Journal of Economic Behavior and Organization xxx (xxxx) xxx

Contents lists available at ScienceDirect

Journal of Economic Behavior and Organization

journal homepage: www.elsevier.com/locate/jebo

Winners and losers of generative AI: Early Evidence of Shifts in Freelancer Demand

Ole Teutloff [a,b], Johanna Einsiedler [a], Otto Kässi [c,d], Fabian Braesemann [b,d,e], Pamela Mishkin [f,1], R. Maria del Rio-Chanona [g,h,i,*]

[a] Copenhagen Center for Social Data Science, University of Copenhagen, Copenhagen, Denmark
[b] DWG Datenwissenschaftliche Gesellschaft Berlin, Berlin, Germany
[c] Etla Economic Research Helsinki, Helsinki, Finland
[d] Oxford Internet Institute, University of Oxford, Oxford, United Kingdom
[e] Einstein Center Digital Future, Berlin, Germany
[f] Independent, San Francisco, United States
[g] Computer Science Department, University College London, London, United Kingdom
[h] Bennett Institute for Public Policy, University of Cambridge, Cambridge, United Kingdom
[i] Complexity Science Hub Vienna, Vienna, Austria

ARTICLE INFO

JEL classification:
C 21
C 55
J 21
J 23
J 24
O 33

Keywords:
Generative AI technologies
Large language models
Automation and employment
Labor market implications of AI
Technological transition
Online labor markets

ABSTRACT

We examine how ChatGPT has changed the demand for freelancers in jobs where generative AI tools can act as substitutes or complements to human labor. Using BERTopic we partition job postings from a leading online freelancing platform into 116 fine-grained skill clusters and with GPT-4o we classify them as substitutable, complementary or unaffected by LLMs. Our analysis reveals that labor demand increased after the launch of ChatGPT, but only in skill clusters that were complementary to or unaffected by the AI tool. In contrast, demand for substitutable skills, such as writing and translation, decreased by 20–50% relative to the counterfactual trend, with the sharpest decline observed for short-term (1-3 week) jobs. Within complementary skill clusters, the results are mixed: demand for machine learning programming grew by 24%, and demand for AI-powered chatbot development nearly tripled, while demand for novice workers declined in general. This result suggests a shift toward more specialized expertise for freelancers rather than uniform growth across all complementary areas.

1. Introduction

The impact of Large Language Models (LLMs) on labor markets remains an open question, with early evidence suggesting a wide range of outcomes. On one hand, LLMs have been shown to reduce demand for certain tasks like writing and translation, where they can substitute human labor (Demirci et al., 2023; Qiao et al., 2023; Hui et al., 2023; Liu et al., 2023). On the other hand, laboratory experiments have shown that LLMs can be used as complementary tools that enhance productivity, particularly for less experienced workers (Brynjolfsson and McAfee, 2014; Noy and Zhang, 2023). These mixed results align with the view of Autor (2024), that the effects of new technology on work depend heavily on how it is deployed by firms, governments, and other stakeholders. To better understand the impact of generative AI on work and employment, we examine the effects of the ChatGPT launch on labor demand

* Corresponding author at: Computer Science Department, University College London, London, United Kingdom.
E-mail address: m.delriochanona@ucl.ac.uk (R.M. del Rio-Chanona).

[1] This paper reflects the author's individual research and recommendations and was not written in her official capacity at OpenAI, United States.

https://doi.org/10.1016/j.jebo.2024.106845
Received 15 February 2024; Received in revised form 25 October 2024; Accepted 25 November 2024

across fine-grained skill clusters that can be substituted or complemented by LLMs.

While LLMs and Generative AI (Gen AI) are fairly new, there is extensive literature estimating the exposure to automation – i.e., the technical feasibility of occupations being replaced by AI – using O*NET task data on occupations and patent data. Frey and Osborne (2017) estimated that half of the US labor force could be in occupations at risk of automation. Arntz et al. (2016) offered a nuanced view of these findings through a task-level analysis and concluded that less than 10% of the jobs in OECD countries were at high risk of automation. In 2017, Brynjolfsson and Mitchell (2017) argued that AI would most likely redefine work within occupations. Webb (2019) used Natural Language Processing (NLP) to study occupation task descriptions and patent documents, suggesting that high-skill tasks are more amenable to AI replacement, potentially decreasing wage inequality between the highest and lowest earners without impacting the top 1% of earners. Albanesi et al. (2023) highlighted a positive association between AI exposure scores and employment growth across 16 European countries from 2011 to 2019.

Several empirical studies have explored the impact of recent technological innovations on the labor market. Autor et al. (2024) found that in recent decades, labor demand-eroding effects of technologies intensified while demand-increasing effects did not. Using online platform data, Stephany and Teutloff (2024) documented that workers with machine learning and AI skills can earn a premium compared to workers who have otherwise similar skill sets but lack AI proficiency. For online freelancers, Lysyakov and Viswanathan (2023) found that in response to the threat of image-generating AI, designers on an online crowd-sourcing platform split into two groups: low-tier designers exited the market, while higher-capability designers moved away from direct competition with AI by focusing on more complex tasks.

Regarding LLMs, a small but rapidly growing literature examines their labor market impacts. Eloundou et al. (2023) estimated that approximately 80% of the U.S. workforce could see at least 10% of their job tasks influenced by the deployment of (LLMs), with around 19% of employees potentially experiencing an impact on at least half of their tasks. Noy and Zhang (2023) conducted an experiment and found that ChatGPT increased the productivity of mid-level professional writing tasks. Similarly, Brynjolfsson et al. (2023) found that access to a generative AI tool led to a 14% increase in productivity, as indicated by the number of issues resolved per hour. Novice and low-skilled workers experienced the highest increase in productivity, while the impact on seasoned, skilled workers was negligible. Experimental evidence also showed that LLM-based code writing assistants such as GitHub Copilot increased developer speed by over 50% (Peng et al., 2023), with less experienced programmers benefiting the most.

Outside laboratory settings, initial evidence shows the short-term effects of LLMs on labor markets from online labor platforms. On the demand side, generative AI significantly reduces the number of freelance projects in automation-prone occupations (Demirci et al., 2023). Demirci et al. (2023) used a skill co-occurrence network approach to create 15 clusters of jobs that fall within three broad groups: manual-intensive, writing automation-prone, and image-generation jobs. They find that writing automation-prone and image-generation jobs decrease compared to manual-intensive jobs. On the supply side, effects vary between workers: those in occupations highly exposed to generative AI substitution suffer reductions in earnings and employment opportunities (Hui et al., 2023; Liu et al., 2023). However, while top-performing freelancers might be disproportionately affected by generative AI (Hui et al., 2023), those who integrate generative AI into their work attract more jobs (Liu et al., 2023). Freelancers in web development whose jobs are complemented by AI may increase their earnings, while translators face fewer opportunities and lower wages (Qiao et al., 2023).

Both experimental studies and those based on online labor platforms offer initial insights into how LLMs may affect the labor market. However, these studies have either focused on controlled laboratory settings, which may not translate to complex labor markets, or on a limited group of broad skills, potentially obscuring heterogeneous effects. Moreover, online labor platform studies have primarily examined substitution effects, while LLMs often serve complementary roles, enhancing certain tasks without fully replacing entire jobs (Eloundou et al., 2023). Given that the effects of new technologies depend significantly on their implementation and the specific skills they impact (Autor, 2024), we need a deeper understanding of the impact across skill heterogeneity and to distinguish between substitutable and complementary skills. However, conducting such a study poses two main challenges. First, skill categories on online labor markets evolve over time, complicating the use of co-occurrence networks at a fine-grained level. Second, it is manually infeasible to categorize the impact of LLMs on millions of job postings into substitutable, complementary, or unaffected categories.

NLP tools, which are frequently used in economic research, can help us overcome these challenges. As Ash and Hansen (2023) pointed out, text algorithms can solve several common problems in applied economics, including document similarity measurement and concept detection. For example, Hoberg and Phillips (2016) quantifies the relatedness of firms based on the similarity of their product offer descriptions. Hansen et al. (2023) and Adams-Prassl et al. (2020) use a language-processing framework to determine whether job postings refer to hybrid or fully remote work. del Rio-Chanona et al. (2023) use topic modeling to study how the prevalence of topics across the work discourse evolved during the Covid-19 pandemic and the Great Resignation. Most recently, LLMs have been shown to perform near the human level on a range of qualitative data analysis tasks, such as e. g. text annotation, interview analysis, or automated lexicography (Karjus, 2023). Following this line of work, Eloundou et al. (2023) used ChatGPT to quantify the extent to which occupations' work activities can be performed by LLMs.

To tackle the first challenge of evolving skill categories, we use a topic modeling approach to group job postings into skill clusters with similar semantic meanings. This allows us to include more than 3 million job postings into our analysis and identify 116 distinct skill clusters. To overcome the second challenge of assigning jobs AI exposure labels, we build upon recent LLM prompt engineering research (Eloundou et al., 2023; OpenAI, 2023; Wei et al., 2022) to classify the exposure of the skill clusters to LLMs into substitutable, complementary or unaffected groups. We identify 71 fine-grained skill clusters to be substitutable or complementary, and 45 to be unaffected. This classification allows us to move beyond existing research that manually defines writing-related services to be the main job types affected by the launch of the ChatGPT (Hui et al., 2023; Qiao et al., 2023; Liu et al., 2023) and instead

allows us to measure the impact across a broader range of substitutable and complementary skill clusters. Finally, we employ a difference-in-differences econometric model using the unaffected clusters as a control group to quantify changes in demand across substitutable and complementary skill clusters after the release of ChatGPT.

Our results show that while the demand for unaffected and complementary clusters increased after the introduction of ChatGPT, the number of job postings in the substitutable clusters decreased by 7%. The difference-in-differences analysis reveals that, compared to the unaffected clusters, substitutable clusters experienced a 25% decline, while complementary clusters as a whole showed no significant effects. When distinguishing projects by their duration, we find that the decrease in substitutable job postings was concentrated in short-term (1–3 weeks) jobs. In terms of worker experience, heterogeneity across substitutable clusters is limited. However, we find a decline in job postings for novice workers in complementary skill clusters.

At a cluster level, those related to writing for Real Estate and 'About Us' pages showed the most substantial decreases, at 52% and 59%, respectively. Translation jobs also saw significant impacts, with the 'Western European Languages' category experiencing a 23% decline. While complementary clusters did not show significant aggregate effects, demand changes were mixed within these clusters. Notably, the demand for jobs in the 'AI-powered chatbots' cluster nearly tripled, with a 179% increase, and the 'Machine Learning' cluster saw a significant rise, with demand increasing by 24%.

Our findings present a nuanced view of the impact of ChatGPT on labor demand. While we observe similar trends to those reported by Demirci et al. (2023), Hui et al. (2023), Liu et al. (2023), Qiao et al. (2023) regarding substitutable tasks, our analysis indicates that these decreases are primarily concentrated in short-term gigs. Moreover, although there is no aggregate decline in complementary skills, we find a reduced demand for novice workers in these clusters. This may appear to contradict the findings of Noy and Zhang (2023) and Brynjolfsson et al. (2023), who found that LLM tools boost the productivity of novice workers. However, this may result from the complex dynamics of the labor market, where increased productivity of novice workers within firms may reduce the need for novice freelancers. Thus, while LLMs usage enhance the efficiency of in-house novice workers, it could diminish external demand for similar freelance skills. This suggests a shift toward specialized expertise for freelancer demand rather than uniform growth across all complementary areas. Our results also show that while some complementary skill clusters experienced significant declines, others, particularly those related to programming and software development, saw substantial increases.

Overall, our analysis underscores the importance of considering the specific nature of tasks and the context of AI implementation when assessing its impact on labor markets. The heterogeneity in outcomes suggests that integrating AI into work processes can create both winners and losers, depending on how AI technologies are utilized and the specific skill sets involved.

2. Data and methods

2.1. Research design

To measure the impact of LLMs on labor demand at a granular skill level, we distinguish between skill clusters that are substitutable by, complementary to, or unaffected by LLMs. We then analyze how demand for these skills changes in an online labor platform around the time of the public release of ChatGPT on November 30, 2022.

Our analysis focuses on data from a leading online labor platform, where job postings list detailed skill requirements. We use the transformer model BERTopic (Grootendorst, 2022) to group job postings into clusters based on their skill requirements. To assess the robustness of our results, we combine hierarchical clustering with a manual inspection of all skill clusters, ensuring that job postings with similar skill requirements are clustered together. This procedure results in 116 skill clusters. The methodology of inferring skill clusters from the data is illustrated in Fig. 1.

Following LLM exposure definitions from previous literature (Qiao et al., 2023; Eloundou et al., 2023) and through prompt engineering techniques, we design prompts for labeling each skill cluster into substitutable, complementary, and unaffected with GPT-4o. We manually inspect each cluster label by examining the reasoning provided by GPT and the label coherence within hierarchical clusters. In the cases where human and GPT labels diverge, we give priority to human labels. This results in 12 skill clusters in the substitutable category, 59 in the complementary, and 45 in the unaffected category.

In the last step, similar to other studies assessing the causal impact of LLMs on the labor market (del Rio-Chanona et al., 2024; Demirci et al., 2023; Qiao et al., 2023; Hui et al., 2023), we use a difference-in-differences (DiD) model to estimate changes in demand of projects in the online labor market across treatment groups (substitution and complementarity) in comparison to the control group (unaffected).

2.2. Data

We use data from one prominent global online freelancing platform, which wished to remain anonymous. Online freelancing platforms are digital marketplaces that connect buyers and sellers of remotely deliverable work. Prominent examples are *Upwork*, *Fiverr*, or *Freelancer*. Jobs on these platforms cover a wide range of professional domains, including writing, translation, software development, design, marketing, or legal services. Employers range from individuals to startups and Fortune 500 companies (Corporaal and Lehdonvirta, 2017). When posting a job, employers specify detailed skill requirements, including a mix of broad skills such as *writing*, *graphic design* or *translation* and specific skills such as *Microsoft Word*, *Adobe Photoshop* or *Chinese English translation*. On average, projects require a median of six skills. In addition, every job posting contains a title and a description specifying the task. For the purpose of identifying skill clusters, we only consider the skill tags, not the job title or the text of the job description.

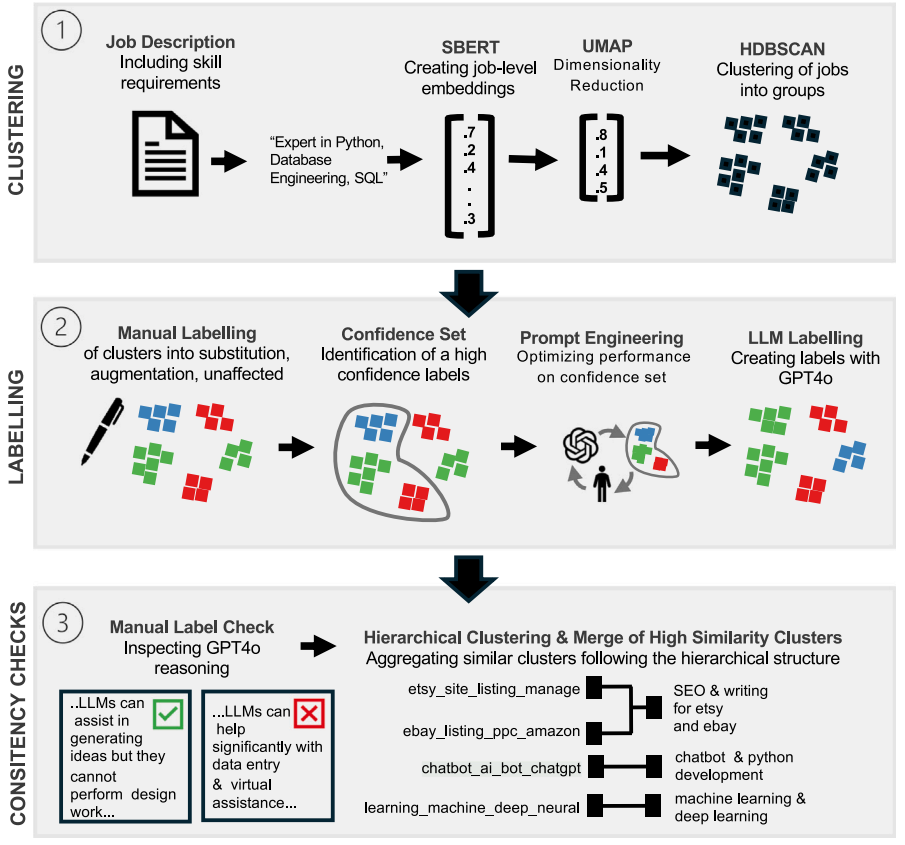


Fig. 1. Methodological overview: From job postings to AI exposure label. Our approach for creating skill clusters from job postings and assigning them to a treatment group consists of three steps. First, we use the large language model SBERT to embed the skill requirements of each job posting, creating numerical high-dimensional job-level embeddings. Through UMAP, we reduce the dimensionality and then apply the HDBSCAN clustering algorithm. Second, we manually label the resulting skill clusters into substitutable, complementary, and unaffected to create a high-confidence set that serves as a benchmark for prompt engineering GPT-4o. The best prompt is used to label all skill clusters. Third, we use GPT-4o's reasoning and the hierarchical structure of the skill clusters to verify the label consistency. Merging similar skill clusters ensures stable observations over time that are robust to the platform renaming skills.

When posting a job, the employer is initially asked to provide a title. Based on this job title, the platform automatically generates a set of skill requirements. Employers have the flexibility to edit this list of skills as they see fit. The platform curates the pool of available skills, adding new ones based on job postings and ongoing discussions with key clients (for instance, when Python2 was replaced by Python3, the relevant skill tags were updated for the new job postings. The old ones still show Python2.). The skill clusters are built in a way that the number of clusters does not change during the analysis period.

The granular information on skill demand and the fast-paced nature of this digital labor market make it an ideal setting to study the labor market implications of LLMs because changes in demand for specific skills should manifest much faster than in traditional labor market settings, making it an ideal real-world "lab" (Horton and Tambe, 2015) for observing the effects of LLMs.

Data for this study was gathered through the Online Labour Index project, as detailed by Kässi and Lehdonvirta (2018), which has been monitoring daily new job listings through the platform's API since January 2017. Further data collection on job specifics took place in multiple phases from November 2021 to August 2023. The freelancing platform, while based in the United States, operates globally, drawing a significant number of workers from countries such as India, Pakistan, the Philippines, and East Europe, as noted by Stephany et al. (2021).

We have collected information on the daily new job postings between January 2021 and September 2023, covering several million job postings. In a few cases, there were gaps in the data collection due to various software errors. These are visible as sudden "blips" in the time series of postings. Reassuringly for us, the data collection worked without errors from May 2022 to September 2023. We also demonstrate that our results remain consistent even when controlling for these fluctuations using week dummies or when excluding the few problematic time periods from our data.

Besides detailed information on the skills required for the job posting, our data includes various other metadata. In particular, we use two variables as control variables when studying effect heterogeneity: expected project duration and worker experience level. When creating a job posting, the employer is asked to provide the expected duration of the job. The duration can be one of the following: under 3 weeks, 3–9 weeks, 9–18 weeks, or 18–52 weeks. This information is not binding for either the employer or the worker; it merely reflects the employer's expectation about the contract length.

Additionally, the employer can specify the experience level of workers they are looking for, choosing from three options: "I am willing to hire an inexperienced freelancer cheaply" (novice), "I prefer an intermediate level freelancer" (intermediate), or "I am willing to pay more for an experienced freelancer" (veteran). This information is presented to workers when they apply for jobs. However, it is important to note that the specified worker level is merely a statement of employer expectations. Novice workers can still apply for and potentially be hired for jobs aimed at higher levels.

The job posting metadata also includes other potential control variables, such as expected hours per week, and the desired worker's home country. These variables were not useful as control variables due to a large share of missing values.

Finally, when creating a job posting, the employers can also declare their expected project budget. We use the employer declared budget as a proxy measure for employers' willingness to pay for labor.

2.3. From job postings to skill-clusters with BERTopic modeling

We use BERTopic to form and name clusters of job postings with similar skill requirements. This process is done in five steps, which we explain and justify below.

Embeddings. We give a text description to each job posting by appending the pre-text 'Expert in' to its list of skills, which are lower-cased and unhyphenated. For instance, a job posting with skills 'ghostwriting', 'writing', 'ebook', 'creative writing', and 'english' becomes 'Expert in ghostwriting, writing, ebook, creative writing, english'. We then embed these text descriptions into a 384-dimensional space using the sentence transformer "all-MiniLM-L6-v2", which is recommended for short texts.

We prefer this approach over skill co-occurrence since skills are not constant over time. For example, while 'news writing style' and 'news writing' refer to the same or very similar skill, 'news writing style' appears as a skill between 2021–2022 but not in 2023, while 'news writing' is a skill present in 2022–2023, but not before. Network co-occurrence algorithms, which are commonly used to model skill networks (Anderson, 2017; Lukac, 2021; Demirci et al., 2023; Stephany and Teutloff, 2024), may struggle to see these two skills as overlapping. Transformer models like BERTopic are able to capture the overlap due to semantic similarity. In addition, compared to bag-of-words embedding approaches such as GloVE or Word2Vec (Pennington et al., 2014; Mikolov et al., 2013), transformers have the advantage of not being restricted to a predefined dictionary of words.

Dimensionality reduction The next step is to reduce the dimensionality of the embeddings using Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) (McInnes et al., 2018). This step is crucial because, in high-dimensional spaces, data points tend to become equidistant from each other, making it difficult for clustering algorithms to identify meaningful groupings (Aggarwal et al., 2001). Moreover, high-dimensional representations often contain noise or irrelevant features, which dimensionality reduction helps filter out. By focusing on the most important features, dimensionality reduction techniques enhance the quality of the clusters (Van Der Maaten et al., 2009).

Clustering We then cluster projects using Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN). We adjust the minimum cluster size for HDBSCAN to 1000, as the BERTopic documentation recommends trying different values and setting parameters larger than 500 for datasets exceeding one million entries (Grootendorst, 2024b). We further use an outlier reduction function to decrease the number of uncategorized topics (Grootendorst, 2024a; McInnes and Healy, 2017).

We use HDBSCAN instead of K-means because K-means assumes clusters to be spherical and of roughly equal size. HDBSCAN, however, utilizes density within the embeddings, accommodating clusters of varying shapes (Bhattacharjee and Mitra, 2021). This approach is generally preferred for complex data distributions, such as text embeddings, and is recommended by the BERTopic documentation (Grootendorst, 2022). HDBSCAN also provides a hierarchy of clusters, allowing us to check cluster coherence manually.

This step resulted in 286 clusters. We manually checked the common skills and several job descriptions in each cluster and concluded that, at least intuitively, the clustering seems to produce meaningful results. Since BERTopic is stochastic in nature, the results depend on the initial, random allocation of topics to clusters. To test the robustness of our clusters, we ran the algorithm multiple times and also tried K-means clustering instead of HDBSCAN. The robustness checks are detailed in Appendix E.

Vectorization The clustering procedures end in the previous step. However, BERTopic facilitates giving intuitive names to each cluster with two more steps. These two steps do not affect the categorization of clusters into substitutable, complementary, or unaffected (see next section) and consequently do not affect our difference-in-difference estimation. Nonetheless, the names help interpret our results. The vectorization step converts text in clusters into numerical vectors, which creates a document-term matrix where each entry represents the frequency of a token in a document.

cTF-IDF Finally, to label each cluster, the most representative words are chosen via continuous term frequency-inverse document frequency (cTF-IDF). This process balances out frequent terms and removes those that are too frequent, ensuring that the most informative terms for each cluster are highlighted. This helps in assigning meaningful labels to the skill clusters.

In Appendix A, we provide a comparison of our clusters to the categories used by the online freelance platforms and additional arguments for our choice of clustering methods and robustness checks.

2.4. From skill clusters to AI exposure labels with GPT-4o

We apply the definitions from Eloundou et al. (2023) and Qiao et al. (2023) to group skill clusters into the following three categories of AI exposure:

Substitutable: A non-expert person can use Large Language Models to complete these tasks with a similar level of quality without the help of an expert.

Complementary: A non-expert cannot use Large Language Models to complete these tasks without significantly compromising the quality of execution. However, an expert freelancer can use Large Language models to reduce the time it takes them to complete these tasks by at least half and without significantly compromising quality.

Unaffected: While non-experts and experts may be able to use Large Language Models to help in these tasks, Large Language Models can only provide limited help. In other words, having access to a large language model will not reduce the time it takes an expert freelancer to complete the job by more than half.

These represent *ex-ante* definitions of the substitution, complemented and unaffected categories. While we conduct the manual and GPT-based labeling in 2024, we only consider LLM capabilities that were available by the time ChatGPT was released in November 2022 (for more details and showcase examples, see appendices B and C.

Manual labeling. To reliably classify skills clusters into these three groups, we combine several steps of manual analysis and GPT prompting as outlined in Fig. 1 sections two and three. First, two authors independently label all skill clusters. Reviewing and discussing the labels allows us to create a high-confidence set of 55 skill clusters for which we are confident about the AI exposure labels. This set includes 17 clusters in substitution, 14 in complementarity, and 24 in unaffected. However, since many of the skill clusters consist of highly specialized skills, reliable manual labeling would require expert knowledge in a variety of different occupational fields. Therefore, we rely on GPT-4o to extend our labels to the entire set of skill clusters using the high confidence set to optimize GPT prompts and benchmark the results.

Prompt engineering. To develop a prompt for GPT-4o that produces high-quality AI exposure labels for all skill clusters, we employ several validated prompt engineering best practices with the high confidence set serving as our *test set* on which to evaluate GPT output. As demonstrated by Brown et al. (2020), LLMs have been found to possess 'few-shot properties', i. e. their performance across most tasks increases when the prompt includes several demonstrative examples. Additionally, so-called 'chain-of-thought' prompting, whereby the user also gives examples of the logical steps that should lead to the correct answer, has been shown to significantly improve the complex reasoning capabilities of LLMs (Wei et al., 2022). We consider these insights by including an example for each category and a precise description of the reasoning behind the classification to the prompt. Further, as suggested by OpenAI (2023), we format our prompt such that the instructions are at the beginning and are separated through quotation marks from the context part. We further explicitly articulate our desired output format (JSON).

For each cluster, our input in the prompt includes the top ten skills as identified by cTF-IDF from BERTopic and the cluster's most frequent skills. We incorporate don't solely rely on the skills identified by cTF-IDF since those may not include relevant skills that are common across many clusters. The final best-performing prompt achieves 93 percent accuracy on the high confidence set. The full prompt text and examples for GPT output can be found in Appendix B. Using this final prompt to generate labels for all skill clusters results in 31 clusters assigned to substitution, 164 to complementarity, and 91 to unaffected.

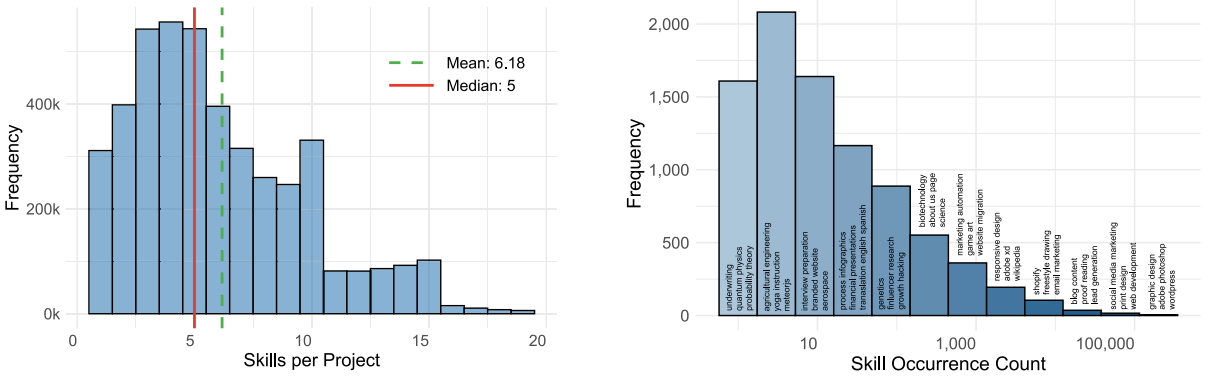
Consistency checks. To ensure consistent, high-quality labels, we manually review all GPT labels and the corresponding reasoning. Moreover, we use hierarchical clustering to aggregate the original 286 skill clusters into higher-level groups. Label coherence within the higher-level skill clusters corroborates the validity of the GPT labels. We change labels where the GPT reasoning is unconvincing and labels are not aligned within higher-level hierarchical skill clusters. Lastly, we manually merge small skill clusters following the hierarchical structure. While high granularity is beneficial when assigning AI exposure labels, small skill clusters are, at times, not stable over time. The online freelancing platform seems to periodically introduce new names for skills while discontinuing old ones. As a result, a few skill clusters suddenly disappear while other very similar skill clusters see sudden surges in their number of jobs. To smooth this noise, we merge *problematic* skill clusters into clusters with similar skills. This approach allows us to maximize the coherence of AI exposure labels while minimizing noise in the job count of skill clusters caused by the renaming of skills by the platform. We obtain a final set of 116 clusters, of which 12 are labeled as substitution, 59 as complementary and 45 as unaffected. Appendix section B provides additional information on how we create the AI exposure labels including details on manual labeling, prompts, label revision, and robustness.

2.5. Difference-in-differences and causal inference

To estimate the causal effect of the ChatGPT launch on labor demand, we use a standard difference-in-differences design. Our control group consists of skill clusters where we assume ChatGPT cannot significantly aid in accomplishing tasks associated with jobs in those clusters. The treated clusters are labeled as either substitutable or complementary (examples of substitutable and complementary clusters can be found in Appendix C). We compare the two treated groups to the control group and estimate how the difference between them evolved since the launch of ChatGPT. More concretely, we express the log count of new job postings in cluster i in week t as:

$$\begin{aligned}\log(Postings_{it}) = \beta_0 + \beta_1 \text{Complementary}_i + \beta_2 \text{Substitutable}_i + \gamma \text{After}_t \\ + \delta_1(\text{Complementary}_i \times \text{After}_t) + \delta_2(\text{Substitutable}_i \times \text{After}_t) + \epsilon_{it}.\end{aligned} \tag{1}$$

After_t is a dummy variable that gets value 1 after the public launch of ChatGPT on the week of the 30th of November 2022 (and zero before that week). The binary variable Complementary_i gets value 1 if cluster i is labeled as Complementary, and analogously, the binary variable Substitutable_i has value 1 if the cluster is considered containing mainly skills relevant for tasks where humans can be substituted by LLMs. As sensitivity checks, we also estimate variations of Eq. (1) with cluster fixed and week fixed effects to account for seasonal variation in market labor demand and possible gaps in data collection.



(a) Distribution of the number of skills per project **(b)** Distribution of skills

Fig. 2. (a) Distribution of the number of skill tags per job posting. (b) Overall distribution of the skill tags in the data set.

To estimate the effect at the skill cluster level we use the following regression model

$$\log(Postings_{it}) = \beta_0 + \beta_1 \text{Treated}_i + \gamma \text{After}_t + \delta_i(\text{Treated}_i \times \text{After}_t) + \epsilon_{it}, \tag{2}$$

where $\text{Treated}_i = 1$ whenever the cluster is in the substitutable or complementary categories and δ_i corresponds to the effect on skill cluster i.

The log transformation on the outcome variables of Eqs. (1) and (2) allows us to interpret the estimated δ's as the percentage changes in postings caused by the launch of ChatGPT. To account for possible correlation within skill clusters, we cluster our standard errors at the skill-cluster level.

For the difference-in-differences model to capture the causal effect of the ChatGPT launch on labor demand, we need to make the standard parallel trends assumption, i. e. that in the absence of treatment, the difference between the two treatment groups and the control group would have remained unchanged. In our setting, this implies that we need to assume that the ChatGPT launch would not have had an effect on the demand for labor in the unaffected cluster. While this is ultimately untestable, we demonstrate that the control group and the two treatment groups evolved almost in parallel before the launch of ChatGPT.

There are some additional threats to our identification strategy that we cannot address. The platform in question may have made marketing decisions or pricing changes after the launch of ChatGPT. If these activities were concentrated on skill clusters labeled as substitutable or complementary, the difference-in-differences design might partly capture these non-ChatGPT effects. However, web analytics data from , indicates that monthly traffic on the platform has remained approximately constant around the ChatGPT launch date. Moreover, there were wide concerns about a labor shortage — i.e. a large number of unfilled vacancies — in the U.S. in months around the launch of ChatGPT. This could have led to an increased demand for platform labor. Given that all three groups showed similar trends before the ChatGPT launch, it is unclear why this business cycle effect would be stronger for skill clusters in the unaffected and complementary categories, though.

3. Results

3.1. Descriptive results

3.1.1. Skill requirements of job postings

Before examining the impact of LLMs on online labor markets, we present a descriptive analysis of our dataset. As illustrated in Fig. 2(a), most job postings list between three to eight skills, with a median of five skills per job posting. The ubiquity of skills across job postings is highly heterogeneous, as depicted in Fig. 2(b). Popular skills such as JavaScript, graphic design, and content writing appear in nearly a million job postings, highlighting their fundamental role in the gig economy. Conversely, specialized skills like English-Dutch translation, PowerPoint expertise, and deep neural network development are featured in approximately a thousand job postings, reflecting their niche demand. Additionally, highly specialized skills such as Montessori education or specific sign languages are mentioned only a handful of times, indicating a very limited but specific demand.

3.1.2. Embedding space and AI exposure labels

Our clustering and labeling methodology results in 116 skill clusters, of which 12 clusters are labeled as 'substitutable' (i. e., those types of jobs can be conducted by a non-expert with the help of LLMs), 59 as 'complementary' (i. e., those jobs in which the use of LLMs increases productivity substantially, but which still require expert knowledge) and 45 as 'unaffected' (i. e., those job types where LLMs do not help freelancers or where they lead only to a non-substantial productivity increase). Appendix C showcases examples of skill clusters, their representative skills, their assigned category, and the reasoning provided by GPT for the category

Table 1
Descriptive Statistics of the data per AI exposure category - Panel A: number of clusters, means, median and standard deviations. Panel B: distribution of expected project duration by category. Panel C - expected worker experience levels within categories.

Panel A: Descriptive Statistics of variable Job postings (logarithmic values in parentheses)				
Category	Number of clusters	Mean	Median	Standard deviation
Unaffected	45	451 (5.3)	208 (5.3)	707 (1.3)
Complementary	59	295 (5.2)	185 (5.2)	316 (1.1)
Substitutable	12	402 (5.2)	166 (5.1)	660 (1.2)
Panel B: Expected Duration Shares				
	Up to 3 weeks	**3-9 weeks**	**9-18 weeks**	**18-52 weeks**
Unaffected	48%	22%	10%	20%
Complementary	47%	25%	10%	18%
Substitutable	51%	22%	11%	16%
Panel C: Worker Experience Level Shares				
	Novice	Intermediate	Veteran	
Unaffected	9%	61%	30%	
Complementary	8%	58%	34%	
Substitutable	12%	60%	28%	

assignment.

The substitutable category contains the fewest assigned clusters, while the complementary category has the most. This distribution is expected, as complementary skill clusters require only a subset of tasks to be influenced by LLMs, whereas substitutable clusters necessitate that most tasks be automatable. Consequently, the combinatorial possibilities for complementary skill clusters are larger, resulting in both more clusters and greater heterogeneity within this category (Arthur, 2010). This pattern is consistent with findings from Eloundou et al. (2023), who show that only 1.8% of tasks in the U.S. labor market can be fully automated, 18.5% of workers have more than 50% of their tasks exposed, but for about 70% of tasks only certain components could be completed by an LLM.

Fig. 3(a) illustrates how the transformer model embeds job postings based on their skill requirements and clusters those with similar skills together. These figures are a two-dimensional projection of the high-dimensional word embedding space. The first panel shows the eight largest clusters (i. e. clusters with the most job postings) within the substitutable category. These clusters are mostly related to writing and translation. Writing-related clusters such as proofreading (red), blog writing (light green), and script writing (purple) are close together on the right side of the embedding space. In contrast, translation topics such as German and Japanese translation are close together on the left side. The second and third panels show the eight largest clusters in the complementary and unaffected categories. These figures intuitively show that more similar clusters are placed closer together (e. g., in the complementary category, programming-related skills clusters are located close to each other).

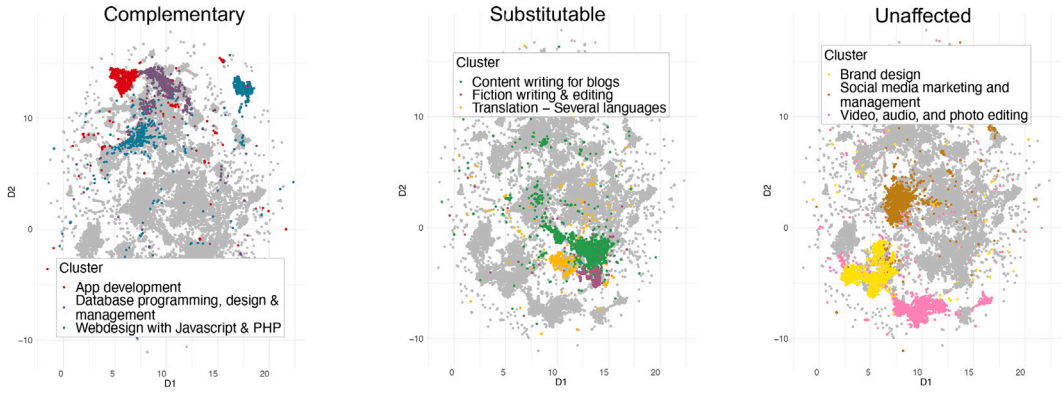
Fig. 3(b) shows the most frequent skills required by job postings in each category. To a large extent, those match our expectations: In the complementary category, we frequently find skills related to software development (e.g. web development, web design, API) as well as references to specific coding languages or tools (e.g. javascript, html, css, php). This aligns with existing research (Moussiades et al., 2024) showing that ChatGPT can efficiently be used to significantly speed up coding. Within the substitutable category, we see different types of writing skills (e. g., blog writing, copywriting, article writing) as well as translation skills (e. g., French, English, Spanish to English translation). Lastly, within the unaffected category, we find many skills that are related to visual capabilities (e. g., graphic design, Adobe Photoshop, illustration), which could not be done with the release version of ChatGPT from November 30th 2022.

3.1.3. Descriptive statistics

After clustering the job postings and labeling the clusters, we aggregate the data on a weekly level. This data forms our analysis sample (Table 1). The average unaffected cluster has 451 job postings per week, the average complementary cluster has 295 job postings per week, and the average substitutable cluster has 402 new job postings per week. Across the three categories, means are much larger than medians, indicating a right-skewed distribution of job postings. After a log transformation, the means and medians of the variables are relatively close to one another.

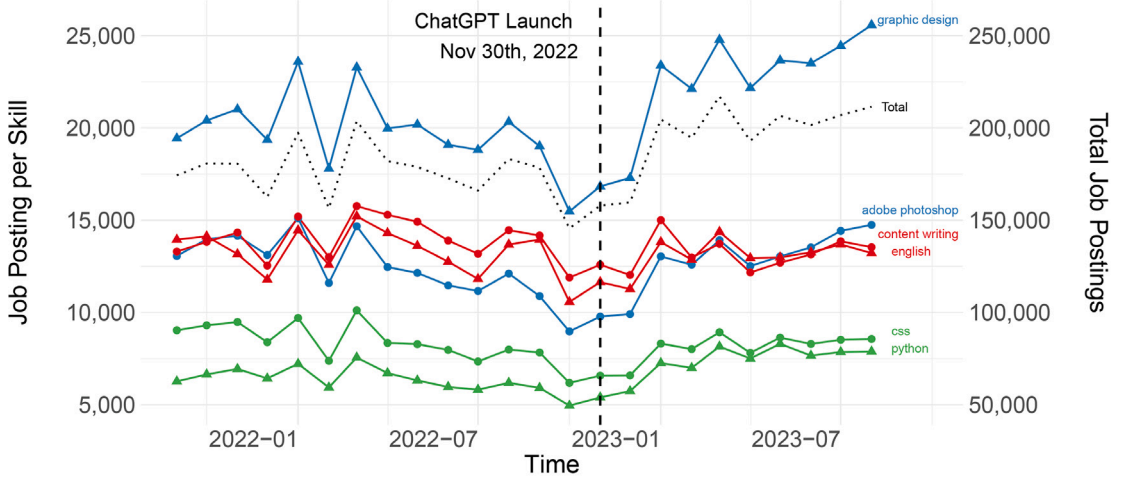
In Panel B, we show the distribution of expected contract lengths across categories. These distributions are very similar to each other. The same applies to the desired worker experience level. In roughly 60% of the job postings, the employer has specified a preference for hiring a worker with intermediate experience. About 30% of job postings seek a veteran, and approximately 10% would consider hiring an inexperienced worker.

The overall number of job postings in the online labor market has grown substantially, as shown in Fig. 3(c). This time series shows the demand dynamics for popular skills, which are either substitutable (red), complementary to (green), or unaffected (blue) by LLMs. The demand for exemplary skills in the complementary and unaffected groups has grown substantially, while it showed a downward trend for skills in the substitutable category.



(a) Skill clusters in the embedding space

(b) Wordclouds of most frequent, important words per category



(c) Development of skill-specific job demand over time

Fig. 3. (a) Two-dimensional representation of the skill embedding space with 1% random sample of jobs plotted and the three most frequent skill clusters per AI exposure category highlighted. (b) Wordclouds of the 30 most observed important skills per category. (c) Development of job postings between September 2021 and August 2023 over time for important skills in each AI exposure category.

Table 2
Difference in Difference regression results - models (1)–(4) show the overall effects, models (5)–(8) the effects by project length, and models (9)–(11) show the results by worker experience level. Overall, job demand in the substitutable categories decreased compared to the counterfactual.

	Overall effects				Effects by project length				Effects by worker experience		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Intercept	5.21***	7.21***			3.95***	3.20***	2.49***	3.09***	2.96***	4.65***	4.01***
	(0.19)	(0.70)			(0.18)	(0.19)	(0.19)	(0.22)	(0.21)	(0.20)	(0.19)
Substitutable	0.05	−0.13	0.05		0.02	−0.21	−0.21	−0.44	0.25	0.04	−0.01
	(0.37)	(0.34)	(0.38)		(0.36)	(0.37)	(0.39)	(0.41)	(0.43)	(0.37)	(0.37)
Complemented	−0.10	−0.05	−0.10		−0.08	0.01	−0.16	−0.21	−0.21	−0.13	−0.05
	(0.24)	(0.21)	(0.24)		(0.22)	(0.25)	(0.24)	(0.27)	(0.26)	(0.24)	(0.24)
After	**0.16****	**0.12****			0.11**	0.14***	0.19***	0.06	0.29***	0.13***	0.17***
	(0.04)	(0.05)			(0.04)	(0.04)	(0.04)	(0.05)	(0.05)	(0.04)	(0.04)
Substitutable ×After	**−0.28****	**−0.29****	**−0.28****	**−0.28****	−0.29**	−0.14	−0.12	−0.06	−0.19*	−0.28**	−0.30**
	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)	(0.10)	(0.07)	(0.09)	(0.08)	(0.09)	(0.10)
Complemented ×After	−0.05	−0.01	−0.05	−0.05	−0.02	0.01	−0.02	−0.06	−0.16*	−0.03	−0.04
	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.06)	(0.07)	(0.07)	(0.05)	(0.05)
Controls	–	✓	–	–	Exp. project duration (x weeks or less)				Worker experience level		
Week fixed effects	–	–	✓	✓	3	9	18	52	Novice	Intermediate	Veteran
Cluster fixed effects	–	–	–	✓							
Observations	11 960	11 524	11 960	11 960	11 955	11 847	11 443	11 615	11 552	11 960	11 942
R^2	0.005	0.13	0.02	0.97	0.004	0.007	0.01	0.01	0.02	0.005	0.004

Notes: *p < 0.05, **p < 0.01, ***p < 0.001.

3.2. Regression results

Here, we turn to investigating the effect the introduction of ChatGPT had on job demand in the three different AI exposure categories: substitutable, complementary, and unaffected. In Fig. 4(a), we plot the total number of job postings in each AI exposure category. All three categories follow approximately similar time trends before the introduction of ChatGPT. This is confirmed in the event study plot shown in Fig. 4(b). It illustrates that the difference between the two treatment labels (complementary and substitutable) and the unaffected control group remained on average constant before the launch of ChatGPT.

Specifically, no significant differences in time trends between the treatment and control groups are evident, and the disparity between both treatment groups and the control group is effectively negligible across all but one week in the pre-treatment period.[2]

We present our difference-in-differences regression results in Table 2 (models 1 – 4). Model (1) shows the results without any fixed effects. In model (2), we add two control variables: the percentages of projects with different durations (under 3 weeks, 3–9 weeks, 9–18 weeks, and 18–52 weeks) and projects aimed at three experience levels of freelancers (novice, intermediate, and veteran). Since our unit of observation is the number of new job postings within each cluster-by-week cell, we introduce these control variables as percentage shares of the number of job postings within each cell. In model (3), we introduce week fixed effects, and in model (4), we further incorporate cluster fixed effects.

Overall, we find the following: After the launch of ChatGPT, the number of job postings in the substitutable clusters decreased by 24% ($100 * e^{.28} - 1$), relative to the unaffected clusters.[3]

The decrease can be decomposed into two effects: (i) the number of new job postings in the unaffected clusters increased by 17% after the introduction of ChatGPT ($100 * e^{.16} - 1$) and (ii) the number of substitutable job postings decreased by 7% (−24% + 17%) after the introduction of GPT.

Additionally, we find that the point estimates on the term Complementary×After are also negative but not statistically significant. As we show below, this is partly due to the fact that some complementary clusters experienced a decrease in demand after the launch of ChatGPT while others experienced an increase.

The difference-in-differences estimates remain highly consistent across the four different model specifications, increasing our confidence that no unobserved confounders are driving our results.

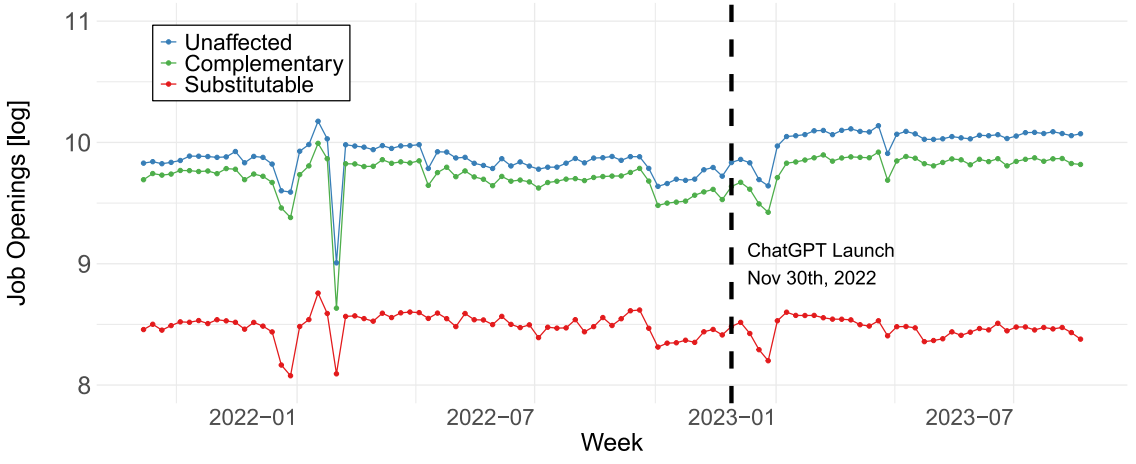
3.2.1. Project duration heterogeneity

To investigate the heterogeneity of the effects of ChatGPT by expected project length, we split the data into subsets based on the expected project length and then aggregate the four data subsets into week × cluster cells. The results are presented in Table 2 (models 5 – 8). We find that the demand reduction in the substitutable clusters is driven by a decrease in short projects lasting three weeks or less.

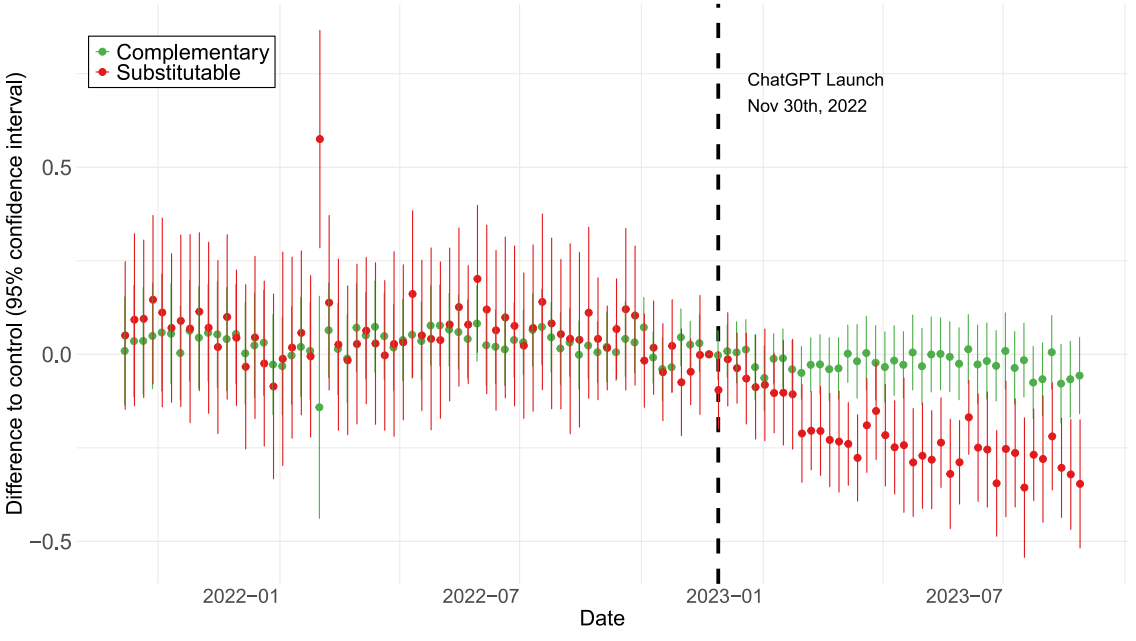
Conversely, we observe an increase in labor demand in the unaffected clusters even for short projects, as evidenced by the positive coefficient on the After indicator variable. The difference between the unaffected and complemented clusters is not statistically

[2] An exception is noted during the week of the 31st of January 2022 when a transitory spike occurred in the substitution group relative to the two other groups. We cannot ascertain whether this is related to an error in the data collection or an exceptional number of job postings on the platform. Nevertheless, as detailed in Appendix E, our findings remain substantially unchanged even when we exclude the data from this anomalous week in our pre-treatment analysis.

[3] Throughout this section, we transform log points into percentages using the formula $100 \times (e^x - 1)$, i. e., $100 * (e^{-0.28} - 1) \approx -24\%$).



(a) Development of demand (job openings) over time



(b) Estimated demand shock in the substitutable and complementary group around the introduction of ChatGPT in November 2022

Fig. 4. (a) Logarithm of aggregated demand per AI exposure category between September 2021 and August 2023. (b) OLS coefficient estimates and their 95% confidence intervals showing the difference between each treatment group and the control.

significant and is also much smaller economically across all project duration categories. These findings indicate that the launch of ChatGPT resulted in a shift of work from short gigs to longer projects in the substitutable clusters.

3.2.2. Worker experience level heterogeneity

To study the heterogeneous effects with respect to worker experience levels, we split our data into three subsets based on the employers' desired worker experience and then aggregate the data into week × cluster cells (analogous to the previous subsection). We present our results in Table 2 (models 9 – 11).

The effect of the ChatGPT launch on the demand for substitutable skills is roughly similar across desired worker experience levels. The regression coefficient on the Substitutable×After term is slightly smaller in absolute value in the novice subset compared to the other two subsets. However, due to large standard errors, this difference is not statistically significant. More interestingly,

we find a negative and statistically significant difference in differences estimate for the complementary group for novice workers. This indicates that the demand for novice labor decreased in the complemented cluster after the launch of ChatGPT. This finding highlights the importance of considering other dimensions of heterogeneity beyond skill requirements when studying the effects of LLMs on the labor market.

3.2.3. Effect heterogeneity by skill cluster

Next, we formally examine the heterogeneity of the impact of the ChatGPT launch across clusters in the substitutable and complementary AI exposure categories by estimating the regression in Eq. (2) presented in the methods section.

Fig. 5 illustrates the estimated changes in demand for each skill cluster in the substitutable and complementary category. Most substitutable clusters exhibit a decline compared to the counterfactual: Writing related to real estate and 'about us' pages show the most substantial decreases, at 52% and 59%, respectively. The largest cluster in our sample of affected skill clusters — content writing for blogs — declined by 20%. Translation is also significantly impacted, with the 'Western European Languages' category experiencing a 23% decline. In the complementary clusters, demand changes are mixed, with some clusters showing significant decreases and others increases. Notably, the demand for jobs in the cluster 'AI-powered chatbots' nearly tripled, with a 179% increase. The cluster of 'Machine learning' also shows a significant rise, with demand increasing by 29%.

Cluster-specific estimates should be interpreted with caution for two reasons. First, while the complementary and substitutable clusters show no statistically significant pre-trends on average, individual clusters may still display notable pre-trends, which could influence some results. Rather than excluding clusters with statistically significant pre-trends, we present results for all clusters. Given the large number of comparisons, some statistically significant pre-trends will occur by chance, and it is difficult to distinguish between true pre-trends and those arising randomly. Second, comparing estimates requires adjusting for multiple hypothesis testing. However, adjusting for every pairwise comparison is impractical, as it would inflate standard errors and make Fig. 5 uninformative.[4] We therefore recommend focusing on the top and bottom of the distribution shown in Fig. 5.

Notwithstanding these caveats, the main takeaway from Fig. 5 is that ChatGPT resulted in substantial demand increases for clusters related to AI for chatbot development, lead generation, and machine learning and significantly decreased demand for writing and translation jobs.[5]

The results from Fig. 5 align with results on the performance of LLMs across languages (Li et al., 2024) and domains (del Rio-Chanona et al., 2024). The significant decline in demand for translation of Western European languages compared to less common languages, such as Arabic and Hebrew, is consistent with the superior performance of LLMs in Western languages, making them more easily substitutable. Clusters related to writing that do not exhibit a significant decrease in demand include Spanish localization and translation, fiction writing, editing, and script writing for explainer videos. These skills require an understanding of language nuances, creativity, and interaction with video content, which explains their relatively stable demand. In programming, the most affected clusters are related to JavaScript, HTML, and CSS, which are also among the most affected programming languages in collaborative programming platforms such as Stack Overflow (del Rio-Chanona et al., 2024). The increase in demand for AI-powered chatbot development and machine learning is intuitive, as businesses increasingly adopt these technologies.

3.2.4. Sensitivity analyses

We have conducted a series of sensitivity analyses to demonstrate that our results are robust to reasonable variations in the regression model and pass standard checks for identification. While the main exhibits related to these exercises are provided in the appendices, we briefly discuss each of them here individually.

Placebo tests. We first provide a standard 'placebo treatment time' test to show that our results do not capture underlying trends that are unrelated to ChatGPT. We do this by entirely dropping the post-ChatGPT period from our data. Thereafter, we introduce a placebo treatment set to 30th May 2022. In Appendix F, we show that the difference in differences estimates are indistinguishable from zero in this case.

Dropping graphic design. So far, we have focused on studying the effects of LLMs on the labor market, ignoring other breakthroughs in generative artificial intelligence technologies related to image generation (such as Dall-E, originally released in 2021, and Midjourney, originally released in 2022). According to our labeling which only focuses on text-based models, graphic design tasks are mostly in unaffected clusters. Nonetheless, one could argue that our results might be affected by graphic design jobs in the control group, which were likely affected by the advent of advanced AI image generation capabilities. We addressed this by dropping all clusters that include 'graphic design' as one of their important skills from the control group. We report the results of this robustness check in Appendix G, which are virtually unchanged.

[4] Depending on the exact method for multiple hypothesis testing correction, we would need to inflate the standard errors by some constant for every 70×69 pairwise comparison we make.

[5] The standard errors on cluster-specific estimates are very similar because the analysis sample consists of week-by-cluster cells, resulting in identical sample sizes across clusters.

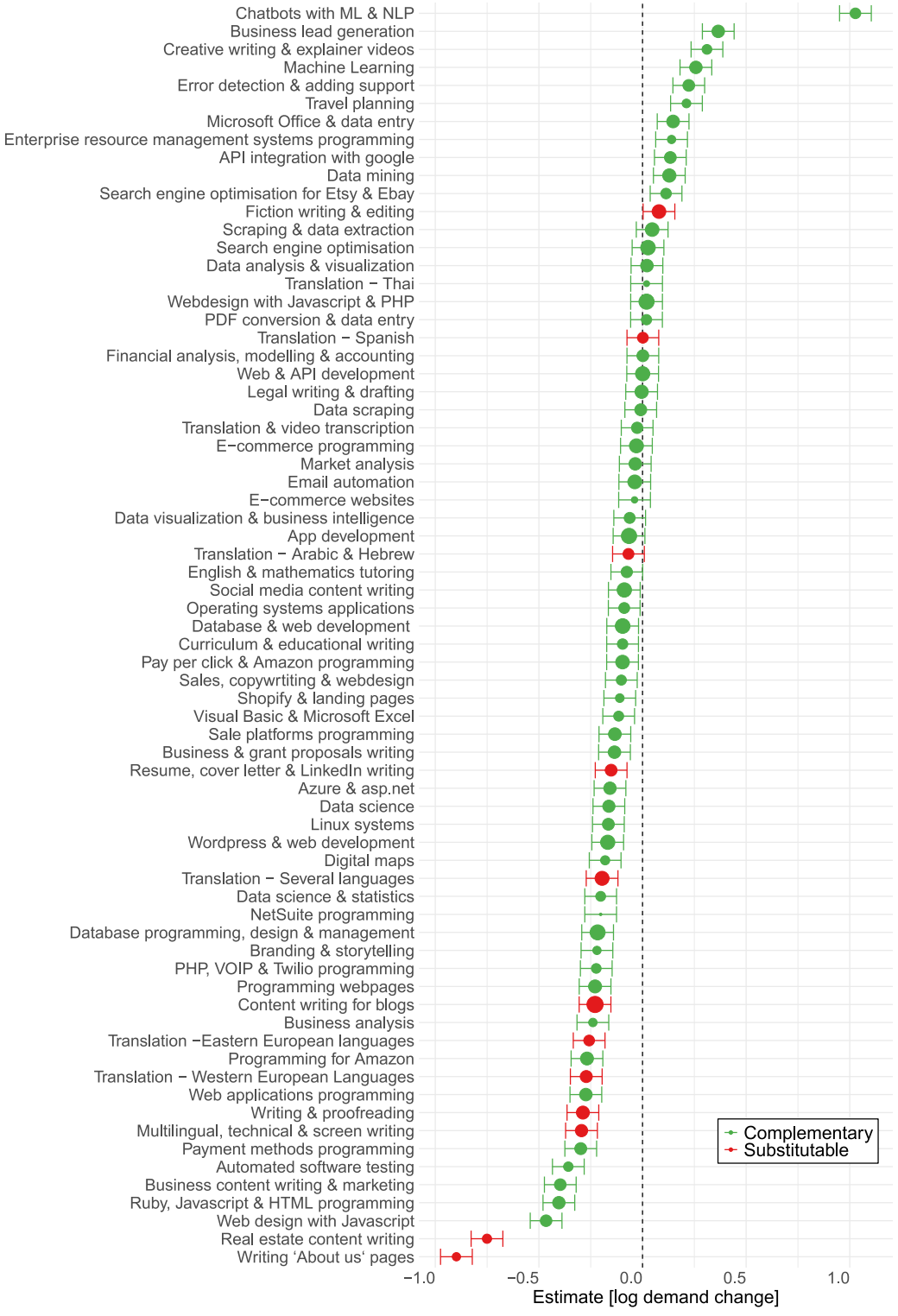


Fig. 5. Estimates of log change in demand at the skill cluster level. Error bars denote standard errors, and the node size is proportional to the log average number of job postings in the given skill cluster.

Randomization inference. As noted by MacKinnon and Webb (2020), randomization inference can provide more accurate p-values when the share of treated clusters is small relative to the control clusters, resulting in either too large or too small standard errors.

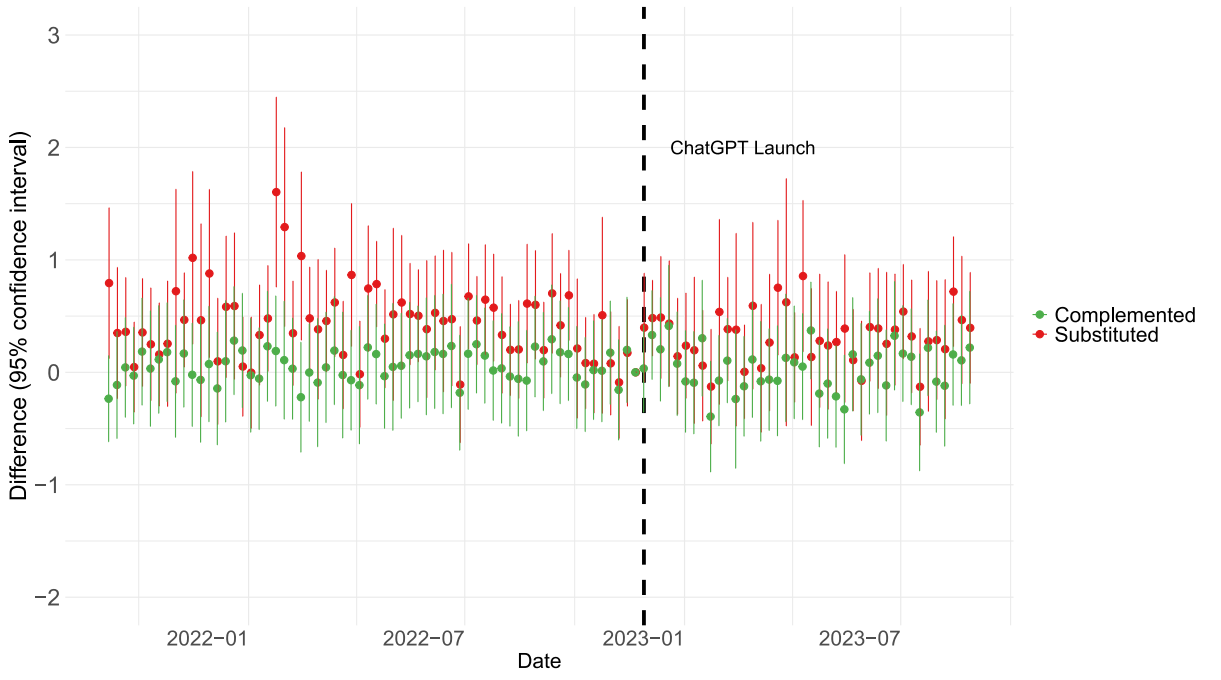


Fig. 6. Event study: log mean budget - This figure plots the difference in log mean budget between the unaffected cluster and substituted and augmented clusters and the corresponding 95% confidence intervals. The vertical line corresponds to ChatGPT launch. Standard errors are clustered at job posting cluster level.

Accordingly, we implemented a randomization inference procedure as follows: we first shuffled the treatment labels (Unaffected, Complementary, Substitutable) randomly. Then, we re-estimated the difference-in-differences model given by Eq. (1) using the shuffled treatment labels, saving the difference-in-differences terms δ_1 and δ_2. Repeating this procedure many times produces a distribution of δ_1 and δ_2 pairs under the null hypothesis that the launch of ChatGPT has no effect on the demand for complementary and substitutable clusters. Comparing the actual estimates of δ_1 and δ_2 to this distribution of placebo estimates allows us to understand how extreme they are in comparison. We report the results of this exercise in Appendix H. The results indicate that the p-values based on randomization inference are very similar to those based on the cluster robust covariance matrix reported in the main text.

3.2.5. Isolating the demand shock from general equilibrium effects

The previous analysis demonstrated a reduction in job postings for substitutable skill clusters following the launch of ChatGPT. Here, we strengthen the argument that this decline reflects a demand-side shock rather than a reduction in labor supply with three key points.

First, we highlight, that using *job postings* as a measure of labor demand is a well-established approach in labor economics (Davis et al., 2013; Marinescu and Wolthoff, 2020; Barnichon, 2010). Job postings are initiated by employers to signal a need for labor, making them a relatively pure indicator of demand, independent of supply-side factors such as the number or characteristics of job seekers.

Second, we show that the mean budget per job posting spent by employers require complementary and substituted skill clusters did not change due to the introduction of ChatGPT. This suggests that employers were willing to pay the same price for labor but demanded less of it, consistent with a leftward shift in the labor demand curve following the ChatGPT launch. Although we lack data on realized wages for all employers, some job postings include a desired employer budget as a part of the job posting. We use the subset data of job postings of employers who reported a budget, and replicate the event study analysis of the previous section, but with the log mean budget as the dependent variable. Fig. 6 and Column (1) of Table 3 show the results of this analysis. There is no significant change in the average employer budget across complementary and substitutable clusters in comparison to the unaffected category.

Third, we show that applications per job posting increased due to a decrease in job postings rather than an increase in applications. To do this, we replicate the event study analysis using the log of the average number of applicants per job posting as the dependent variable. As Fig. 7 and Column (2) of Table 3 show, the number of applicants per project increased after the launch of ChatGPT. Moreover, we can decompose the increase in applications per job posting as follows

$$\log\left(\frac{\sum_j Applicants_{jit}}{Postings_{it}}\right) = \log\left(\sum_j Applicants_{jit}\right) - \log(Postings_{it}). \tag{3}$$

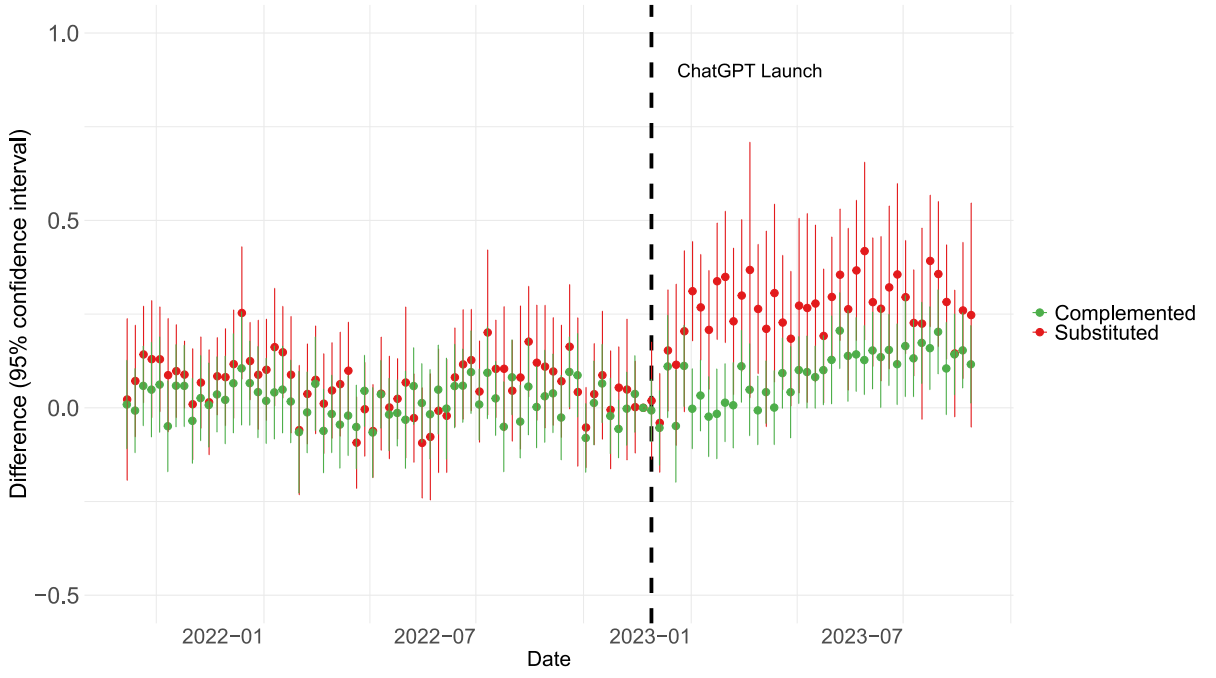


Fig. 7. Event study: log applicants per job posting - This figure plots the difference in log mean number of applicants between the unaffected cluster and substituted and augmented clusters and the corresponding 95% confidence intervals. The vertical line corresponds to ChatGPT launch. Standard errors are clustered at job posting cluster level.

Table 3
Difference-in-Difference Results - Additional Outcomes: Column (1) reports the difference-in-difference estimates for the effect of ChatGPT on employers' reported budgets (measured in log-$) across the two treatment groups compared to the control group. Column (2) reports the effect of ChatGPT launch on average applicants per job posting (log) across the two treatment groups compared to the control group. Standard errors are clustered at the cluster level.

	(1)	(2)
Dependent Variable	log(Avg. budget)	log(Avg. Applicants/Job posting)
Intercept	6.44***	2.97***
	(0.13)	(0.05)
Substituted	−0.50**	0.05
	(0.18)	(0.10)
Complemented	−0.02	0.04
	(0.15)	(0.08)
After	0.08*	0.06***
	(0.03)	(0.02)
Substituted ×After	−0.13	0.20***
	(0.07)	(0.05)
Complemented ×After	−0.03	0.07**
	(0.04)	(0.03)
Observations	11 930	11 957
R^2	0.02	0.03

Notes: *p < 0.05, **p < 0.01, ***p < 0.001.

From Column (1) of Table 2 and Column (2) of Table 3 we know that $\log(Postings_{it}) = -0.28$ and $\log\left(\frac{\sum_j Applicants_{jit}}{Postings_{it}}\right) = 0.20$, with respective standard errors 0.09 and 0.05. The difference between the estimates is within the standard errors, and therefore we cannot attribute the increase in applications per job posting to an increase in the supply.[6]

[6] More formally: let $R = \log\left(\frac{\sum_j Applicants_{jit}}{Postings_{it}}\right)$, $A = \log\left(\sum_j Applicants_{jit}\right)$, and $P = \log\left(Postings_{it}\right)$. The empirical estimates imply, that $\Delta R = 0.2$, and $\delta P = -.28$. Solving for ΔA yields $-.08$. To assess the statistical significance of ΔA, we can solve for its standard error: $SE(\Delta A) = \sqrt{SE(\Delta R)^2 + SE(\Delta P)^2} \approx .10$ (assuming independence between R and P).

In summary, we have demonstrated that the launch of ChatGPT led to a decrease in labor demand without reducing average employer budgets. Furthermore, the number of applicants per job posting increased after the launch, and this increase can be almost completely explained by the reduction in the number of job postings. Taken together, these findings suggest that the launch of ChatGPT led to a reduction in labor demand, with the same number of applicants competing for a smaller pool of substitutable jobs.

While our data does not capture wage changes, it is possible that increased competition may have reduced equilibrium wages. If this is the case, our demand estimates should be interpreted as a lower bound on the true effect. However, Liu et al. (2023), using a similar dataset and research design, did not find evidence of reduced project values, suggesting that the impact on workers' wages may be limited.

One limitation of the previous analysis is that we lack information about the identity or characteristics of the applicants, which may have changed over time. If the launch of ChatGPT led to a less qualified applicant pool, this could have also contributed to a reduction in labor demand. This shift in applicant quality — along with the possibility of ChatGPT directly replacing certain types of work — could be one reason for the observed decrease in labor demand following the launch of ChatGPT. However, we think this is unlikely for two reasons. First, we do not find an increase in supply, meaning that if there was an influx of less qualified applicants, more qualified workers must have left the platform. We have no reason to suspect this would have happened given that the average budget per job remained constant. Moreover, the platform in question allows employers to closely monitor workers' performance and even withhold payment if the deliverable is deemed insufficient in quality. This level of oversight reduces the likelihood that employers would decrease labor demand purely due to uncertainty about worker quality.

4. Discussion

To understand how AI is transforming labor markets, it is crucial to understand its impact on demand across diverse skill sets, distinguishing between jobs that AI can substitute and those it can complement. In this paper, we analyze job posting data from online labor markets, which are highly flexible and quick to react to external changes. We use transformer-based topic modeling to cluster over 3 million job postings into 116 fine-grained skill clusters. Additionally, we employ prompt engineering to label these clusters as substitutable by, complementary to, or unaffected by AI. Finally, we employ a difference in differences analysis to study the impact of ChatGPT's release on demand across 71 skill clusters which we have categorized as either substitutable by or complementary to AI.

Following the introduction of ChatGPT, we find a reduction in demand of approximately 24% for skills deemed substitutable by LLMs compared to unaffected skill clusters, illustrating the potential of LLMs to automate certain tasks historically performed by humans. Conversely, the overall demand for skills complementary to LLMs did not change significantly. However, some complementary skill clusters, such as those related to chatbot development and machine learning, showed substantial increases in demand, while others experienced slight decreases. Regarding wages and geography, substitutable skills are associated with higher wages and are typically performed by freelancers in high-wage countries. In contrast, complementary skills exhibit a broader wage spectrum and are often performed by freelancers from middle-income countries. The impact varies by project duration and worker experience levels, with demand decreasing mainly in short-term projects within the substitutable category. Moreover, while complementary skill clusters did not see a reduction in demand as a whole, novice workers within complementary skill clusters experienced a significant contraction.

Our results provide a nuanced perspective on the labor market effects of generative AI technologies like ChatGPT and other LLMs. Unlike previous literature, which focused primarily on a limited subset of skill clusters in the online freelancing market (Hui et al., 2023; Demirci et al., 2023; Qiao et al., 2023; Liu et al., 2023), our analysis reveals both job destruction in the substitution category and job creation in certain complementary skill clusters. Within the writing and translation sectors, our study uncovers significant heterogeneity: for instance, demand for "about us" page writing has decreased more than for multilingual and technical writing, while translation demand is most affected for Western European languages. Although the decline in demand for substitutable work is concerning, it is predominantly concentrated in short-term gigs rather than long-term projects. Additionally, we highlight the role of new technologies in driving job creation, exemplified by the increased demand for AI-powered chatbots. The advancements in LLMs have significantly improved chatbot quality, significantly expanding their possible use cases and consequently driving higher demand for chatbot development.

The decrease in demand for novice workers within complementary skill clusters may seem to contradict experimental evidence suggesting that AI tools benefit inexperienced workers the most (Noy and Zhang, 2023; Brynjolfsson et al., 2023). However, while these experimental studies are conducted in controlled environments where workers are directly given the AI tools, in the real labor market, it is typically companies or clients who outsource work to freelancers. Thus, novice workers within firms may benefit from AI tools, reducing the need to hire external freelancers. This underscores the fact that the impact of technologies on workers depends on how tools are implemented and who is provided access to them.

This paper's analysis has limitations. First, online labor platforms evolve over time, introducing new features and disabling others, which introduces a degree of noise into our data collection. As outlined in Section 2.3, skill names change over time (e.g., *Photoshop* to *Adobe Photoshop*), and as discussed in Section 2.2, we have occasional gaps in our data due to software errors and platform interface changes. However, using transformer-based hierarchical topic modeling, we create a stable time series of skill clusters. Our results remained consistent when controlling for or excluding the few problematic periods, indicating that the missing data adds only random noise rather than systematic bias. Another limitation is our reliance on GPT-4 for labeling skill clusters exposure to automation. This is a highly challenging task, even for expert human labelers, as it requires specialized knowledge

across a vast array of skills. By using GPT-4, we were able to efficiently label a wide range of skills, allowing for a comprehensive and consistent categorization of all job postings in the online labor market. We minimized errors using state-of-the-art prompt engineering, measuring accuracy with a test set, and performing manual checks.

In addition, this paper focuses on the partial equilibrium effects of new technology and does not capture potential productivity increases outside of the platform under study. Technological improvements can decrease worker demand, but productivity gains can lead to new capital and organizational investments, potentially increasing labor demand Acemoglu and Restrepo (2018), Autor and Salomons (2018). To better understand how substitution of freelancers with GenAI tools affects in house labor, one would need to model the fragmentation of production into tasks in more detail (Rubbo, 2023).

A further limitation is that we only observe labor demand but not wages. Thus, we cannot definitively attribute changes in job postings to shifts in the demand curve. Workers might decrease their wage bids if they face more competition, or they might increase their wage bids if they believe ChatGPT enhances their value. However, previous research (Dube et al., 2020; Duch-Brown et al., 2022; Horton, 2021) suggests that online labor market workers have low bargaining power and act as price-takers. Moreover, our analysis indicates, that our results are not substantially biased by simultaneity issues arising from labor supply responses to the ChatGPT launch. Specifically, we demonstrate that while the launch of ChatGPT led to a decrease in labor demand, average employer budgets remained unchanged. Additionally, we observe an increase in the number of applicants per job posting is of the same magnitude than the reduction in job postings. These findings collectively support the conclusion that ChatGPT's launch led to reduced labor demand, with workers competing for a smaller pool of substitutable jobs.

While this analysis focuses on the short-term impacts of LLMs, understanding the long-term effects requires examining how these impacts vary based on workers' skills (Teutloff et al., 2023; Mealy et al., 2018), location (Farinha et al., 2020; Braesemann et al., 2022), and the alignment between current skills and those needed for emerging jobs (Neffke et al., 2024; Neffke and Henning, 2013). Our results provide evidence of first-order demand changes, which can be integrated into more complex models of worker adaptation (del Rio-Chanona et al., 2021; Acemoglu and Autor, 2011) to answer the broader question on the long-term effects of LLMs. Future research could examine the timing of demand shifts more precisely, investigating how long it takes employers to incorporate AI-driven tools and what the transition period looks like. Looking more at the firm side could reveal more about how to generalize from findings on contract work platforms to traditional labor markets. In addition, further work could explore whether the tasks whose demand decreased are now performed more effectively, performed differently, or not performed at all.

Our study reveals that a greater proportion of skill clusters are complemented by AI rather than substituted. However, while some of these complementary skill clusters experienced an increase in demand, others saw a decline. This underscores the critical role of public and private decision-makers in shaping the socio-economic impact of AI (Ritala et al., 2023; Gerling et al., 2024). Managers and policy-makers need to support workers in adapting and reskilling and should empower them to use LLMs in complementary ways (Retkowsky et al., 2024; Mollick et al., 2024). Our findings from online labor markets, which adapt quickly to technological changes, provide insights into potential future effects on traditional labor markets. Although traditional markets may adapt more slowly due to contract and severance laws, they will likely follow similar patterns over time. Policymakers must implement reskilling and upskilling programs to help workers transition into roles that complement AI technologies, focusing on foundational skills (Hosseinioun et al., 2023) and analytical thinking (Humburg and Van der Velden, 2017). Overall, our study demonstrates that while AI tools like ChatGPT can substitute some jobs, it also drives the development of new products and services, exemplifying creative destruction (Klimek et al., 2012) and contributing to the economy's evolution towards greater complexity (Balland et al., 2022).

Statements

Declaration of generative AI in scientific writing

During the preparation of this work the authors used ChatGPT 4.0 in order to improve the readability and language of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Funding information

Otto Kässi's work was funded by TT-säätiö (project title: "Tekoäly työelämässä – Miten käy Suomen kilpailukyvyn?").

R. Maria del Rio-Chanona's research was supported by the James S. McDonnell Foundation.

Fabian Braesemann is supported by funding from the Oxford Internet Institute's Research Programme on AI & Work, funded by the Dieter Schwarz Stiftung gGmbH.

We are grateful for the support obtained by Open AI in providing access to the GPT API.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix. Online appendix: supplementary analyses

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jebo.2024.106845.

Data availability

The authors do not have permission to share data.

References

Acemoglu, D., Autor, D., 2011. Skills, tasks and technologies: Implications for employment and earnings. In: Handbook of Labor Economics, vol. 4, Elsevier, pp. 1043–1171.

Acemoglu, D., Restrepo, P., 2018. The race between man and machine: Implications of technology for growth, factor shares, and employment. Am. Econ. Rev. 108 (6), 1488–1542.

Adams-Prassl, A., Balgova, M., Qian, M., 2020. Flexible work arrangements in low wage jobs: Evidence from job vacancy data.

Aggarwal, C.C., Hinneburg, A., Keim, D.A., 2001. On the surprising behavior of distance metrics in high dimensional space. In: Database Theory—ICDT 2001: 8th International Conference London, UK, January 4–6, 2001 Proceedings 8. Springer, pp. 420–434.

Albanesi, S., da Silva, A.D., Jimeno, J.F., Lamo, A., Wabitsch, A., 2023. New Technologies and Jobs in Europe. Technical Report, National Bureau of Economic Research.

Anderson, K.A., 2017. Skill networks and measures of complex human capital. Proc. Natl. Acad. Sci. 114 (48), 12720–12724.

Arntz, M., Gregory, T., Zierahn, U., 2016. The Risk of Automation for Jobs in OECD Countries: A Comparative Analysis. oecd.

Arthur, W.B., 2010. The Nature of Technology: What it is and How It Evolves. Penguin UK.

Ash, E., Hansen, S., 2023. Text algorithms in economics. Annu. Rev. Econ. (ISSN: 1941-1391) 15 (1), 659–688. http://dx.doi.org/10.1146/annurev-economics-082222-074352.

Autor, D., 2024. Applying AI to Rebuild Middle Class Jobs. Technical Report, National Bureau of Economic Research.

Autor, D., Chin, C., Salomons, A., Seegmiller, B., 2024. New frontiers: The origins and content of new work, 1940–2018. Q. J. Econ. qjae008.

Autor, D., Salomons, A., 2018. Is Automation Labor-Displacing? Productivity Growth, Employment, and the Labor Share. Technical Report, National Bureau of Economic Research.

Balland, P.-A., Broekel, T., Diodato, D., Giuliani, E., Hausmann, R., O'Clery, N., Rigby, D., 2022. The new paradigm of economic complexity. Res. Policy 51 (3), 104450.

Barnichon, R., 2010. Building a composite help-wanted index. Econom. Lett. 109 (3), 175–178.

Bhattacharjee, P., Mitra, P., 2021. A survey of density based clustering algorithms. Front. Comput. Sci. 15, 1–27.

Braesemann, F., Stephany, F., Teutloff, O., Kässi, O., Graham, M., Lehdonvirta, V., 2022. The global polarisation of remote work. Plos one 17 (10), e0274630.

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D., 2020. Language models are few-shot learners. CoRR, abs/2005.14165. URL: https://arxiv.org/abs/2005.14165.

Brynjolfsson, E., Li, D., Raymond, L.R., 2023. Generative AI at Work. Technical Report, National Bureau of Economic Research.

Brynjolfsson, E., McAfee, A., 2014. The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies. WW Norton & Company.

Brynjolfsson, E., Mitchell, T., 2017. What can machine learning do? Workforce implications. Science 358 (6370), 1530–1534.

Corporaal, G.F., Lehdonvirta, V., 2017. Platform Sourcing: How Fortune 500 Firms are Adopting Online Freelancing Platforms. University of Oxford.

Davis, S.J., Faberman, R.J., Haltiwanger, J.C., 2013. The establishment-level behavior of vacancies and hiring. Q. J. Econ. 128 (2), 581–622.

del Rio-Chanona, R.M., Hermida-Carrillo, A., Sepahpour-Fard, M., Sun, L., Topinkova, R., Nedelkoska, L., 2023. Mental health concerns precede quits: shifts in the work discourse during the Covid-19 pandemic and great resignation. EPJ Data Sci. 12 (1), 49.

del Rio-Chanona, R.M., Laurentsyeva, N., Wachs, J., 2024. Large language models reduce public knowledge sharing on online Q&A platforms. PNAS nexus 3 (9), pgae400.

del Rio-Chanona, R.M., Mealy, P., Beguerisse-Díaz, M., Lafond, F., Farmer, J.D., 2021. Occupational mobility and automation: a data-driven network model. J. R. Soc. Interface 18 (174), 20200898.

Demirci, O., Hannane, J., Zhu, X., 2023. Who is AI replacing? The impact of ChatGPT on online freelancing platforms. The Impact of ChatGPT on Online Freelancing Platforms (October 15, 2023).

Dube, A., Jacobs, J., Naidu, S., Suri, S., 2020. Monopsony in online labor markets. Am. Econ. Rev. Insights 2 (1), 33–46.

Duch-Brown, N., Gomez-Herrera, E., Mueller-Langer, F., Tolan, S., 2022. Market power and artificial intelligence work on online labour markets. Res. Policy 51 (3), 104446.

Eloundou, T., Manning, S., Mishkin, P., Rock, D., 2023. Gpts are gpts: An early look at the labor market impact potential of large language models. arXiv preprint arXiv:2303.10130.

Farinha, T., et al., 2020. Impacts From Automation Diffuse Locally Â€"A Novel Approach To Estimate Jobs Risk In Us Cities. Technical Report, Utrecht University, Department of Human Geography and Spatial Planning ....

Frey, C.B., Osborne, M.A., 2017. The future of employment: How susceptible are jobs to computerisation? Technol. Forecast. Soc. Change 114, 254–280.

Gerling, C., Teubner, T., Braesemann, F., 2024. Who uses generative AI, how, and why? A cluster analysis on motives, perceptions, and use patterns of ChatGPT. Electronic Markets (In revision).

Grootendorst, M., 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794.

Grootendorst, M.P., 2024a. Outlier reduction - BERTopic — maartengr.github.io. https://maartengr.github.io/BERTopic/getting_started/outlier_reduction/outlier_reduction.html. (Accessed 20 May 2024).

Grootendorst, M.P., 2024b. Parameter tuning - bertopic — maartengr.github.io. https://maartengr.github.io/BERTopic/getting_started/parameter%20tuning/parametertuning.html#top_n_words. (Accessed 20 May 2024).

Hansen, S., Lambert, P.J., Bloom, N., Davis, S., Sadun, R., Taska, B., 2023. Remote Work across Jobs, Companies, and Space. National Bureau of Economic Research, http://dx.doi.org/10.3386/w31007.

Hoberg, G., Phillips, G., 2016. Text-based network industries and endogenous product differentiation. J. Polit. Econ. (ISSN: 1537-534X) 124 (5), 1423–1465. http://dx.doi.org/10.1086/688176.

Horton, J.J., 2021. The Ruble Collapse in an Online Marketplace: Some Lessons for Market Designers. Technical Report, National Bureau of Economic Research.

Horton, J.J., Tambe, P., 2015. Labor economists get their microscope: Big data and labor market analysis. Big data 3 (3), 130–137.

Hosseinioun, M., Neffke, F., Youn, H., et al., 2023. Nested skills in labor ecosystems: A hidden dimension of human capital. arXiv preprint arXiv:2303.15629.

Hui, X., Reshef, O., Zhou, L., 2023. The short-term effects of generative artificial intelligence on employment: Evidence from an online labor market. Available at SSRN 4527336.

Humburg, M., Van der Velden, R., 2017. What is expected of higher education graduates in the twenty-first century? In: The Oxford Handbook of Skills and Training. Oxford University Press US, p. 201.

Karjus, A., 2023. Machine-assisted mixed methods: augmenting humanities and social sciences with artificial intelligence. arXiv:2309.14379.

Kässi, O., Lehdonvirta, V., 2018. Online labour index: Measuring the online gig economy for policy and research. Technol. Forecast. Soc. Change 137, 241–248.

Klimek, P., Hausmann, R., Thurner, S., 2012. Empirical confirmation of creative destruction from world trade data. PLoS One 7 (6), e38924.

Li, Z., Shi, Y., Liu, Z., Yang, F., Liu, N., Du, M., 2024. Quantifying multilingual performance of large language models across languages. arXiv preprint arXiv:2404.11553.

Liu, J., Xu, X., Li, Y., Tan, Y., 2023. " generate" the future of work through AI: Empirical evidence from online labor markets. arXiv preprint arXiv:2308.05201.

Lukac, M., 2021. Two worlds of online labour markets: Exploring segmentation using finite mixture models and a network of skill co-occurrence.

Lysyakov, M., Viswanathan, S., 2023. Threatened by AI: Analyzing users' responses to the introduction of AI in a crowd-sourcing platform. Inf. Syst. Res. 34 (3), 1191–1210.

MacKinnon, J.G., Webb, M.D., 2020. Randomization inference for difference-in-differences with few treated clusters. J. Econometrics (ISSN: 0304-4076) 218 (2), 435–450. http://dx.doi.org/10.1016/j.jeconom.2020.04.024, URL: https://www.sciencedirect.com/science/article/pii/S0304407620301445.

Marinescu, I., Wolthoff, R., 2020. Opening the black box of the matching function: The power of words. J. Labor Econ. 38 (2), 535–568.

McInnes, L., Healy, J., 2017. Accelerated hierarchical density based clustering. In: 2017 IEEE International Conference on Data Mining Workshops. ICDMW, IEEE, pp. 33–42.

McInnes, L., Healy, J., Melville, J., 2018. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426.

Mealy, P., del Rio-Chanona, R.M., Farmer, J.D., 2018. What you do at work matters: new lenses on labour. What You Do at Work Matters: New Lenses on Labour (March 18, 2018).

Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Mollick, E., Mollick, L., Bach, N., Ciccarelli, L., Przystanski, B., Ravipinto, D., 2024. AI agents and education: Simulated practice at scale. arXiv preprint arXiv:2407.12796.

Moussiades, L., Zografos, G., Papakostas, G., 2024. GPT-4 vs. GPT-3.5 As Coding Assistants. Research Square Platform LLC, http://dx.doi.org/10.21203/rs.3.rs-3920214/v1.

Neffke, F., Henning, M., 2013. Skill relatedness and firm diversification. Strateg. Manag. J. 34 (3), 297–316.

Neffke, F., Nedelkoska, L., Wiederhold, S., 2024. Skill mismatch and the costs of job displacement. Res. Policy 53 (2), 104933.

Noy, S., Zhang, W., 2023. Experimental evidence on the productivity effects of generative artificial intelligence. Available at SSRN 4375283.

OpenAI, 2023. Best practices for prompt engineering with openai API. https://help.openai.com/en/articles/6654000-best-practices-for-prompt-engineering-with-openai-api. (Accessed 12 February 2024).

Peng, S., Kalliamvakou, E., Cihon, P., Demirer, M., 2023. The impact of ai on developer productivity: Evidence from github copilot. arXiv preprint arXiv:2302.06590.

Pennington, J., Socher, R., Manning, C.D., 2014. Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. EMNLP, pp. 1532–1543.

Qiao, D., Rui, H., Xiong, Q., 2023. AI and jobs: Has the inflection point arrived? Evidence from an online labor platform. arXiv preprint arXiv:2312.04180.

Retkowsky, J., Hafermalz, E., Huysman, M., 2024. Managing a ChatGPT-empowered workforce: Understanding its affordances and side effects. Bus. Horiz..

Ritala, P., Ruokonen, M., Ramaul, L., 2023. Transforming boundaries: how does ChatGPT change knowledge work? J. Bus. Strategy (ahead-of-print).

Rubbo, E., 2023. Fragmentation of production and the wage distribution. Working Paper.

Stephany, F., Kässi, O., Rani, U., Lehdonvirta, V., 2021. Online labour index 2020: New ways to measure the world's remote freelancing market. Big Data Soc. 8 (2), 20539517211043240.

Stephany, F., Teutloff, O., 2024. What is the price of a skill? The value of complementarity. Res. Policy 53 (1), 104898.

Teutloff, O., Stenzhorn, E., Kässi, O., 2023. Skills, job application behavior and the gender wage gap: Evidence from online freelancing. Available at SSRN.

Van Der Maaten, L., Postma, E.O., van den Herik, H.J., et al., 2009. Dimensionality reduction: A comparative review. J. Mach. Learn. Res. 10 (66–71), 13.

Webb, M., 2019. The impact of artificial intelligence on the labor market. Available at SSRN 3482150.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E.H., Le, Q., Zhou, D., 2022. Chain of thought prompting elicits reasoning in large language models. CoRR abs/2201.11903. URL: https://arxiv.org/abs/2201.11903.