Task Demand

	Routine Cognitive Task Demand, 2020–24					
	(1)	(2)	(3)	(4)	(5)	(6)
	A. High Exposure to ChatGPT					
Task Mentions	-0.037	-0.023	-0.026	-0.084	-0.026	-0.016
	(0.006)	(0.011)	(0.011)	(0.006)	(0.011)	(0.011)
Establishment share	-0.059	-0.064	-0.051	-0.092	-0.050	-0.041
	(0.010)	(0.014)	(0.013)	(0.010)	(0.013)	(0.013)
	B. Medium Exposure to ChatGPT					
Task Mentions	-0.010	0.007	0.007	-0.010	0.007	0.007
	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Establishment share	-0.010	0.007	0.007	-0.009	0.007	0.007
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Observations	322,934	321,998	321,967	321,967	321,967	321,967
Covariates:						
Vacancies				✓	✓	✓
Word Count			✓	✓	✓	✓
Salary		✓	✓	✓	✓	✓
% Jobs with bachelor's or some college		✓	✓	✓	✓	✓
% Jobs with postgrad education		✓	✓	✓	✓	✓
Time_FE		✓	✓		✓	✓
Industry x Time_FE						✓

Note: Task mentions, vacancies, word count, and salary are all log-transformed. "High exposure" refers to establishments above the 60th percentile of AI exposure, and "Medium exposure" refers to those between the 40th and 50th percentiles. Estimates are based on difference GMM (Arellano-Bond), instrumenting for lagged dependent variables. Establishment and industry fixed effects are differenced out and already controlled for in this framework. Industry refers to broad 2-digit NAICS codes.

Table 6: Effects of ChatGPT on Routine Manual Task Demand

	Routine Manual Task Demand, 2020–24					
	(1)	(2)	(3)	(4)	(5)	(6)
	A. High Exposure to ChatGPT					
Task Mentions	0.049	0.078	0.022	-0.085	0.021	0.023
	(0.009)	(0.016)	(0.011)	(0.005)	(0.011)	(0.011)
Establishment share	0.017	0.017	0.031	-0.064	0.032	0.035
	(0.016)	(0.020)	(0.013)	(0.009)	(0.013)	(0.013)
	B. Medium Exposure to ChatGPT					
Task Mentions	-0.019	0.038	0.038	-0.009	0.038	0.031
	(0.001)	(0.002)	(0.002)	(0.001)	(0.002)	(0.002)
Establishment share	-0.018	0.040	0.040	-0.007	0.040	0.035
	(0.002)	(0.003)	(0.003)	(0.002)	(0.003)	(0.003)
Observations	322,934	321,998	321,967	321,967	321,967	321,967
Covariates:						
Vacancies				✓	✓	✓
Word Count			✓	✓	✓	✓
Salary		✓	✓	✓	✓	✓
% Jobs with bachelor's or some college		✓	✓	✓	✓	✓
% Jobs with postgrad education		✓	✓	✓	✓	✓
Time_FE		✓	✓		✓	✓
Industry x Time_FE						✓

Note: Task mentions, vacancies, word count, and salary are all log-transformed. "High exposure" refers to establishments above the 60th percentile of AI exposure, and "Medium exposure" refers to those between the 40th and 50th percentiles. Estimates are based on difference GMM (Arellano-Bond), instrumenting for lagged dependent variables. Establishment and industry fixed effects are differenced out and already controlled for in this framework. Industry refers to broad 2-digit NAICS codes.

nificant change in either exposure group when compared to the control group, suggesting that this category has not yet been meaningfully affected by generative AI in relative terms.

Routine cognitive (RC) task shares declined by 4.1 percentage points in the high-exposure group relative to the control group, conditional on prior trends and covariates. There is no significant change in the medium group. Since only the high-exposure group satisfies the parallel trends assumption, this decline is more plausibly causal and reflects a shift away from tasks that rely on structured knowledge processing.

Finally, routine manual (RM) task shares increased by 3.5 percentage points in both the medium- and high-exposure groups relative to the low-exposure group, with both estimates passing pre-trend tests. While initially puzzling, given that ChatGPT is not expected to raise demand for manual labor, one possible explanation is compositional: if AI automates a wide range of nonroutine tasks, establishments may retain a greater share of workers who perform essential physical upkeep, logistics, or facilities support. As other task types decline more sharply than expected, routine manual work may constitute a larger share of remaining labor needs, even if its absolute demand is flat or declining. This relative increase is measured against a control group and conditional on pre-treatment trends and covariates, suggesting that routine manual work declined less than anticipated.

Overall, the estimates point to a compelling shift: generative AI appears to reduce shares demanded for nonroutine analytical and interactive tasks—tasks that were previously augmented by technology—and increase demand for routine manual work, which had historically declined. At the same time, it continues the downward pressure on routine cognitive tasks, suggesting that while some effects of generative AI mirror those of prior waves of automation, others represent a profound departure. One potential explanation is offered by Toner-Rodgers (2024), who finds in a randomized controlled trial that AI adoption within establishments tends to shift workers toward evaluating AI-generated output rather than producing novel content themselves. In that study, workers increasingly performed more routine tasks such as checking, verifying, or implementing AI-generated results. This may imply that generative AI substitutes more strongly for idea generation—a core component of nonroutine work—thereby reducing overall demand for such tasks at scale. Routine cognitive tasks, which involve predictable rule-based processing, may also be directly automated by AI systems trained on structured data. In contrast, routine manual work often requires physical presence or dexterity, making it less substitutable by current forms of generative AI and potentially complementary if AI frees up labor that can be reallocated to these roles. Additionally, if generative AI tools require physical setup, oversight, or human involvement in operational workflows, the rise in routine manual task shares may partly reflect the labor needed to manage AI-related capital on-site. Although my analysis does not directly test these mechanisms, these patterns raise important questions about the underlying channels—such as task displacement, reallocation, or complementarity—and highlight the need for future work to disentangle their implications for workers and firms.

VI. Conclusion

There is growing interest in understanding how AI is reshaping the labor market, yet research on its impact at the task level remains limited. While prior studies have explored AI's broader economic effects, few have systematically examined shifts in task demand across the U.S. labor market. This paper builds on the existing literature by providing new evidence on how LLMs, exemplified by ChatGPT, are associated with changes in task demand at the establishment level. While the findings suggest significant shifts in certain task types, the extent to which these reflect causal effects varies depending on the task category.

While my estimates remain stable across different AI exposure thresholds, there are important limitations to consider. One key limitation is the use of task mentions as a proxy for task demand at the establishment level. Just because a task is mentioned more frequently in job descriptions does not necessarily mean it is performed more often on the job. Employers may list tasks they want workers to perform rather than those that are actually central to the job role. Additionally, the classification of tasks is based on text analysis, which may not fully capture roles where character-based attributes (e.g., leadership, adaptability) are emphasized over specific technical tasks. Some of these attributes do not fit neatly into the task categories used in this analysis, potentially leaving gaps in understanding how AI-driven labor shifts are occurring.

Despite these limitations, the findings offer valuable insights into how generative AI is reshaping labor demand. The results suggest that AI's impact on the labor market is markedly different from past technological shifts. While computerization primarily complemented nonroutine tasks and automated routine ones, the introduction of LLMs has not only reduced demand for routine cognitive tasks but also displaced nonroutine analytical and interactive tasks—something prior waves of automation largely complemented. At the same time, there is a notable increase in demand for routine manual tasks, suggesting that AI may be shifting human workers back into routine roles.

This paper documents that the demand for workplace tasks shifted differently depending on the extent to which an establishment is exposed to LLMs. The core findings—declines in nonroutine analytical, nonroutine interactive, and routine cognitive task shares, and a null effect on nonroutine manual tasks—are broadly robust across alternative definitions of exposure thresholds. While the magnitude of effects varies slightly, these directional patterns hold under multiple specifications. These results are not directly comparable to earlier studies on computerization, which typically examine long-run occupational-level changes over several decades. In contrast, my estimates capture short-run, within-establishment responses to the release of generative AI, accounting for adjustment frictions and conditional on pre-treatment trends and observable characteristics. Additionally, while much of the computerization literature emphasizes decompositions of task or employment trends, my approach focuses on identifying a causal effect by comparing establishments with differing levels of AI exposure—offering a complementary, more micro-level view of how new AI technologies are reshaping task demand. Although these estimates capture the short-run causal effects of exposure to generative AI, the long-run equilibrium impacts may differ as firms continue to adapt over time. However, the results offer important information on how firms with greater exposure to generative AI responded differently in the immediate aftermath of ChatGPT's release, highlighting relative changes in task composition that can foreshadow broader transformations in the nature of work.

Overall, these findings underscore the distinct impact of generative AI compared to past technological shifts, raising significant questions about how the workforce should adapt. The fact that AI is displacing nonroutine analytical tasks—something previous waves of automation tended to complement—suggests we're in uncharted territory. More research is needed to understand how to make the most of these changes, whether through better job training, smarter AI policies, or new ways of integrating human and machine work. Figuring out how to maximize AI's benefits while minimizing disruption will be crucial as this technology continues to evolve. If accompanied by thoughtful policies that address potential inequalities and support reskilling efforts, these technologies could ultimately unlock new opportunities and foster greater innovation in the labor market.

Appendix

A1. CBOW Task Word Expansion

This section presents the key terms identified using the CBOW model for each type of task. These keywords illustrate how task types were identified based on the language found in the job postings.

Table 7: Task Type Word Lists

Task Type	Initial List	Expanded List After CBOW	Words Added
Non-Routine Analytical	analyze, analyzing, design, designing, devising, evaluate, evaluating, interpreting, plan, planning, research, researching, sketch, sketching	interprets, interpreting, planning, sketching, prototype, analyzes, summarize, analyze, researching, design, interpret, plan, sketch, layouts, devising, evaluating, evaluate, research, analyzing, layout, blueprint, drawing, designing, deploying	interprets, summarizes, prototype, analyzes, layouts, layout, blueprint, drawing, deploying
Non-Routine Interactive	advertise, advertising, advise, advising, buying, coordinate, coordinating, entertain, entertaining, lobby, lobbying, managing, negotiate, negotiating, organize, organizing, presentation, presentations, presenting, purchase, sell, selling, teaching	organizing, advisement, hallway, teach, conducting, advising, negotiate, advise, succession, negotiates, coordinate, teaching, lobby, entertain, marketing, restrooms, negotiation, negotiating, subcontract, publicity, lobbying, campaign, presentation, advertising, purchase, presenting, coordinating, managing, manage, advertise, organize, selling, presentations, directing, reception, contract, briefing, sell, buying, entertaining	advisement, hallway, conducting, succession, negotiates, teach, marketing, restrooms, subcontract, publicity, campaign, directing, reception, contract, briefing
Non-Routine Manual	accommodate, accommodating, accommodation, renovate, renovating, repair, repairing, restore, restoring, serving	instal, accommodate, restore, accommodation, overhaul, renovate, repair, restoring, accommodating, repair, repairing, renovating, serving, repairs, rebuilding	instal, overhaul, repairs, rebuilding
Routine Cognitive	bookkeeping, calculate, calculating, correcting, corrections, measurement, measuring	calculates, bookkeeping, calculating, correcting, measuring, measurement, corrections, calculate	calculates
Routine Manual	control, controlling, equip, equipment, equipping, operate, operating	equipment, control, operate, equipping, controlling, operating, equip	(No new words added)

Note: The "Words Added" column highlights only new words introduced by the CBOW model that did not appear in the original seed lists. These additions help capture task-relevant vocabulary not explicitly listed in prior classifications.

A2. Robustness Checks by Threshold

Table 8: Effects of ChatGPT on Task Demand, 2020–24

	Task Demand, 2020–24				
	NRA	NRI	NRM	RC	RM
	A. High Exposure to ChatGPT				
Task Mentions	-0.081	-0.015	-0.033	-0.016	0.023
	(0.010)	(0.010)	(0.014)	(0.011)	(0.011)
Establishment share	-0.043	-0.026	0.001	-0.041	0.035
	(0.003)	(0.013)	(0.017)	(0.013)	(0.013)
	B. Medium Exposure to ChatGPT				
Task Mentions	-0.099	-0.008	0.006	0.007	0.031
	(0.012)	(0.003)	(0.001)	(0.001)	(0.002)
Establishment share	-0.035	-0.019	0.007	0.007	0.035
	(0.003)	(0.003)	(0.002)	(0.001)	(0.003)
Observations	321,967	321,967	321,967	321,967	321,967

Note: Task mentions, vacancies, word count, and salary are all log-transformed. “High exposure” refers to establishments above the 60th percentile of AI exposure, and “Medium exposure” refers to those between the 40th and 50th percentiles. Estimates are based on difference GMM (Arellano-Bond), instrumenting for lagged dependent variables. Establishment and industry fixed effects are differenced out and already controlled for in this framework. Industry refers to broad 2-digit NAICS codes.

Table 9: Effects of ChatGPT on Task Demand, 2020–24

	Task Demand, 2020–24				
	NRA	NRI	NRM	RC	RM
	A. High Exposure to ChatGPT				
Task Mentions	-0.088	-0.042	-0.020	0.013	-0.012
	(0.020)	(0.022)	(0.028)	(0.001)	(0.022)
Establishment share	-0.069	-0.028	0.004	0.013	0.065
	(0.005)	(0.005)	(0.003)	(0.001)	(0.005)
	B. Medium Exposure to ChatGPT				
Task Mentions	-0.135	-0.024	-0.111	0.009	-0.028
	(0.015)	(0.015)	(0.021)	(0.001)	(0.017)
Establishment share	-0.048	-0.034	0.005	0.009	0.052
	(0.004)	(0.004)	(0.002)	(0.001)	(0.004)
Observations	321,967	321,967	321,967	321,967	321,967

Note: Task mentions, vacancies, word count, and salary are all log-transformed. "High exposure" refers to establishments above the 90th percentile of AI exposure, and "Medium exposure" refers to those between the 10th and 90th percentiles. Estimates are based on difference GMM (Arellano-Bond), instrumenting for lagged dependent variables. Establishment and industry fixed effects are differenced out and already controlled for in this framework. Industry refers to broad 2-digit NAICS codes.

Table 10: Effects of ChatGPT on Task Demand, 2020–24

	Task Demand, 2020–24				
	NRA	NRI	NRM	RC	RM
	A. High Exposure to ChatGPT				
Task Mentions	-0.094	-0.020	-0.079	0.013	-0.012
	(0.014)	(0.015)	(0.020)	(0.016)	(0.016)
Establishment share	-0.059	-0.021	0.006	0.010	0.050
	(0.004)	(0.004)	(0.002)	(0.001)	(0.003)
	B. Medium Exposure to ChatGPT				
Task Mentions	-0.128	-0.016	-0.126	-0.038	-0.003
	(0.011)	(0.011)	(0.016)	(0.013)	(0.013)
Establishment share	-0.041	-0.031	0.005	0.009	0.047
	(0.003)	(0.003)	(0.002)	(0.001)	(0.003)
Observations	321,967	321,967	321,967	321,967	321,967

Note: Task mentions, vacancies, word count, and salary are all log-transformed. "High exposure" refers to establishments above the 80th percentile of AI exposure, and "Medium exposure" refers to those between the 20th and 80th percentiles. Estimates are based on difference GMM (Arellano-Bond), instrumenting for lagged dependent variables. Establishment and industry fixed effects are differenced out and already controlled for in this framework. Industry refers to broad 2-digit NAICS codes.

Table 11: Effects of ChatGPT on Task Demand, 2020–24

	Task Demand, 2020–24				
	NRA	NRI	NRM	RC	RM
	A. High Exposure to ChatGPT				
Task Mentions	-0.084	-0.020	-0.042	0.002	-0.013
	(0.013)	(0.015)	(0.018)	(0.015)	(0.014)
Establishment share	-0.053	-0.016	0.006	0.009	0.042
	(0.003)	(0.003)	(0.002)	(0.001)	(0.003)
	B. Medium Exposure to ChatGPT				
Task Mentions	-0.110	-0.010	-0.060	-0.044	0.012
	(0.010)	(0.011)	(0.016)	(0.012)	(0.012)
Establishment share	-0.037	-0.027	0.005	0.008	0.042
	(0.003)	(0.003)	(0.002)	(0.001)	(0.003)
Observations	321,967	321,967	321,967	321,967	321,967

Note: Task mentions, vacancies, word count, and salary are all log-transformed. “High exposure” refers to establishments above the 75th percentile of AI exposure, and “Medium exposure” refers to those between the 25th and 75th percentiles. Estimates are based on difference GMM (Arellano-Bond), instrumenting for lagged dependent variables. Establishment and industry fixed effects are differenced out and already controlled for in this framework. Industry refers to broad 2-digit NAICS codes.

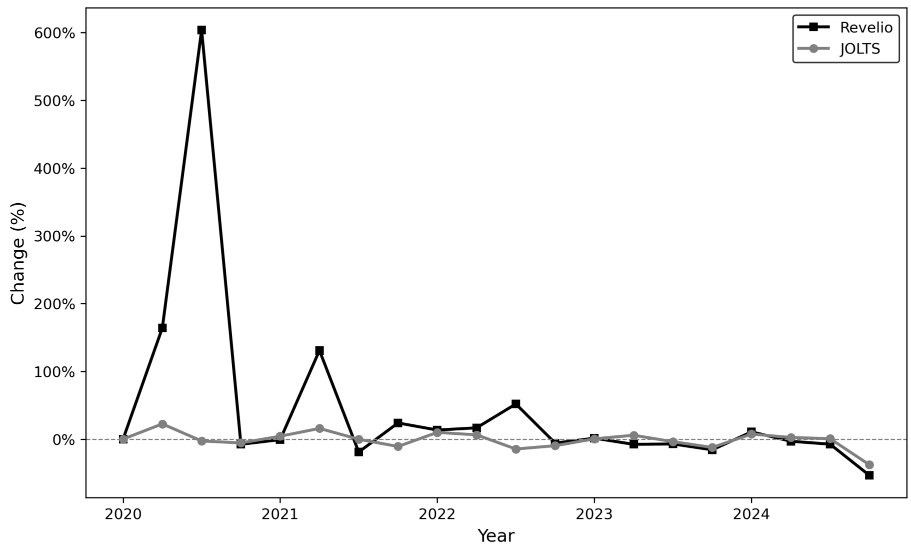
Table 12: Effects of ChatGPT on Task Demand, 2020–24

	Task Demand, 2020–24				
	NRA	NRI	NRM	RC	RM
	A. High Exposure to ChatGPT				
Task Mentions	-0.084	-0.015	-0.028	-0.024	0.007
	(0.011)	(0.012)	(0.016)	(0.013)	(0.013)
Establishment share	-0.052	-0.010	0.006	0.008	0.039
	(0.003)	(0.003)	(0.002)	(0.001)	(0.003)
	B. Medium Exposure to ChatGPT				
Task Mentions	-0.110	-0.011	-0.034	-0.048	0.018
	(0.010)	(0.011)	(0.015)	(0.012)	(0.012)
Establishment share	-0.039	-0.021	0.005	0.009	0.039
	(0.003)	(0.003)	(0.002)	(0.001)	(0.003)
Observations	321,967	321,967	321,967	321,967	321,967

Note: Task mentions, vacancies, word count, and salary are all log-transformed. “High exposure” refers to establishments above the 70th percentile of AI exposure, and “Medium exposure” refers to those between the 30th and 70th percentiles. Estimates are based on difference GMM (Arellano-Bond), instrumenting for lagged dependent variables. Establishment and industry fixed effects are differenced out and already controlled for in this framework. Industry refers to broad 2-digit NAICS codes.

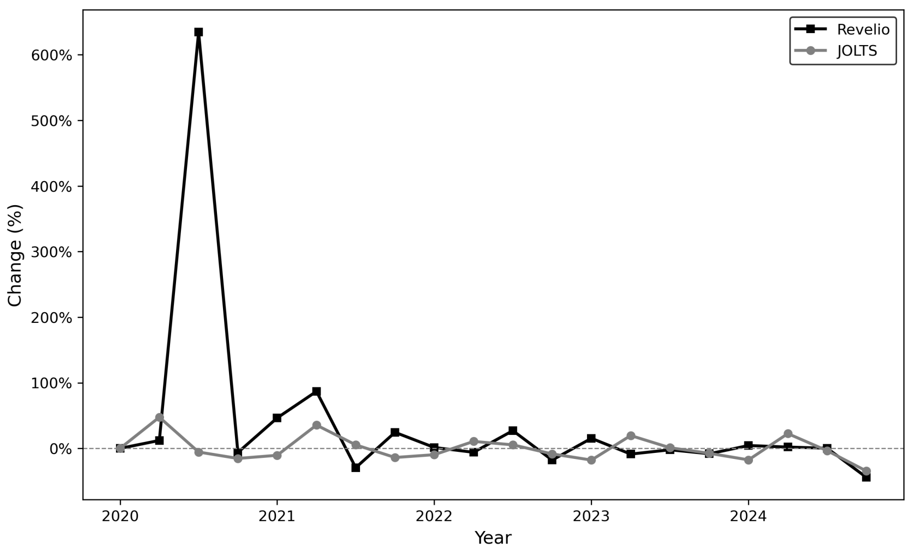
A3. Industry-Level Comparisons Between Revelio Job Postings and JOLTS Hires

Figure 8: Comparison of Revelio Postings and JOLTS Hires in Professional and Business Services



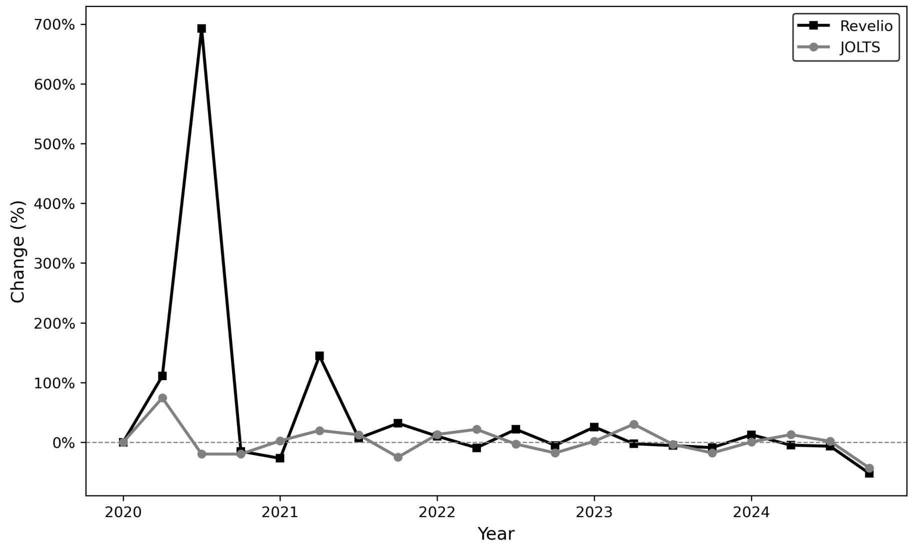
Note: Data are quarterly. JOLTS hires are seasonally unadjusted to maintain comparability with Revelio postings, which are not seasonally adjusted.

Figure 9: Comparison of Revelio Postings and JOLTS Hires in Trade, Transportation, and Utilities



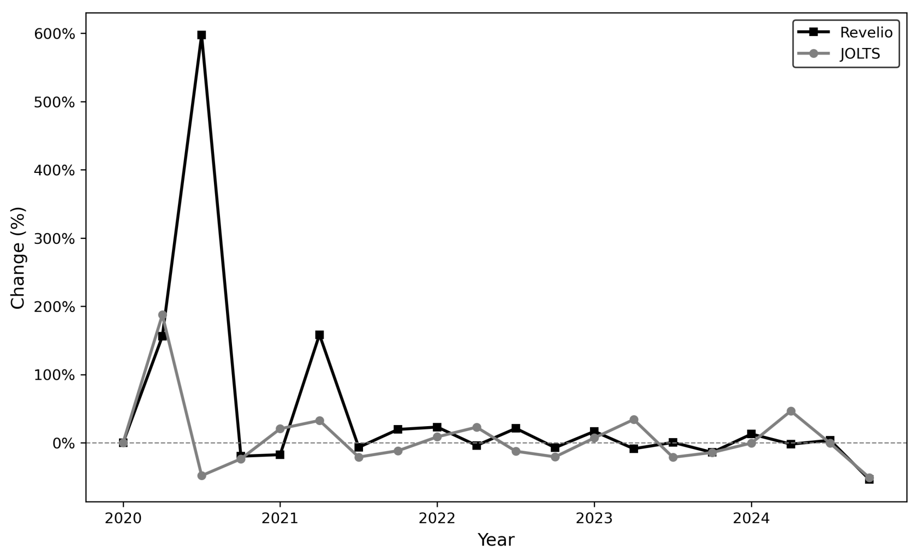
Note: Data are quarterly. JOLTS hires are seasonally unadjusted to maintain comparability with Revelio postings, which are not seasonally adjusted.

Figure 10: Comparison of Revelio Postings and JOLTS Hires in Private Education and Health Services



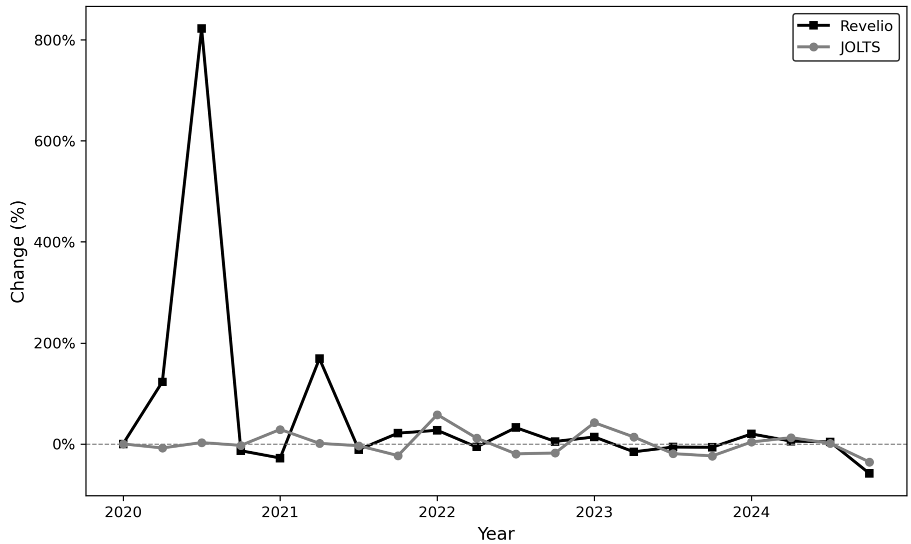
Note: Data are quarterly. JOLTS hires are seasonally unadjusted to maintain comparability with Revelio postings, which are not seasonally adjusted.

Figure 11: Comparison of Revelio Postings and JOLTS Hires in Other Services



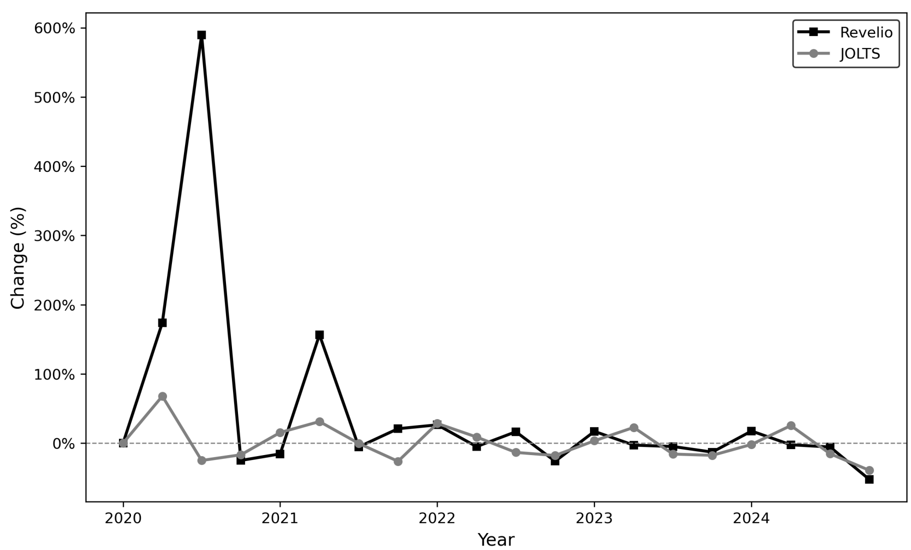
Note: Data are quarterly. JOLTS hires are seasonally unadjusted to maintain comparability with Revelio postings, which are not seasonally adjusted.

Figure 12: Comparison of Revelio Postings and JOLTS Hires in Mining and Logging



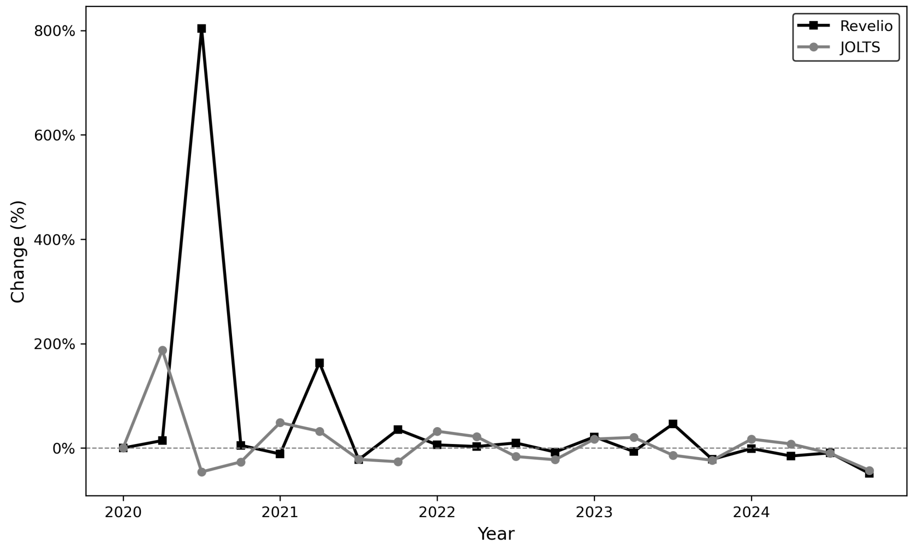
Note: Data are quarterly. JOLTS hires are seasonally unadjusted to maintain comparability with Revelio postings, which are not seasonally adjusted.

Figure 13: Comparison of Revelio Postings and JOLTS Hires in Manufacturing



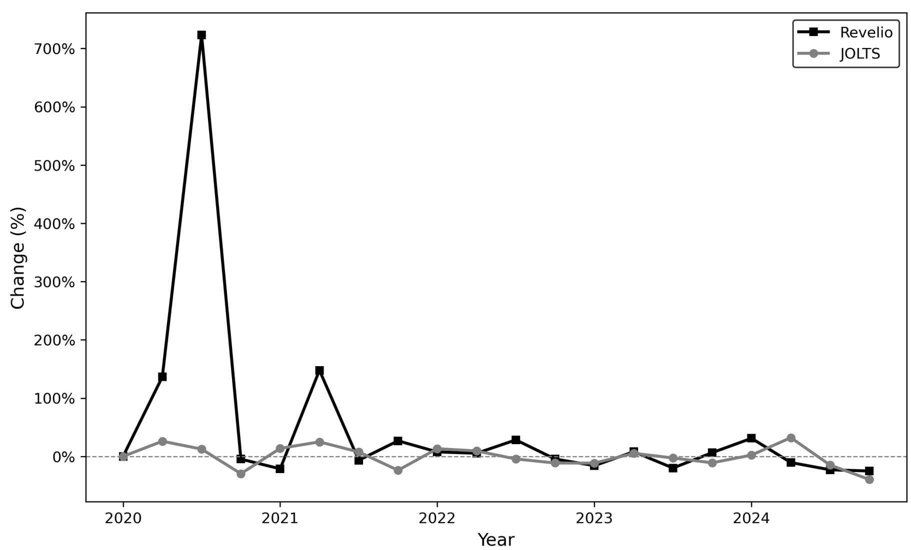
Note: Data are quarterly. JOLTS hires are seasonally unadjusted to maintain comparability with Revelio postings, which are not seasonally adjusted.

Figure 14: Comparison of Revelio Postings and JOLTS Hires in Leisure and Hospitality



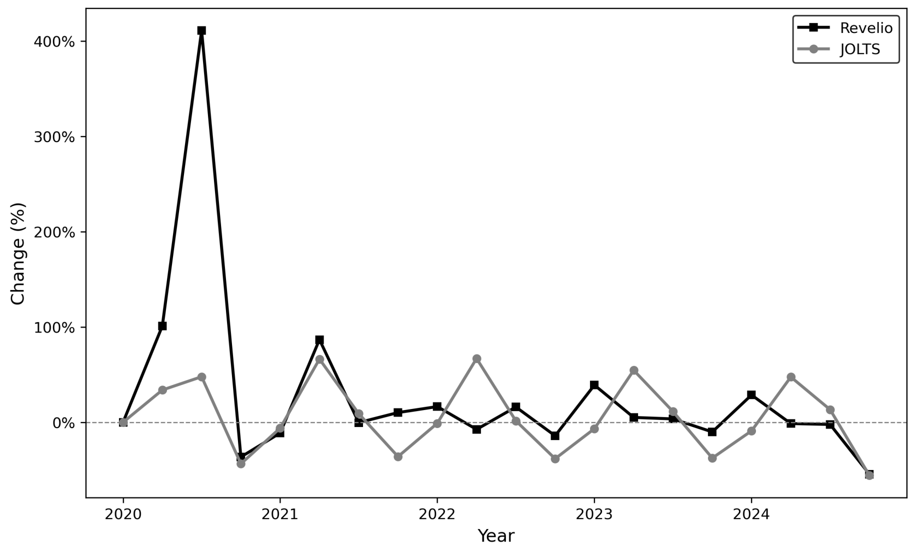
Note: Data are quarterly. JOLTS hires are seasonally unadjusted to maintain comparability with Revelio postings, which are not seasonally adjusted.

Figure 15: Comparison of Revelio Postings and JOLTS Hires in Information



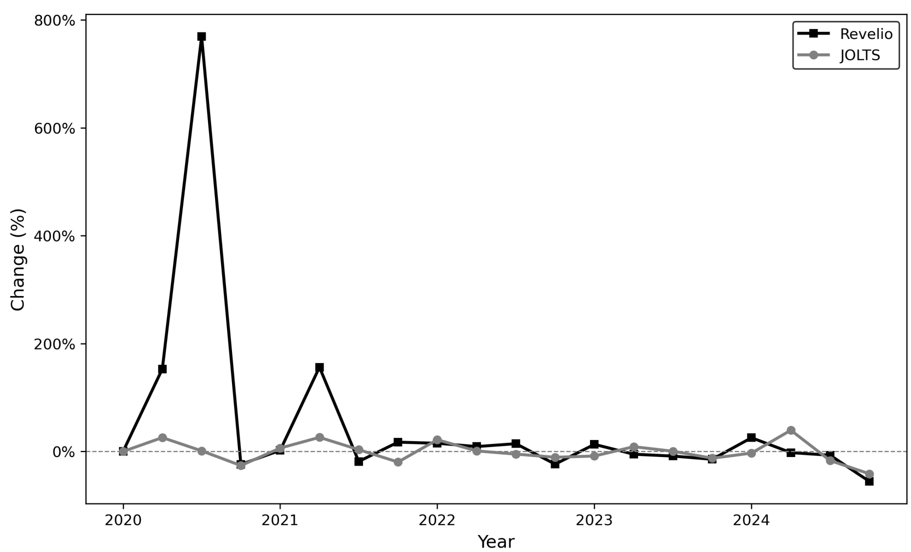
Note: Data are quarterly. JOLTS hires are seasonally unadjusted to maintain comparability with Revelio postings, which are not seasonally adjusted.

Figure 16: Comparison of Revelio Postings and JOLTS Hires in Government



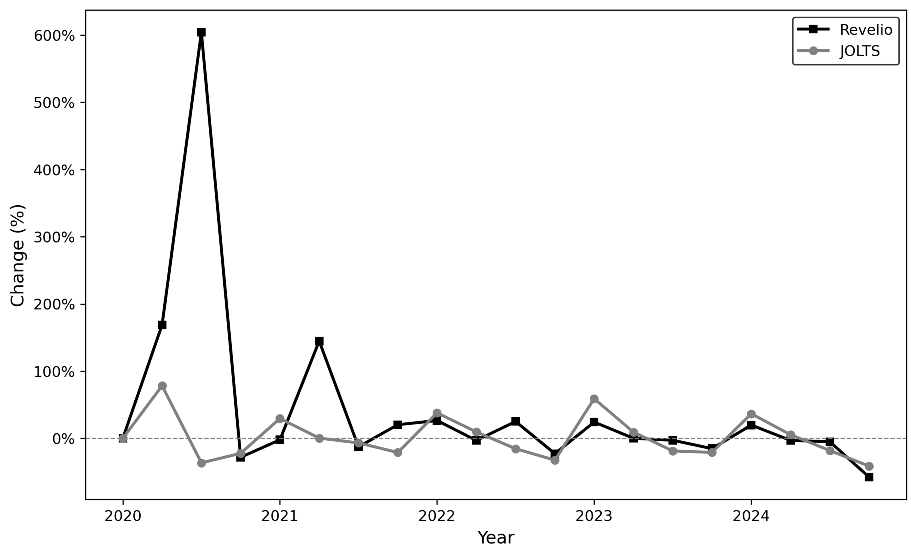
Note: Data are quarterly. JOLTS hires are seasonally unadjusted to maintain comparability with Revelio postings, which are not seasonally adjusted.

Figure 17: Comparison of Revelio Postings and JOLTS Hires in Financial Activities



Note: Data are quarterly. JOLTS hires are seasonally unadjusted to maintain comparability with Revelio postings, which are not seasonally adjusted.

Figure 18: Comparison of Revelio Postings and JOLTS Hires in Construction



Note: Data are quarterly. JOLTS hires are seasonally unadjusted to maintain comparability with Revelio postings, which are not seasonally adjusted.

A4. Conditional Error Correlation

An important question for the model specification is whether the conditional error terms from the regressions are correlated across task types, which would justify the use of a Seemingly Unrelated Regression (SUR) model to account for potential efficiency gains from jointly estimating the system. Additionally, it is of interest to assess whether the errors exhibit serial correlation within the same task type across time, which could suggest the presence of autocorrelation in the residuals. Identifying cross-equation correlation and serial correlation would help refine the modeling approach and ensure that standard errors are appropriately adjusted to account for these dependencies.

To assess serial correlation in the idiosyncratic error terms, I first estimated the following fixed-effects panel regression for each task type:

$$\ln(\text{Task}_{it}) = \beta_1 \ln(\text{Vacancies}_{it}) + \beta_2 \ln(\text{WordCount}_{it}) + \gamma_t + \alpha_i + u_{it} \tag{2}$$

where $\ln(\text{Task}_{it})$ is the log of task mentions for establishment i at time t, $\ln(\text{Vacancies}_{it})$ and $\ln(\text{WordCount}_{it})$ are control variables for establishment vacancies and job description length, γ_t represents time fixed effects, α_i captures establishment fixed effects, and u_{it} is the idiosyncratic error term.

After estimating each regression, I generated the residuals (u_{it}) and their first lags ($u_{i,t-1}$) for each task type. The table below reports the correlation coefficients between the residuals and their lags across all task types.

Table 13: Correlation Matrix of Residuals and Their Lags

	AR	L.AR	NM	L.NM	IR	L.IR	RC	L.RC	RM	L.RM
AR	1.000	0.396	0.097	0.054	0.256	0.088	0.127	0.050	0.102	0.022
L.AR	0.396	1.000	0.053	0.096	0.259	0.260	0.062	0.128	0.022	0.102
NM	0.097	0.053	1.000	0.096	0.097	0.058	0.144	0.088	0.184	0.078
L.NM	0.054	0.096	0.096	1.000	0.091	0.098	0.096	0.146	0.062	0.187
IR	0.256	0.259	0.097	0.091	1.000	0.364	0.108	0.052	0.160	0.086
L.IR	0.088	0.260	0.058	0.098	0.364	1.000	0.053	0.110	0.086	0.164
RC	0.127	0.062	0.144	0.096	0.108	0.053	1.000	0.404	0.119	0.049
L.RC	0.050	0.128	0.088	0.146	0.052	0.110	0.404	1.000	0.059	0.119
RM	0.102	0.022	0.184	0.062	0.160	0.086	0.119	0.059	1.000	0.388
L.RM	0.022	0.102	0.078	0.187	0.086	0.164	0.049	0.119	0.388	1.000

Notes: AR = Analytical Residual, L.AR = Lagged Analytical Residual,
NM = Non-routine Manual Residual, L.NM = Lagged Non-routine Manual Residual,
IR = Interactive Residual, L.IR = Lagged Interactive Residual,
RC = Routine Cognitive Residual, L.RC = Lagged Routine Cognitive Residual,
RM = Routine Manual Residual, L.RM = Lagged Routine Manual Residual.

The correlation matrix of residuals and their lags indicates that the error terms do not appear to be strongly correlated across task types, which is reassuring for the validity of the separate regressions. Most cross-task correlations are relatively small, suggesting limited dependence between the residuals of different task types. However, there is clear evidence of serial correlation within task types, as each residual is moderately correlated with its own one-period lag. For example, the correlation between the analytical residual and its lag is approximately 0.40, while similar patterns appear for the other task types. This serial correlation implies that residuals within task types are persistent over time, which suggests that future specifications should account for autocorrelation to ensure valid inference.

A5. Pre-Trend Tests for Parallel Trends

Table 14: F-Test p-Values for Pre-Trend Tests

Outcome Variable	High Exposure	Medium Exposure
Log of Task Totals		
NRA	< 0.001	< 0.001
NRM	0.003	< 0.001
NRI	< 0.001	< 0.001
RC	< 0.001	< 0.001
RM	0.015	0.012
Task Shares		
NRA	0.060	< 0.001
NRM	< 0.001	< 0.001
NRI	0.285	< 0.001
RC	0.417	< 0.001
RM	0.736	0.119

Notes: p-values are from joint F-tests for pre-trend balance. For each outcome, I estimate a dynamic difference-in-differences model as in Equation 1, excluding the post-treatment term and testing the joint significance of the pre-treatment interaction coefficients for the high and medium exposure groups, respectively.

References

Brynjolfsson, E., Li, D., and Raymond, L. R. (2023). Generative AI at Work.

Choi, J. H., Monahan, A., and Schwarcz, D. (2023). Lawyering in the Age of Artificial Intelligence.

Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., and Lakhani, K. R. (2023). Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality. *SSRN Electronic Journal*. Publisher: Elsevier BV.

Eloundou, T., Manning, S., Mishkin, P., and Rock, D. (2024). GPTs are GPTs: Labor market impact potential of LLMs. *Science*, 384(6702):1306–1308. Publisher: American Association for the Advancement of Science.

Felten, E. W., Raj, M., and Seamans, R. (2023). Occupational Heterogeneity in Exposure to Generative AI.

Noy, S. and Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654):187–192. Publisher: American Association for the Advancement of Science.

Peng, S., Kalliamvakou, E., Cihon, P., and Demirer, M. (2023). The Impact of AI on Developer Productivity: Evidence from GitHub Copilot. arXiv:2302.06590 [cs].