Preliminary Draft

Remote Work across Jobs, Companies, and Space

Stephen Hansen
University College London

Peter John Lambert
London School of Economics

Nick Bloom
Stanford University

Steven J. Davis
Hoover Institution & Chicago Booth

Raffaella Sadun
Harvard University

Bledi Taska
Lightcast

February 22, 2023*

Abstract

The pandemic catalyzed an enduring shift to hybrid and fully-remote work arrangements. To study this shift, we examine 250 million vacancy postings in five English-speaking countries. Our measurements rely on a state-of-the-art language-processing framework that we fit, test, and refine using 30,000 human classifications. We achieve 99% accuracy in flagging job postings that advertise hybrid or fully remote work, greatly outperforming dictionary-based methods and also outperforming other machine-learning methods. From 2019 to 2022, the share of postings that say new employees can work remotely one or more days per week rose three-fold in the U.S and by a factor of five or more in Australia, Canada, New Zealand and the U.K. These developments are highly non-uniform across and within cities, industries, occupations, and companies. Even when zooming in on employers in the same industry that compete for talent in the same occupations, we find huge differences in the share of postings that say the job allows remote work.

Keywords: remote work, hybrid work, work from home, job vacancies, text classifiers, BERT, pandemic impact, labour markets
JEL Codes: E24, O33, R3, M54, C55

*Thanks for outstanding research assistance to Yabra Muvdi, who built and estimated the classification algorithm, and to Miaomiao Zhang and Kelsey Shipman, who supported the data analysis. Hansen gratefully acknowledges financial support from ERC Consolidator Grant 864863, Lambert from the London School of Economics STICERD PhD research grant, Bloom from the Smith Richardson and John Templeton Foundations, and Davis from the Templeton Foundation and the Booth School of Business at the University of Chicago. Selected visualisations and downloadable data accompanying this paper can be found at www.WFHmap.com.

1 Introduction

The COVID-19 pandemic propelled an enormous uptake in hybrid and fully-remote work. Over time, it has become clear that this shift will endure long after the initial forcing event. Looking forward, U.S. survey data say that one-quarter of full workdays will happen at home or other remote location after the pandemic ends, five times the pre-pandemic rate (Barrero et al. 2021). The pandemic also drove large, enduring increases in remote work in dozens of other countries (Criscuolo et al. 2021, Aksoy et al. 2022). There are few, if any, modern precedents for such an abrupt, large-scale shift in working arrangements.

Most previous efforts to quantify and characterize this shift rely on surveys of workers and employers and assessments of remote-work feasibility by occupation. We rely instead on the information contained in job vacancy postings. Specifically, we consider the full text of 250 million postings in five English-speaking countries. In doing so, we apply state-of-the-art language-processing methods to analyze the text and determine whether the job allows for remote work. We fit, test, and refine our language-processing model using 30,000 classifications generated by human readings. We also identify the city, employer, industry, occupation, and other attributes associated with each job vacancy.

Vacancy postings pertain to the flow of new jobs rather than the stock[1]. In addition, postings that promise remote work two days a week, for example, entail a commitment–or at least a statement of intent–that extends into the future. For both reasons, postings need not show the same pattern of remote work as the currently employed. Indeed, the remote-work share of postings lags far behind the remote-work share of employment in the pandemic's early stages. And while the incidence of remote work among the employed fell markedly in the two years after spring 2020, we show that the remote-work share of postings rose sharply over the same period.

Our approach to studying the remote-work phenomenon has several noteworthy strengths. First, our data cover almost all vacancies posted online by job boards, employer websites, and vacancy aggregators from 2014 to 2022 in our five countries. Coverage on this scale is infeasible with survey methods. Second, postings typically describe the job and its attributes in considerable detail, as suggested by a median posting length of 347 words. Comparable

[1]Vacancy distributions by industry, employer size, and worker turnover rates differ greatly from the corresponding employment distributions, and the differences are highly sensitive to labor market conditions. See Davis et al. (2012).

detail is hard to obtain from other sources, especially at scale[2]. Third, we apply frontier methods to develop a language-processing model that reads postings and classifies remote-work status in an automated manner. The model achieves a 99% accuracy rate in flagging jobs that allow for remote work, greatly outperforming dictionary-based methods. Our model also outperforms a variety of other methods. Fourth, the combination of scale, rich text data, and automation lets us characterize the shift to remote work in a highly granular manner. We track the evolution of remote work at a monthly frequency in hundreds of occupations, thousands of cities, tens-of-thousands of employers, and in city-by-occupation and employer-by-occupation cells. We post many of these statistics at www.WFHmap.com, with frequent updates.

The share of postings that say new employees can work remotely one or more days per week was tiny before the pandemic: 1% or less in Australia, Canada and New Zealand as of 2019, about 3% in the U.K., and about 4% in the U.S. From 2019 to 2022, this remote-work share rose three-fold in the U.S and five-fold or more more in the other countries. As of November 2022, the remote-work share exceeds 10% of postings in Australia, Canada, the U.K, and the U.S. and appears to remain on an upward trajectory in all five countries.

Remote-work posting shares vary greatly across occupations, industries, and cities. Looking across occupations, the remote-work share correlates positively with computer use, education, and earnings. Finance, Insurance, Information and Communications have especially high remote-work shares. Chicago, London, New York, San Francisco, Toronto, and other cities that function as business service hubs have high remote-work shares. These differences have widened since the pandemic struck. According to a linear least-squares regression, 63% of the variation across occupations in 2022 remote-work shares is accounted for by their 2019 shares. In contrast, just 19% of city-level variation in 2022 remote-work shares is accounted for by 2019 shares. These patterns suggest that city-level variation in work-from-home adoption rates depends on more than the occupational structure of the city's workforce.

We also find that the shift to remote work is highly non-uniform across same-industry employers–even when recruiting in the same occupational category. This emergent heterogeneity on the demand side expands opportunities to satisfy preferences over remote work on the supply side[3]. Our non-uniformity result also carries another important message: in many

[2]Previous research exploits the detail in vacancy postings to study technical change, the cyclicality of skill requirements, their relationship to wages, how compensation and other job attributes affect applicant flows and, of course, to classify jobs in a fine-grained manner. Examples include Modestino et al. (2016), Deming and Kahn (2018), Hershbein and Kahn (2018), Davis and Samaniego de la Parra (2020), Forsythe et al. (2020), Marinescu and Wolthoff (2020), and Acemoglu et al. (2022).

[3]On preference heterogeneity in regards to remote work, see Bloom et al. (2015), Mas and Pallais (2017), Wiswall and Zafar (2018), Barrero et al. (2021), and Aksoy et al. (2022).

occupations, it is misleading to think of remote-work suitability as a purely technological constraint. Remote-work intensity is, instead, an outcome of choices about job design and how to operate an organization. These choices are influenced by the external environment and subject to shock-induced shifts. In line with this view, Aksoy et al. (2022) find that employers plan higher levels of work from home after the pandemic ends in countries that experienced longer and stricter government-mandated lockdowns during the pandemic.

Several recent papers use job postings to study the remote-work phenomenon. See Draca et al. (2022) for the U.K., Alipour et al. (2020) for Germany, and Bamieh and Ziegler (2022) for Austria. Bai et al. (2021) use pre-pandemic postings in the U.S. to construct firm-level indexes of remote-work feasibility, which they relate to post-pandemic performance as measured by sales, net income, and equity returns. Perhaps the closest forerunner to our paper is Adrjan et al. (2021), who use postings data at the country-sector-month level to study remote work from January 2019 to September 2021.

Previous studies use dictionary methods (keyword search criteria) to identify postings that allow for remote work. We take a different approach that implements a pipeline for text analysis that is popular in computer science but rarely used in economics. First, we sample 10,000 job posting fragments and ask humans to label each as positive (the text indicates the possibility to work-from-home one or more days per week) or negative (otherwise). Each fragment receives three labels from workers on Amazon Mechanical Turk (AMT). Second, we use the DistilBERT model (Sanh et al. 2020) to estimate the relationship between job posting text and the human labels[4]. Finally, we take the estimated model out-of-sample to impute a classification to all job postings in the corpus.

We compare our estimated model to other approaches in terms of performance in a held-out set of labeled texts not included in estimation. These include dictionary methods, logistic regression, and GPT-3, the model that heavily influenced the development of the recently released and popular ChatGPT tool. Our method substantially outperforms other methods with respect to accuracy rates, the F1 score (geometric mean of true-positive and true-negative classification rates), and other metrics. GPT-3 yields the second-based test-set classification performance. Reasons for this out performance of DistilBERT over dictionary approach includes negative mentions (e.g. phrases like "Work from home: not possible" or "teleworking not currently permitted"), phrase variations (e.g. "work comfortably from home" or "home or office working possible") and compound terms (e.g. "work from home

[4]DistilBERT is a smaller, faster version of the BERT model introduced by Devlin et al. (2019a), which has 57 thousand Google Scholar citations as of January 2023. BERT and DistilBERT exploit machine-learning tools and are pre-trained on the full English-language Wikipedia corpus and the Toronto Book Corpus. For a helpful non-technical overview of BERT, see Luktevich (2022).

come facilities" or "requires a Home Office work permit"). After extensive experimentation we found the frequency, range and non-random nature of these types of challenging terms necessitated large language methods trained on a human coded training sample.

To our knowledge, we are the first to deploy modern, large language models at scale to predict human annotations for a concrete economic measurement problem and to establish the performance gains from doing so[5]. (Shapiro et al. 2022) finds little gain to using BERT relative to dictionaries for detecting the sentiment of news articles but uses fewer than 1,000 labeled examples for estimation on a corpus of text with far less negation, phrase variation and compound challenges. By scaling up labeling with AMT, we can harvest an order of magnitude more labels which allows flexible language models to reach their full potential. In principle, this approach can be deployed on arbitrary measurement problems involving the extraction of economic concepts from text. To facilitate the adoption of our approach, we make available the code for estimation as well as the set of labeled example at `https://github.com/yabramuvdi/WHAM`. This code is efficient enough to generate 1.5 billion out-of-sample classifications in 36 hours for less than $1500 on Google Cloud.

A prominent line of research classifies occupations as suitable or unsuitable for remote work based on descriptions of work activities and experiences[6]. We highlight some limitations of this approach. First, remote-work intensity is a malleable feature of jobs, occupations, and organizations. Second, classifications based on suitability assessments explain little of the variation in remote-work posting shares. For occupations that Dingel and Neiman (2020) classify as unsuitable to be done entirely from home, the remote-work share of U.S. postings in 2022 ranges from 0 to 51% with a mean of 5% and standard deviation of 7%. For occupations they classify as suitable for work from home, the share ranges from 0.3 to 74% with a mean of 18% and standard deviation of 12%.

Another prominent line of research surveys workers and employers to study working arrangements. Barrero et al. (2020), Bartik et al. (2020), Bick et al. (2022) and Brynjolfsson et al. (2020) document and characterize the enormous uptake in work from home in spring 2020. Bartik et al. (2020), Barrero et al. (2021) and Ozimek (2020) use employer plans and other forward-looking survey data to forecast that the big shift to remote work will endure. Relative to our approach, the survey-based approach is more useful for eliciting the perceptions, attitudes, and expectations of workers and employers. Our approach offers

[5]Bajari et al. (2021) and Bana (2022) use BERT to predict prices from Amazon product reviews and wages from job posting text, respectively. Each paper achieves high predictive performance. Our innovation is to pair a language model with human labels as the target variable.

[6]Dingel and Neiman (2020) is the most influential example. Other examples include del Rio-Chanona et al. (2020), Mongey et al. (2021), and Adams-Prassl et al. (2022). Like us, Adams-Prassl et al. (2022) concludes that remote-work intensity varies greatly across jobs within occupations.

several other distinct advantages, as discussed above.

The next section describes our vacancy posting data and develops our classification model. Section 3 assesses the model's performance in absolute terms, and relative to other approaches. Section 4 sets forth our main findings related to remote-work intensity over time and across countries, cities, occupations, and more. We also compare our remote-working posting shares to survey-based measures of remote work. Section 5 concludes.

2 Data and Measurement

To measure remote-work posting shares, we exploit a near-universe of online job postings from January 2014 through November 2022 for our five countries[7]. [inline]do we do any cleaning on the data? We extract 10,000 text sequences from selected postings and ask humans to read them. Each sequence is about 45 words long, and the average posting has about six sequences. Breaking postings into sequences facilitates human and algorithmic classification at scale, as we discuss below. Our human readers answer this question: 'Does this text explicitly offer an employee the right to remote-work one or more days a week?', yielding a binary classification. The pairwise agreement rate between readers exceeds 90 percent.

We turn to the 'DistilBERT' language processing framework to fit a specific language-processing model for our purposes[8]. The framework is pre-trained on thousands of books and the English-language Wikipedia corpus, which helps the framework interpret the intended meaning of a given document or passage. We further pre-train on roughly one million text sequences drawn from our corpus of vacancy postings. This second pre-training step familiarizes the framework with the nature of the text in vacancy postings.

After pre-training, we use the human classifications to train, or fit, a specific language-processing model. We call this model the 'Working-(from)-Home Algorithmic Measure' (*WHAM*). We will show that *WHAM* achieves near-human performance in its classification task, and that it outperforms a variety of other approaches. We describe our approach in some detail, because we think it has useful applications to many other text-analysis tasks in economics and other fields.

[7]This draft works with a 5% random sample of postings before January 2019 and the universe of postings thereafter.

[8]DistilBERT is a direct descendant of BERT, which is widely used in industry. BERT stands for Bidirectional Encoder Representations from Transformers. Transformers are a deep-learning method in which every output element is connected to every input element of a text sequence, for example, with weights on each element dynamically calculated as the text is processed. See **?** for an overview of how transformers work. **?** is the seminal contribution.

2.1 Job Vacancy Data

We use data from online job vacancies, collected by Lightcast (formerly called Burning Glass Technologies), an employment analytics and labor market information firm[9]. Lightcast web-scrape information from over fifty-thousand different online sources, including vacancy aggregators, government job boards, and employer websites. They claim to cover the "near-universe" of all online job vacancy postings.

For each online job vacancy posting in our dataset, we have have access to a plain text document scrapped from the description of each job listing. Additionally, we can observe the posting date, employer name, occupation, location of the employer, industry and more.

We restrict our attention to job vacancy postings listed between January 2014 and December 2022. We further drop vacancy posting were the occupation is unknown[10]. For the period prior to January 2019 we utilise a 5% sub-sample. In total our dataset contains nearly 250 million online job vacancy postings, spanning five countries, 5.2 million employers, and nearly 40 thousand cities. Table 1 describes the number of vacancies, as well as the count of unique cities and employer strings for each country.

2.2 Representativeness of Our Sample of Online Vacancy

Our sample of online vacancy posting represents the universe of online vacancy postings, as distinct from the flow of all new hires. Further, data on job vacancy postings pertains to the flow of new jobs, rather than the stock. Burke et al. (2020) show that there is a high degree of alignment between the fraction of new job openings across occupations (as measured using the Job Openings and Labor Turnover Survey (JOLTS) data, collected by BLS) and the number of job vacancy postings scraped by Lightcast. To specifically address concerns over representativeness, we perform a number of re-weighting exercises to match our data to various population moments. Our baseline results match data from all countries and time periods to the occupation-distribution of new online vacancy postings from 2019 in the United States. We also report results based on alternative weighting methods (including the raw unweighted series). A more detailed discussion of these re-weighting exercises, along with references to the national statistical data used for each of our five countries, is found in Appendix **??**.

[9]We provide a rich description of these data, and documentation on how we pre-processed the data, in Appendix 6. We believe this documentation will help support other researchers when using these data.

[10]The main reason we drop postings where the Lightcast data did not provide an occupation is that it typically suggests that the raw text is compromised, or missing crucial information. Less than 1% of vacancies do not have an occupation

2.3 Measurement Problem

The measurement problem we face is to determine whether each job posting allows a new hire to work remotely. We adopt a binary classification approach, and refer to a 'positive' posting as one that mentions the ability to work remotely, and a 'negative' posting as one that does not. For positions that offer partial remote working, we use a threshold of at least one day per week for our positive classification[11] This approach effectively measures an employer's willingness to commit *ex ante* to offering flexibility in work location. Negative postings may in fact be associated with work-from-home positions, for example because the ability to work from home is assumed by market participants to be feasible in particular jobs, or because the employer prefers to bargain over work arrangements during the hiring process rather than make a prior binding commitment. We return below to discuss the interpretation of our measure, and first focus on developing an accurate and robust classification.

The most precise way of classifying postings is arguably via direct human reading. Given the size of our data, however, this approach is not feasible to scale and some means of automated classification is required. The most standard approach adopted in the text-as-data literature in economics is to use a dictionary of keywords whose presence is assumed to indicate a positive classification. As an initial step, we use the keywords in Table 8.6 to classify job postings as positive or negative. While we do not claim the dictionary of terms is fully optimized, it is in line with others in the literature for classifying postings as work-from-home (Adrjan et al. 2021).

An issue that becomes immediately apparent upon inspecting job postings that are classified as positive and negative by the keywords is the presence of notable errors, which Table 2 illustrates. False positives include references to companies' home offices and working in homes dedicated to health-care provision. A second, and perhaps most worrying, source of falso positives is that the structure of job ads shifts during COVID-19 in a way correlated with the presence of false positives. This is due to the fact that after early 2020, many postings feature a new text field indicating whether home work is allowed, and then explicitly state it is not–a naive application of the dictionary method would infer from this text field that the job posting allows working from home[12]. Table 2 also lists examples of false negatives, which illustrates the many and complex ways that companies can use to describe remote work. Accounting for this linguistic variety with a fixed set of keywords is a major

[11] In principle our measurement approach could be extended to the intensive margin (days per week), but for simplicity we begin with the this binary classification.

[12] One approach to correcting this problem is to extend the dictionary to incorporate negation (e.g. to treat as a negative classification the phrase 'this is not a remote work position'). In section 3 we show that this indeed improves measurement accuracy but not by as much as our proposed solution below.

challenge.

2.4 Our Approach to Classification

Our approach to address the classification errors in the dictionary approach has three steps. First, we use at least three human auditors to read and classify 30,000 pieces of text extracted from job ads. Second, we train a modern machine learning algorithm using these human classifications. Third, we take this predictive model out-of-sample to classify each job ad as either positive or negative. The hope is to scale the accuracy of human reading—which can only be deployed on a small fraction of data—to the entire dataset. While this approach is common in the machine learning literature, it is not often used in economics, even though it appears to hold great promise. We call the final model used in this paper the *Working-from-Home Algorithmic Measure (or WHAM) model*.

The main text contains an overview of our methodology, with further details in Appendix 8.

2.4.1 Breaking Up Job-Ad Text into Sequences

While we ultimately wish to classify job postings, we initially label and classify smaller units of text we refer to as *sequences*. The first reason for doing so is that human labeling of entire job postings is prone to a high error rate because of their length and complexity. The second reason is that the typical posting has a great deal of information unrelated to remote work.[inline]can we provide an example? Mixing text relevant for work-from-home with a great deal of irrelevant text introduces noise into the classification algorithm.

The procedure for generating sequences has three salient features[13]. First, postings always begin with a job title, e.g. "Software Programmer familiar with R and Python." We extract these as a single sequence. Second, the beginning of each posting typically has a number of bullet points or other structured fields. In most cases, these also form a single sequence[14]. Finally, the remainder of a posting is typically structured like standard prose with a succession of paragraphs. Each paragraph is taken as a single sequence, unless it passes a length threshold. In this case, we break it into multiple sequences of consecutive sentences.

This procedure produces approximately 1.6 billion sequences out of the nearly 250 million job postings.

[13]The code for sequence extraction will be made publicly available.

[14]The exception is if the number of distinct structured fields is too large, in which case we split them into multiple sequences.

2.4.2 Human Labels for Training and Evaluation

From the sample of sequences, we first chose 10,000 to label manually. One quarter of these sequences was chosen at random from the set of sequences that contained a set of dictionary terms listed in Table 8.6. Another quarter was chosen to contain a broad set of terms that might reflect work-from-home language, including the generic terms 'remote', 'home', 'work', 'location'; any word that begins with 'tele'; and any two-word sequence that begins with 'remote'. Another quarter consisted of sequences that might confound a classifier, including 'home repairs', 'nursing home', 'remote construction', etc. The final quarter was a random sample of sequences not satisfying the three aforementioned criteria. Each portion of the label sample is balanced across year-quarter from 2014Q1 through 2021Q3. We also balance the sample evenly across countries[15] to account for varying English idioms in different geographic locations.

We used Amazon Mechanical Turk to generate labels. To ensure high-quality workers, we set up an initial screening test that required prospective workers to label 20 sequences that we had previously manually classified. Only workers that made at most one error were allowed to proceed to label the full set. Another quality control strategy was to pay around 25% above typical market rates for labeling tasks. This motivated workers who passed initial screening to continue on the project[16].

Each of the 10,000 sequences were labeled by three distinct workers. There is a high agreement rate among workers: 66.9% of examples are unanimous negative examples and 25.5% are unanimous positive examples. The remaining 7.6% examples are evenly balanced between one dissenting vote for either positive or negative. Note that, while half of the sample was chosen to contain a word with the potential to denote work-from-home, only 29.2% of the sequences receive a majority of positive votes.

2.4.3 Developing the WHAM model

In the last five years, the field of natural language processing has been revolutionized by models that allow the meaning of word sequences to arise by how they interact. Consider the sentences 'Some of the deep-sea wells we operate are in remote locations' and 'We are pleased to offer opportunities for remote work'. Each includes the word 'remote' but only the latter is a positive example of remote work. The important point is that the interaction of 'remote'

[15] We draw one quarter of this dataset from each of USA, UK, Canada, and a further one quarter from the pooled Australia and New Zealand data.

[16] In general workers appeared engaged and focused on the labeling task. We received communication from multiple workers seeking to clarify ambiguous cases, which went above and beyond what AMT required for payment.

with surrounding context words determines the overall meaning of these sentences. Moreover, not all context words are equally informative: for example, in the first sentence 'deep-sea', 'wells', and 'locations' are more important than 'some' and 'we' in understanding the meaning of 'remote'. *Self-attention* (Vaswani et al. 2017) is a mathematical construct that allows vector representations of individual words to interact with each other to form new vectors that encode the meaning of sequences. These interaction weights effectively determine which words should be "paid attention to" in resolving these meanings. Self-attention is the key idea that powers models such as *BERT*, *RoBERTa*, *GPT*, *GPT-3*, *PALM*, and, most famously, *ChatGPT*. For more details on these models, collectively called Transformers, see Ash and Hansen (2023).

The particular Transformer model we adopt in the development of WHAM is *DistilBERT* (Sanh et al. 2020). DistilBERT is based on Google's BERT model (Bidirectional Encoder Representations from Transformers, Devlin et al. 2019b), which when originally released set important new performance benchmarks for common NLP tasks (since eclipsed by larger-scale models). Since 2020, Google has used BERT to process its online search queries. DistilBERT 'distills' the information in BERT by training a simplified model to reproduce the same output as BERT. The main advantage is that DistilBERT obtains an expressive language model with far fewer parameters which reduces our estimation costs.

We make two main modifications to the off-the-shelf DistilBERT model to build *WHAM*. First, one set of parameters of off-the-shelf DistilBERT is obtained by predicting randomly deleted words in generic English from surrounding context words. We instead update these parameters to predict randomly deleted words in a sample of 900,000 job posting sequences which is balanced across all years and countries. This step creates word representations that are specific to the language of job postings.

Second, we further modify off-the-shelf DistilBERT to predict human labels from vector representations of job posting sequences. We split our labeled sequences into training and test sets of 5,950 and 4,050 sequences, respectively. The prediction problem is conducted at the label rather than sequence level, so there are 3 * 5,950 = 17,850 total observations in the core training sample of labeled data[17]. Appendix 8 details how we specify the prediction model's hyper-parameters. Figure 1 provides an illustration of which words are influential in the classification problem. In the following section, we compare the test-set accuracy of the estimated model with that of other algorithms in the literature and show its performance is

[17]During an initial exploratory phase, we labeled a sample of around 10,000 additional sentences (rather than sequences) using a combination of Mechanical Turk, hired research assistants, and ourselves. Since these are also potentially informative, we include them in the training set. In most cases, these sentences only received a single label and so in total generate 11,574 additional labels in the training set.

outstanding.

2.4.4 Predicting Remote Work Language at Scale

Finally, we use the estimated prediction model to assign a continuous probability to all sequences in our corpus. The higher the probability, the more confidence the model has that this sequence denotes an offer of remote work arrangements. Figure 7.1 plots a histogram of the share of sequences that fall in different probability intervals. The distribution is bimodal at the lowest and highest probability bins, with the former dominating the distribution. As expected, most sequences do not contain work-from-home language because, as we show below, most job postings do not explicitly mention the possibility to work from home and, among those that do, the majority of sequences discuss other features of the position. The bimodality of the distribution shows that the classification algorithm typically produces a clear prediction, in line with human labelers' high agreement rates. e use an 0.5 threshold for assigning a sequence a positive classification according to WHAM's predicted probability, but the properties of the predicted probability distribution imply that our results are not sensitive to this particular cutoff.

2.4.5 Aggregating Measurement back to Job Postings

We have conducted all the analysis so far at the sequence level, but are ultimately interested in a job posting-level classification. For this, we use a simple 'max' rule and positively classify a job posting if it contains one or more positive sequences. Table 7.3 shows the number of positively classified sequences in each job ad. We can see that among the positive job ads (those with one or more positive sequence), the majority have just a single positive sequence. This reduces concern that the algorithm produces correlated false positive hits at the posting level[18]. This posting-level classification constitutes the final output from our WHAM model, which we use to study the adoption of remote work.

2.4.6 Public Access to WHAM

To allow researchers to interact with and study the properties of our model, we make available a simple online tool that allows one to input arbitrary text and receive a predicted probability

[18] We have manually read a number of randomly drawn postings with more than five positive sequences, and found no instance of the algorithm failing. In some cases, the scraping procedure that gathers data from online job portals appears to have identified as a single job ad a succession of postings by recruitment agencies. In other words, the measurement error arises from the data itself rather than the classification approach.

as output. The URL is https://huggingface.co/spaces/yabramuvdi/wfh-app-v2, which will reproduce the same probabilities as in the paper[19]. One can verify that the examples in table 2 that confound dictionary approaches are correctly classified by WHAM.

2.4.7 Computational Performance of WHAM

One constraint on implementing large-scale NLP models is computational...provide statistics once Yabra has them. DONE: Last table of the paper contains this information.

3 Assessing the Performance of WHAM

Above we highlight instances in which the presence or absence of keywords is insufficient to correctly classify a selection of job posting texts due to the complexity of surrounding context. In order to quantify the gains from adopting our approach, we now undertake a systematic comparison of the ability of different algorithms to correctly classify unseen texts. To do so, we adopt a standard approach in the machine learning literature and randomly split the 10,000 human-labeled sequences into training and test sets (of sizes 5,950 and 4,050, respectively). We then train WHAM just on the training data and use the fitted model to assign a predicted value to each test-set observation. By way of comparison, we also use the following alternative methods for classifying test-set observations[20]:

1. *All Zero.* Each test-set observation is assigned a 0 to match the modal outcome.

2. *Dictionary.* We use term the set from Adrjan et al. (2021)[21], and count an observation as positive if it contains a term from this set.

3. *Dictionary with Negation.* Shapiro et al. (2022) shows that accounting for negation can improve the performance of dictionaries. We adopt a similar method and only count the presence of a dictionary term as indicating remote work when a negation term does not appear in the surrounding context.[inline]how many words before/after?

4. *Logistic Regression.* Adams-Prassl et al. (2020) uses Lightcast data from the UK to measure the prevalence of flexible work schedules, i.e. the times at which work must be

[19] Many thanks to Yabra Muvdi for estimating the model and making it accessible. The model is subject to revision, at which point the predicted probabilities for a given text may change. Users who find systematic biases in the predictions are welcome to contact the authors with their findings, which can be incorporated into future work.

[20] More details of each approach are in Appendix 8

[21] The terms are reported in Table 8.6. The remote work measures in Adrjan et al. (2021) are based on data from Indeed which potentially has a different structure from the Lightcast data.

completed, from job posting text. The paper uses humans to manually annotate 7,000 texts, and fits a (penalized) logistic regression model for classification. The features of the logistic regression are the word frequencies in a given document. We implement a similar logistic model on our training data and use it to classify test data.

5. *Logistic Regression with Negation.* We expand the feature set of the logistic regression to incorporate negation and re-estimate it on the training data.

6. *GPT-3.* Brown et al. (2020) introduced GPT-3, a large language model capable of performing a variety of natural language tasks with limited or no training examples to learn from. We query the model with the prompt "Identify if the text offers the possibility of remote work at least one day per week" and convert the answer into a 0/1 classification[22].

7. *WHAM with Generic English.* Rather than pre-train DistilBERT using job posting text, we use its off-the-shelf word embeddings trained on general English.

Table 3 reports the test-set performance for all methods. A straightforward metric is the error rate, i.e. the fraction of mis-classified texts. On this measure, WHAM outperforms all other methods. GPT-3 has an error rate three times that of the baseline model, while the dictionary method's error rate is eight times higher. On the other hand, the pre-training of the model with job posting text generates only a modest an improvement over generic English.

A more standard performance metric in the machine learning literature is the F_1 score which accounts for both a classifier's ability to recover the true positives (*recall*) as well as the share of predicted positives that are true positives (*precision*).[inline]maybe we explain these metrics in footnotes? This varies between 0 and 1, where higher values indicate better performance. Again, we observe that WHAM substantially outperforms all other measures[23].

One concern is that the distribution of positive and negative postings in the test data does not correspond to that of the full population of job postings: the data extracted for labeling

[22]Deploying GPT-3 on the full Lightcast dataset would be prohibitively expensive at current costs, but we still report its test set performance for benchmarking purposes. More recently, ChatGPT, a successor model to GPT-3, has generated a great deal of public interest. ChatGPT is largely built on an underlying model that OpenAI calls *text-davinci-03* whereas GPT-3 is built on *text-davinci-02*. In our experiments, the former outperforms the latter, so we only report results for GPT-3.

[23]An alternative dictionary for measuring remote work adoption is proposed in Draca et al. (2022) which uses our same UK Lightcast sample. The overall error rate of this dictionary in the full test data set is 0.19 and for the test data set arising from the UK is 0.17. Interestingly, the F1 score we obtain for logistic regression (0.81) is similar to that reported by Adams-Prassl et al. (2020) for classifying flexible work scheduling (0.83, see Table 3 of that paper).

is specifically designed to over-represent positive cases. To obtain a sense of classification accuracy on the full population, we create a simulated dataset of 1000 * 4,050 = 4,050,000 observations, 3% (97%) of which are sampled with replacement from the set of positive (negative) test set examples. Table 7.4 reports the same metrics as Table 3 but computed on this more unbalanced dataset. Again, we find that WHAM outperforms all other methods, but in this case the difference in F_1 scores is even starker. Our baseline WHAM achieves a 0.85 F_1 score, while the F_1 score of GPT-3 falls to 0.52 and other methods have even worse performance. Moreover, pre-training becomes more important as the F_1 score for WHAM with generic embeddings drops to 0.78. This is because, as Table 3 shows, WHAM has a particularly low false positive (FP) rate compared to other methods. When negative examples dominate the evaluation sample, correctly classifying them becomes particularly important for overall performance and WHAM is strong in this dimension. Since this sample's label composition is more in line with the expected composition of the universe of job postings, our findings highlight the potential gains in accuracy of using our approach.

We view these results as methodologically important because they are among the first, to our knowledge, to quantify the gains of adopting modern NLP methods for text classification in economic environments. There are very few papers in the economics literature that systematically compare different classification approaches, and those that do have not found that large language models outperform simpler approaches. For example, Shapiro et al. (2022) does not report large gains from using BERT over simpler models for classifying news sentiment. One reason that we, in contrast, do find large gains is the size of our training data. Shapiro et al. (2022) trains BERT on 800 labeled articles whereas we have an order of magnitude more training data, which provides more information for estimating the complex ways in which word sequences map into outcomes. We conjecture that other prediction problems using text in economics might similarly benefit from a large training sample combined with sequence embedding models.

A separate question is how WHAM compares to alternative methods on the full data sample. Rather than consider all alternatives, we focus on how WHAM compares to the Dictionary method, which is most common in the literature measuring remote work adoption from job posting text. Figure 2 plots monthly time series of the share of remote work postings in the US sample from 2019 through 2022. The patterns present in both series differ markedly. According to the Dictionary method, the remote work share surged at the onset of the COVID-19 pandemic, peaked in early 2021, and fell markedly throughout 2021 before stabilizing in 2022. In contrast, the WHAM method suggests a more modest immediate reaction to the pandemic followed by a steady growth rate thereafter. Clearly,

then, the choice of measurement approach can have important quantitative implications even in aggregate[24].

Of course, aggregate comparisons between methods can mask underlying differences at more granular levels. To illustrate this, we compute the growth rate in remote work adoption according to the Dictionary method and WHAM from 2019 to 2022 for individual SOC2 occupations, pooling all 2019 postings and 2022 postings together. In these two years, the Dictionary method appears similar to WHAM but with an upward shift of around five percentage points. However, as figure 3 reveals, there are large differences in the specific occupations that each method associated with growth in remote work adoption. According to the Dictionary method, the 'Food Preparation and Serving' occupation has experienced highest growth in adoption, while for WHAM the highest-growth occupation is 'Computer and Mathematical'. Moreover, according to WHAM all occupations experienced positive growth in adoption, whereas adoption rates fall for the 'Farming, Fishing, and Forestry' occupation according to the Dictionary method. The higher accuracy of WHAM in the sample of human labels suggests its ranking of occupations is more reliable. In the next section we provide a more in-depth analysis of occupation-level heterogeneity according to WHAM.

In sum, the WHAM model displays a very high classification accuracy—relative to human labels—and differs markedly from the most popular alternative approach in the literature based on keyword search. This difference is especially pronounced since 2020, even at the aggregate level. We believe our approach to measurement provides a highly accurate classification of remote work offers in the text of job postings, and base the remainder of the paper on analyzing its output.

4 Results

In this section we document how the *percent of remote work vacancies*–the fraction of all new vacancies which explicitly offer the right to work remotely one or more days per week–has changed over time. We document this across countries, occupations, cities, and employers.

This section is organised as follows: First, we look at the percent of remote work vacancies across countries in a monthly time series, spanning 2014-to-2022. Second, we compare remote work percentages within both broad and narrowly defined occupations, contrasting our measurements in 2019 vs 2022. We show that the substantial rise since the onset

[24]The patterns in the Dictionary series need not match those from Adrjan et al. (2021) even though we use the same set of keywords, as the structure of the Lightcast data could differ in important ways from that of the Indeed data that Adrjan et al. (2021) use.

of COVID is highly uneven across occupations, and find that occupations with the highest 2019 percentage of remote work were the most likely to top the list in 2022. We also compare occupation-level classifications used in the literature to our measurement. Third, we compare the percentage of new jobs offering remote work across cities. We show that cities with higher remote work percentages in 2019 do not strongly predict higher percentages by 2022 (unlike occupations). This suggests that additional confounding city-level characteristics have played an important role in the adoption of remote work. We also compare a monthly time series of the remote work percentage across a selection US cities. Fifth, we compare our measures of the percent of job ads advertising remote work to survey information from the American Communities Survey (ACS). We show that MSA's which have a high remote work share of vacancies in our data also have high fractions of the population responding that they 'mostly work from home' in the ACS. Fifth, we show that the percentage of remote work vacancies posted by employers who operate in the same industry, and search for the same talent, can vary widely.

4.1 Remote Work across Countries

How did the share of advertised remote work differ across countries prior to, during and after the pandemic? In 4 we plot the monthly time series of the share of advertised remote work for the US, UK, Canada, Australia and New Zealand. For each country and in each month, this figure reports the weighted-mean of the percent of remote work vacancies across nearly 800 narrow occupation groups. We weight each group based on the share of vacancies in this group in the USA during 2019. Our baseline results utilise this method to reduce the impact of compositional differences, both across time and across countries[25]. Three high-level facts emerge:

1. **Unprecedented and sharp increase of advertised remote work at the onset of COVID-19**

   In March-April 2020, the share of new job vacancies which advertised remote work saw a sharp rise across all countries. On average, the increase from February 2020 to April 2020 was 200%. While this immediate increase occurred across all our countries, the level-change was most pronounced in countries with a more severe initial COVID outbreak (USA, UK and Canada).

[25] We provide alternatives to 4 in the Appendix: Results, including taking the raw averages as well as reweighting using administrative data on employment. These alternative approaches to constructing the aggregate time series within countries are consistent with all the qualitative results discussed.

2. **Sustained growth thereafter**
   Since the large spike in March-April 2020, there has been sustained growth in the percent of advertised remote work. In level-terms, this growth has been most pronounced in the UK (where COVID lockdowns most lingered and were most severe relative to the other countries in the sample).[inline]wasn't NZ also in a strict lockdown? We also see evidence of higher growth rates (indeed, convex growth paths) in Australia and NZ, as their pandemic experience worsened. In all countries, the growth in advertised remote work has continued long after the forcing event of the pandemic subsided. An additional reason for this high and persistent growth may be that our measure of new job vacancies lags the stock of employees actually working remotely. As employers learn about these arrangements, and are increasingly forced to compete for talent who value this amenity, they become more willing to ex-ante commit to offering remote work explicitly in their recruitment material.[inline]do we really want to speculate?

3. **Substantial heterogeneity across countries, even before the pandemic**
   The USA had nearly 4% advertised remote work share in 2019, the highest of any country. The UK was only marginally lower, where as Australia, Canada and New Zealand had respectively half, a third, and a tenth the share of the US in 2019. By mid 2022 the spread in levels is much greater, but proportional differences have reduced.

In our robustness exercises, we also look at the raw shares of remote work, i.e. without the re-weighting applied to our baseline 4. Comparing the unweighted Figure AXXX to 4 tells us the direction and magnitude of the impact that occupation composition plays in our results. For example, in mid 2022 the difference between the UK and USA is 8 percentage points using the raw data and 4 percentage points after re-weighting. This suggests that roughly half of the difference in advertised remote work shares between the US and UK is accounted for by differences in the types of jobs being advertised, which is unsurprising as the UK is on the whole more skewed towards white-collar jobs with a higher propensity to be worked from home.

4.2 Remote Work across Jobs

We first show the share of advertised remote work by grouping job ads into broad occupation groups (based on two-digit SOC 2010 classifications), which 5 reports. For this, we look only at data from the United States. The differences across broad occupation groups varies greatly. In 2019, we see that just one-in-twenty positions of all job ads in 'Computer and Mathematical' occupations explicitly offered remote work arrangements in their postings,

whereas in 2022 this share raises to a third advertising remote work. As one might expect, the share of advertised remote work correlates positively with computer use, education, and earnings and is lower in occupation groups which require specialised equipment or customer interactions. Lastly, 5 provides some evidence that the pre-pandemic shares of remote work correlate with post-pandemic shares.

To investigate the relationship between pre-pandemic and post-pandemic shares further we next turn to an analysis at the detailed ONET occupation-level. We group our US job vacancies into granular occupations (using O*NET definitions), and plot both the 2019 and 2022 percent of advertised remote work (on a natural-log-scale), presented in 6. After dropping a handful of data points with fewer than 250 postings in 2019, we retain 875 O*NET occupations[26]. 6 also shows the feasibility classification according to Dingel and Neiman (2020). A black circle represents jobs which these authors classify as 'not suitable for full-time telework', and an orange triangle denotes the opposite [27]. An unweighted ordinary-least-squares trend line is also depicted in blue. Our main takeaways from 6 are:

- The bivariate unweighted-OLS fit using a log-log specification yields an R^2 of 0.63, which shows that–for a given ONET occupation–the share of vacancies which advertised remote work pre-pandemic was strongly predictive of the share post-pandemic.
- The slope coefficient from the bivariate unweighted-OLS model shows that the elasticity of 2019 percent to 2022 percent was 0.75%[28].
- Across all ONET occupations depicted, the mean share of new postings which advertised remote work was 4% in 2019 and 10% in 2022.
- There is substantial variation in the share of advertised remote work across occupations, which grows over time. Across all ONET occupations, the standard deviation in the shares of advertised remote work was 5 in 2019 and 11 post-pandemic.
- Dingel and Neiman (2020)'s classification can account for a small part of the variation in the 2020 levels of advertised remote work. For occupations that they classify as unsuitable to be done entirely remotely, the share of advertised remote work post-pandemic ranges from 0 to 51% with a mean of 5% and standard deviation of 6%. For

[26] In total, there are 1,016 O*NET occupations. Our sample of O*NET codes which have greater than 250 vacancy postings in 2019 is 875. This attrition is expected, for example a number of military occupations are not present in our data.

[27] These are taken from the authors replication data, accessed April 2022, which can be found here.

[28] Our ordinary least-squares estimates impose a power-law coefficient, given the log-log specification.

occupations they classify as suitable for telework, the share ranges from 0.3 to 74% with a mean of 18% and standard deviation of 12%[29].

We view three key points of difference between our measurement approach and those measures which assess telework feasibility for each occupation. First, since our measurement works at the job vacancy level and not the occupation level per se, our measure offers more variation and a signal variability in remote work feasibility within occupations across firms. Second, whereas the feasibility measures treat each job as a collection of tasks, our measure combines both task-feasibility as well as employer and employee preferences, labour market forces, past experience with remote arrangements, and so on[30]. The third reason for the discrepancy is that our measurement exercise will likely have some amount of under-reporting, as employers may not explicitly advertise remote work in their vacancies but none-the-less allow such arrangements. We discuss this further in Section **??**, and also show some evidence in the proceeding section that this bias is stable in the cross section, ensuring granular analysis is well informed by our approach.

4.3 Remote Work across Cities

Next we compare the percent of new vacancy postings which advertised remote work across cities. Job posting are matched to a city based on specific locations listed on the website from which it was scrapped, or else mentioned in-text[31].

7 shows the percent of advertised remote work across a selection of large international cities, both for 2019 and 2022. We see that the percentages vary widely. For example, in 2022, 1-in-4 new job postings in Washington (DC) advertised remote work arrangements,

[29] In a few cases, the D&N machine classification appears very inaccurate. For example, travel agents have been classified as 'not teleworkable', although both before and after the pandemic roughly 1-in-3 jobs advertised remote work. This is likewise the case for 'Advertising Sales Agents' and 'Interpreters & Translators'. Some of these outliers appear to be resolved by the hand coded measure, but these data are only available at a higher level of occupational-aggregation.

[30] A clear example of the differences between our measurement approach and Dingle and Neiman (2020) is for teaching jobs. For example, while D&N correctly classify jobs for "physical education teaching" as being *feasible* for working from home (i.e. via a virtual class room), we know anecdotally that this arrangement was very taxing on staff and avoided as soon as normal schooling resumed. We find that teaching jobs in general (and "physical education teachers" in particular) have some of the lowest shares of advertised remote work of any job, highlighting that feasibility and actual behaviour can vary markedly.

[31] Since the predominant remote work arrangements are hybrid, the location of the work site remains a key feature of most jobs. However, in the extreme case of a 'fully remote' position this analysis becomes less precise and relies on [inline]what do we rely on in these cases?. We plan to refine our measurement approach in future work to distinctly classify 'hybrid' vs 'fully remote' work arrangements, but have thus far concluded that the vast majority of jobs offer hybrid arrangements

compared to 1-in-14 in Perth, Australia. The substantial increases as well as the large heterogeneity in these shifts can be seen both across- and within-countries.

Further evidence of the large shift in both levels and spread of remote work job ads is shown in 8, which plots all cities in our data with more than 250 new vacancies in 2019. The mean (standard deviation) increased from 4% (2%) in 2019 to 11% (6%) in 2022. An unweighted OLS regression line fitted to these city-level data show a much lower coefficient of determination (R^2) of 0.19, compared to the value of 0.63 when running the same exercise across occupation-groups. This highlights that the 2019 shares are far less informative predictor of post-pandemic shares at the city level.

This sizable increase in the levels and spread of remote work across cities, as well as the weak relationship between 2019 and 2022 shares, poses an interesting question: What are the city-level determinants of remote work adoption? We hypothesize that a mix of institutional features, infrastructure quality, pandemic severity (both in disease and policy) and the composition of jobs and firms in each city are all important factors. We leave a more formal tests of these predictions to future work.

We next turn to more granular monthly time series for selected US cities, shown in 9. As well as illustrating the granularity of our data, a number of interesting features emerge from these time series:

- Cities from the North-East and West regions (e.g. San Francisco, Boston, New York) all experience similar increases at the outset of the pandemic, but have very different growth levels subsequently. By December 2022, these differential growth rates result in very dispersed levels.

- We also see substantial fluctuations over time in these North-East and Western cities. These fluctuations appear to be correlated across series, for example the July 2021 dip occurs in SF, Boston, Colorado, and to a lesser extent NYC.

- By contrast, cities from the South show far less growth since COVID and also less volatility. Only Savannah and Miami Beach appear to have partially reverted back to pre-pandemic shares of remote work arrangements.

- Note that in this exercise, we do not re-weight the data, such that much of the variation across cities is likely to be driven by differences in occupation and industry composition. We leave as future work a mapping from our time-series measures and forcing events, such as shelter-in-place orders.

4.4 Comparing Job Advert Measurement to Survey Responses across MSAs

Our measurement of remote working utilises new job postings, which is conceptually a very different empirical object to measures of the share of employees / work days conducted in peoples homes. To understand how these different measurement approaches might relate (if at all) to one another, we utilise recent survey evidence form the American Communities Survey (ACS)[32]. Specifically, we use the (survey weighted) share of 2021 employees across Metropolitan Statistical Area's (MSAs) who report that they 'mostly work from home'. We compare this to the fraction of new job ads from each MSA which advertised remote work in 2022.

10 compares the fraction of employees who were 'mostly working from home' in the ACS with our measure. A least-squares regression line (shown in blue) has a slope coefficient of 0.36. Strictly interpreted, this suggests that *ceteris paribus* an MSA with 10% more employees who respond to the ACS that they are 'mostly working at home' would accompany a 3.5% increase in the percent of new job ads offering remote work arrangements. The overall fit of this least squares regression line is rather high, with a coefficient of determination (R^2) of 0.55. Taken in tandem, this evidence suggests that there is at least some reason to suggest our measurement approach relates to the stock of remote workers. This, along with the many advantages of working with job postings (large scale, very high granularity, long historical time series) support the use-case for our data.

4.5 Remote Work across Companies

Ultimately, the decision to advertise remote work arrangements is made by each employer who is searching for talent. By-and-large, workers value the flexibility to work some days remotely, with survey evidence estimating that a typical worker would sacrifice 6% of their salary to receive this amenity (Barrero et al. 2021). Thus, one important reason why employers have increasingly chosen to offer remote work arrangements is to attract workers. Similarly, remote work arrangements can also lessen the burden of distance and allow firms to recruit for talent in wider geographic areas. Again, this deepens the labour market and may facilitate matching with better candidates. Another reason why we see that employers are offering remote work arrangements in their vacancy listings might be due to learning.

[32]The American Community Survey (ACS) is a demographics survey program conducted by the U.S. Census Bureau. The ACS regularly gathers information previously contained only in the long form of the decennial census, such as ancestry, citizenship, educational attainment, income, language proficiency, migration, disability, employment, and housing characteristics.

Most CEO's comment that mass remote-work of staff would have been unthinkable prior to 2020, yet the forced experimentation during COVID-19 has left many with at least an indifference to such practices and at most tangible evidence of the productivity benefits these bring. Finally, firms–especially those who are expanding quickly–may see remote work arrangements as a way to including office space and energy consumption. Finally, there may simply be positive productivity impacts.[inline]Rather than listing all the positives, shouldn't we highlight that there is likely to be heterogeneity driven by a) the relevance of these attributes for the target worker population; b) firms' ability willingness to adjust their working processes to WFH?

Our analysis of employers is by no means exhaustive, and we leave for future work a more in-depth match to firm-level covariates. The first piece of analysis illustrates that the prevalence of employers who offer explicitly offer remote work arrangements in their vacancy postings varies greatly, even among same-industry firms recruiting in the same occupational category. 11 takes selected employers, and finds:

- 11: Panel A shows the share of remote work vacancies posted by four large aerospace manufacturing firms (NAICS code 3364). We consider only management occupations in this panel, and find that both Boeing and Lockheed Martin explicitly offer remote arrangements in half of their postings in 2022[33]. We further see that Northrop Grumman makes explicit offers of remote work in less than one-in-four management job vacancy postings. In contrast, SpaceX made no explicit offers for such arrangements in any of its new job listings in 2022. All of these firms explicitly offered meaningful amounts of remote work in 2019.

- 11: Panel B shows selected insurance firms who advertise vacancies for workers in the mathematical science occupations[inline]how did we choose these specific occupations?]. We see that United Health had a sizable fraction (52%) of vacancies which explicitly offered remote work, even pre-pandemic, which grew to 80% by 2022. Mutual of Omaha had a more modest pre-pandemic remote work share, but by 2022 mentions such practices in nearly all vacancies targeting mathematicians. Humana saw more than a doubling in its remote work share, but remains substantially lower than its peers.

[33]Without further analysis, we cannot say if a 50% share of remote work vacancies in 2022 results from some switch in behaviour (e.g. if firms post uniformly in time, and switch to fully remote halfway through the year, we would calculate a 50% share) or else if this is driven by some more granular cross-sectional difference between jobs with and without remote arrangements on offer. This will be addressed in future revisions.

- 11: Panel C conducts the same exercise for selected auto manufacturing firms who hire engineers. Almost no explicit offers of remote work were made in 2019. By 2022, Honda explicitly offers 1-in-2 new engineering hires the right to work remotely at least one day per week. GM offers less than half this number, and Ford less than a 6th. Tesla offers almost no remote work in either 2019 or 2022.

We next make more general observations about the share of remote work vacancies across all employers. We first see in 12 that there is no discernible relationship (depicted using a binscatter plot) between the number of vacancies posted by each employer in 2019 and the share of remote work on offer. 13 does a similar calculation in 2022, and now shows a strong positive relationship between the count of total new vacancy postings from an employer and the fraction of these postings that explicitly offer remote work arrangements.[inline]perhaps use this as first set of results before going into heterogenenity discussed above

5 Conclusion

References

Acemoglu, D., Autor, D., Hazell, J., and Restrepo, P. (2022). Artificial Intelligence and Jobs: Evidence from Online Vacancies. *Journal of Labor Economics*, 40(S1):S293–S340.

Adams-Prassl, A., Balgova, M., and Qian, M. (2020). Flexible Work Arrangements in Low Wage Jobs: Evidence from Job Vacancy Data. *SSRN Electronic Journal*.

Adams-Prassl, A., Boneva, T., Golin, M., and Rauh, C. (2022). Work that can be done from home: Evidence on variation within and across occupations and industries. *Labour Economics*, 74:102083.

Adrjan, P., Ciminelli, G., Judes, A., Koelle, M., Schwellnus, C., and Sinclair, T. (2021). Will it stay or will it go? Analysing developments in telework during COVID-19 using online job postings data.

Aksoy, C. G., Barrero, J. M., Bloom, N., Davis, S. J., Dolls, M., and Zarate, P. (2022). Working From Home Around the World.

Alipour, J. V., Falck, O., and Schüller, S. (2020). Germany's Capacities to Work from Home.

Ash, E. and Hansen, S. (2023). Text Algorithms in Economics. Unpublished Manuscript.

Bai, J. J., Brynjolfsson, E., Jin, W., Steffen, S., and Wan, C. (2021). Digital Resilience: How Work-From-Home Feasibility Affects Firm Performance.

Bajari, P., Cen, Z., Chernozhukov, V., Manukonda, M., Wang, J., Huerta, R., Li, J., Leng, L., Monokroussos, G., Vijaykunar, S., and Wan, S. (2021). Hedonic prices and quality adjusted price indices powered by AI. Working Paper CWP04/21, Cemmap.

Bamieh, O. and Ziegler, L. (2022). Are remote work options the new standard? Evidence from vacancy postings during the COVID-19 crisis. *Labour Economics*, 76:102179.

Bana, S. H. (2022). Work2vec: Using Language Models to Understand Wage Premia. Unpublished Manuscript.

Barrero, J. M., Bloom, N., and Davis, S. J. (2020). COVID-19 Is Also a Reallocation Shock.

Barrero, J. M., Bloom, N., and Davis, S. J. (2021). Why Working from Home Will Stick.

Bartik, A. W., Cullen, Z. B., Glaeser, E. L., Luca, M., and Stanton, C. T. (2020). What Jobs are Being Done at Home During the Covid-19 Crisis? Evidence from Firm-Level Surveys.

Bick, A., Blandin, A., and Mertens, K. (2022). Work from Home Before and After the COVID-19 Outbreak.

Bloom, N., Liang, J., Roberts, J., and Ying, Z. J. (2015). Does Working from Home Work? Evidence from a Chinese Experiment. *The Quarterly Journal of Economics*, 130(1):165–

218.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., 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., 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., and Amodei, D. (2020). Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Brynjolfsson, E., Horton, J. J., Ozimek, A., Rock, D., Sharma, G., and TuYe, H.-Y. (2020). COVID-19 and Remote Work: An Early Look at US Data.

Burke, M., Sasser Modestino, A., Sadighi, S., Sederberg, R., and Taska, B. (2020). No Longer Qualified? Changes in the Supply and Demand for Skills within Occupations. Federal Reserve Bank of Boston Research Department Working Papers, Federal Reserve Bank of Boston.

Criscuolo, C., Gal, P., Leidecker, T., Losma, F., and Nicoletti, G. (2021). The role of telework for productivity during and post-COVID-19. (31).

Davis, S. J., Faberman, R. J., and Haltiwanger, J. C. (2012). Recruiting Intensity during and after the Great Recession: National and Industry Evidence. *American Economic Review*, 102(3):584–588.

Davis, S. J. and Samaniego de la Parra, B. (2020). Application Flows.

del Rio-Chanona, R. M., Mealy, P., Pichler, A., Lafond, F., and Farmer, J. D. (2020). Supply and demand shocks in the COVID-19 pandemic: An industry and occupation perspective. *Oxford Review of Economic Policy*, 36(Supplement_1):S94–S137.

Deming, D. and Kahn, L. B. (2018). Skill Requirements across Firms and Labor Markets: Evidence from Job Postings for Professionals. *Journal of Labor Economics*, 36(S1):S337–S369.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019a). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019b). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dingel, J. I. and Neiman, B. (2020). How many jobs can be done at home? *Journal of Public Economics*, 189:104235.

Draca, M., Duchini, E., Rathelot, R., Turrell, A., and Vattuone, G. (2022). Revolution in Progress? The Rise of Remote Work in the UK.

Forsythe, E., Kahn, L. B., Lange, F., and Wiczer, D. (2020). Labor demand in the time of COVID-19: Evidence from vacancy postings and UI claims. *Journal of Public Economics*, 189:104238.

Hershbein, B. and Kahn, L. B. (2018). Do Recessions Accelerate Routine-Biased Technological Change? Evidence from Vacancy Postings. *American Economic Review*, 108(7):1737–1772.

Luktevich, B. (2022). BERT Language Model. https://www.techtarget.com/searchenterpriseai/definition/E language-model.

Marinescu, I. and Wolthoff, R. (2020). Opening the Black Box of the Matching Function: The Power of Words. *Journal of Labor Economics*, 38(2):535–568.

Mas, A. and Pallais, A. (2017). Valuing Alternative Work Arrangements. *American Economic Review*, 107(12):3722–3759.

Modestino, A. S., Shoag, D., and Ballance, J. (2016). Downskilling: Changes in employer skill requirements over the business cycle. *Labour Economics*, 41:333–347.

Mongey, S., Pilossoph, L., and Weinberg, A. (2021). Which workers bear the burden of social distancing? *The Journal of Economic Inequality*, 19(3):509–526.

Ozimek, A. (2020). The Future of Remote Work.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2020). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv:1910.01108 [cs]*.

Shapiro, A. H., Sudhof, M., and Wilson, D. J. (2022). Measuring news sentiment. *Journal of Econometrics*, 228(2):221–243.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Wiswall, M. and Zafar, B. (2018). Preference for the Workplace, Investment in Human Capital, and Gender*. *The Quarterly Journal of Economics*, 133(1):457–507.

Table 1: Sample Size of Online Job Vacancy Postings

(1) Country	(2) Vacancies	(3) Employers	(4) Cities
New Zealand	1,661,036	34,808	67
Australia	8,417,562	191,659	59
Canada	11,240,014	694,456	3,684
United Kingdom	72,636,221	850,958	2,267
United States	155,005,465	3,411,302	31,618
Total	248,960,298	5,183,183	37,695

Note: This table reports the size of our corpus of vacancy postings, which spans from January 2014 to November 2022. For the period 2014-2018 we use a 5% subsample, drawn uniformly across countries and months. For the period 2019-2022 we use the full set of new job vacancy postings. The resulting numbers are pooled from these two periods. Employer names and cities come pre-extracted from our data-provider, using a proprietary algorithm which finds city and employer names in the raw text, and does very mild clustering on these names. Numbers reported denote the final sample, after a few data-cleaning steps which drop approximately 6% of job ads from the raw sample. All cleaning steps are documented in text, and a discussion of the motivations for each step is found in Appendix 6.

Table 2: Simple Dictionary Methods Generate Clear Classification Errors when Applied to Vacancy Postings

False-Positive Examples:	False-Negative Examples:
We are looking for a Deputy Home Manager with domiciliary care experience to join our company. You will **work from home** care facilities with a strong track record of quality service.	We encourage our people to explore new ways of working - including part-time, job-share or working from different kinds of locations, including their home. Everyone can ask about it.
Schedule: * 10 Hour Shift * 8 Hour Shift **Work remotely**: * No	With a hybrid mix of time at home as well as our corporate office, this role will suit an analytical, process orientated and people focused payroll professional who thrives in a fast-paced environment.
Applicants must also have: * Ability to work as part of a team, in a fast paced environment * Experience in a 4 or 5 star hotel * Previous experience working in **remote** locations	We see the value in work-life balance, so whether you like to get a surf in before work, like to head home in time to pick up the kids or you just like working from the comfort of your own home now and then, we want to support you.
You may work on renovation projects, store reorganizations, new store openings, and store closings. May respond to managerial or **Home Office** requests for special reports, information, or for help on special projects.	The interviews for this role are likely to be conducted remotely using Microsoft Teams or Zoom. It is also expected that relevant work within these roles may be done remotely, within the UK.

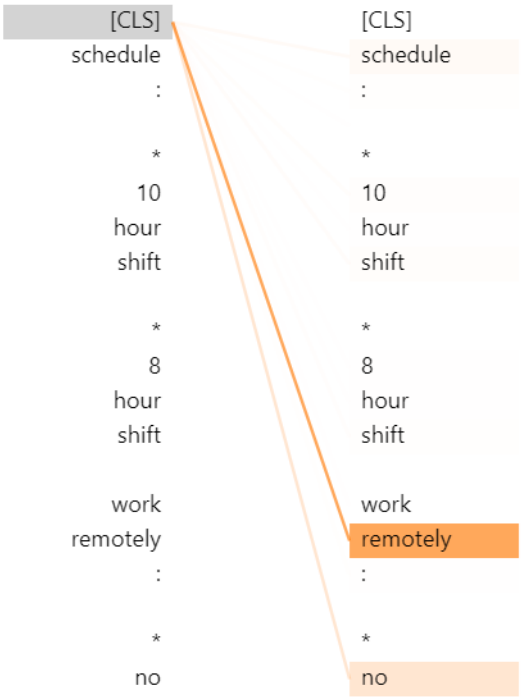
Note: These examples are based on real text sequences from job ads. The light-yellow highlighting provides the relevant context for correct classification of remote work arrangements. We use a bold font-face to show which keywords a dictionary approach to classification would tag. The left column shows 'False-Positive' examples, where the highlighted context suggests these jobs do not offer remote work, despite the presence of bold dictionary keywords. The right column shows 'false-negative' examples, where the highlighted context is suggestive of remote work options, despite no bold keyword hits.

Table 3: Our WHAM Baseline Model Outperforms All Other Methods Across Multiple Classification Metrics

	(1) Error Rate $\frac{FP+FN}{TP+TN+FP+FN}$	(2) TP Rate $\frac{TP}{TP+FN}$	(3) TN Rate $\frac{TN}{TN+FP}$	(4) FP Rate $\frac{FP}{FP+TN}$	(5) FN Rate $\frac{FN}{FN+TP}$	(6) Precision $\frac{TP}{TP+FP}$	(7) F1 Score $2\frac{\text{Precision}*\text{TP Rate}}{\text{Precision}+\text{TP Rate}}$
All Zero	.28	.00	1.00	.00	1.00	.00	.00
Dictionary	.16	.81	.85	.15	.19	.68	.74
Dictionary w/ Negation	.12	.74	.94	.06	.26	.82	.78
Logistic Regression	.11	.81	.93	.07	.19	.81	.81
Logistic Regression w/ Negation	.08	.83	.95	.05	.17	.87	.85
GPT-3	.06	.92	.94	.06	.08	.87	.89
WHAM (Generic English)	.03	.96	.98	.02	.04	.95	.95
WHAM (Baseline)	.02	.97	.99	.01	.03	.97	.97

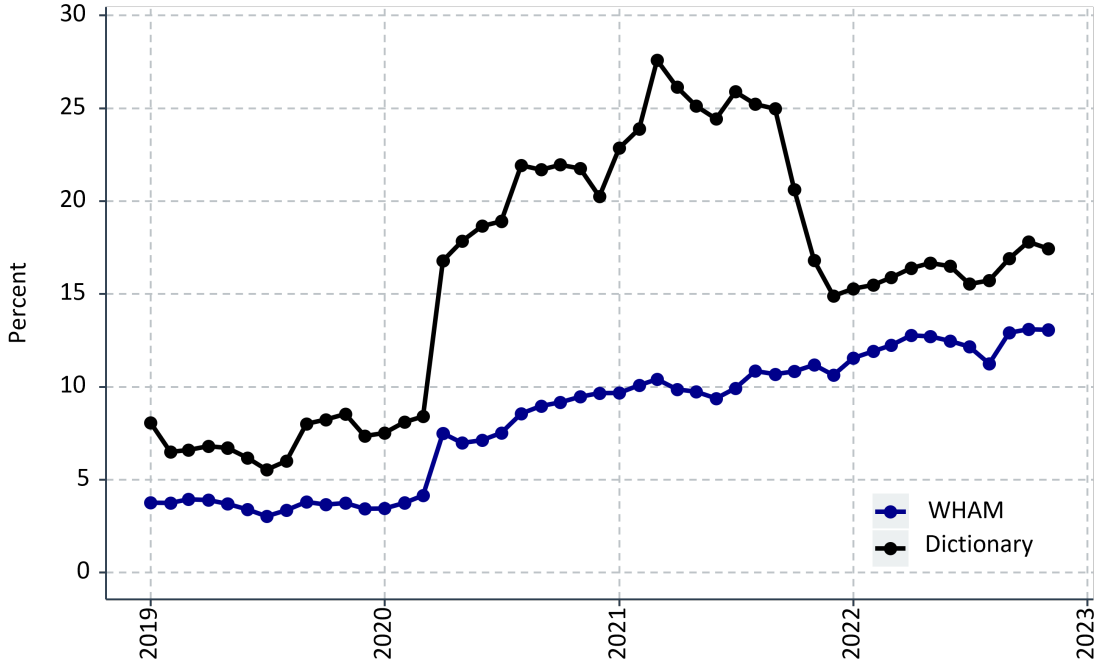
Note: This table reports the performance of alternative classification algorithms for detecting remote work in a held-out test set of 4,050 unique text sequences. Each test set observation is assigned a true value of 0 or 1 according to the majority vote of human labelers, and a predicted value of 0 or 1 according to the particular algorithm. FP is the number of false positives, e.g. the number of observations with a true 0 and predicted 1. Similarly, FN is the number of false negatives, TP the number of true positives, and TN the number of true negatives. All performance measures are built from these four counts. A description of each algorithm is provided in Appendix 8.

Figure 1: Example of Attention Weights for Remote Work Prediction



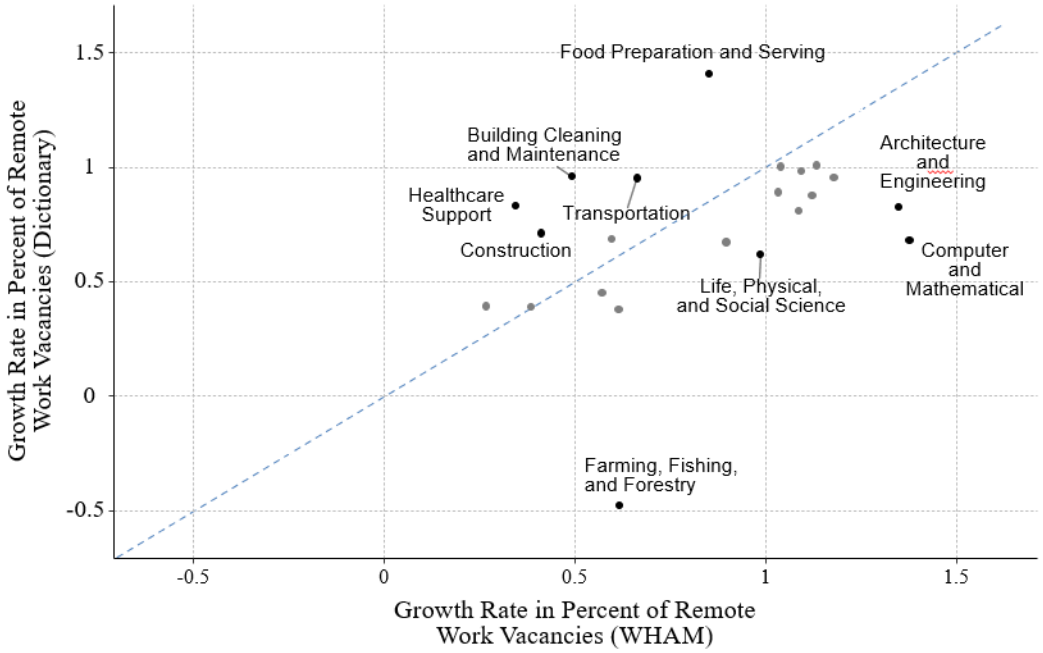
Note: This figure illustrates the attention weights estimated for predicting whether a text sequence extracted from an online job posting offers remote work. The original text underlying this example is "schedule: (newline) * 10 hour shift (newline) * 8 hour shift (newline) work remotely: (newline) * no" which is coded by humans as not offering remote work. The prediction model prepends the text with a [CLS] vector that feeds into a logistic function for classification. The shading in the right column indicates which terms are given more weight in the prediction problem.

Figure 2: Share of Remote Work job ads using "Dictionary" measurement deviates substantially from our WHAM measurement, using U.S. Data



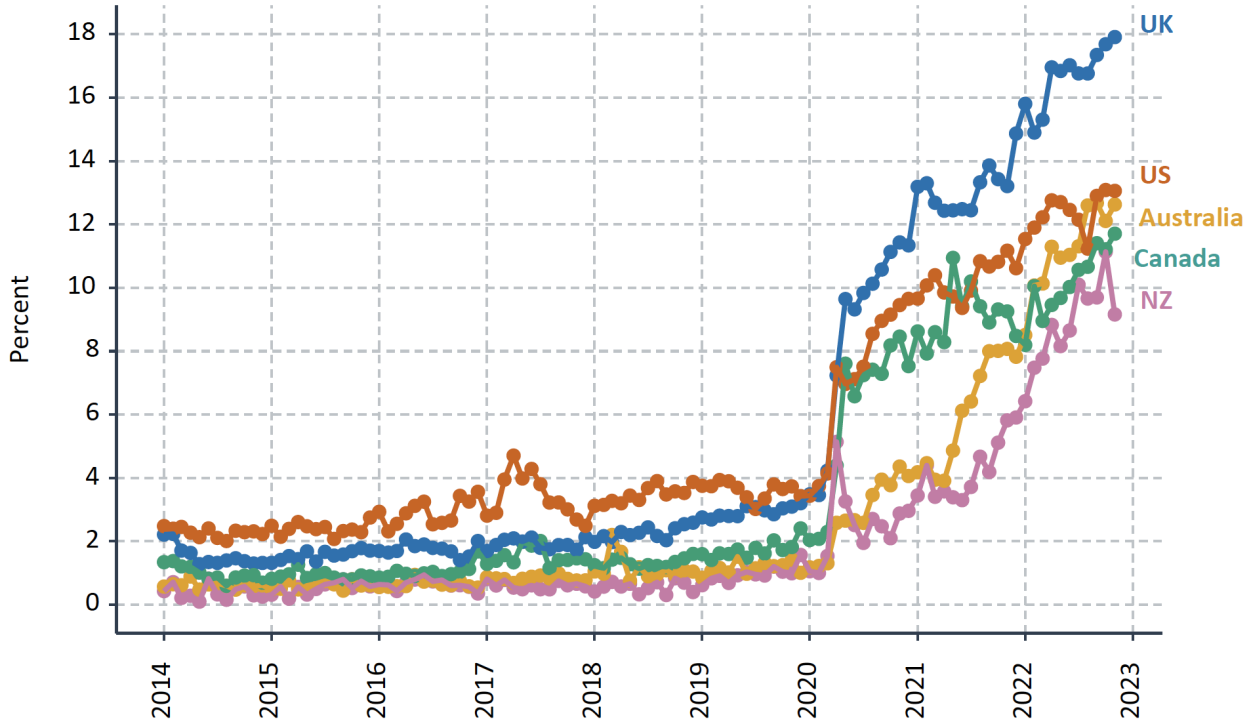
Note: This figure shows the share of vacancy postings that say the job allows one or more remote workdays per week. We compute the "Dictionary" series using the method of keyword search based on a dictionary of keywords in Adrjan et al. (2021). The "WHAM" series calculates these shares using our measurement approach, based on our WHAM algorithm trained on over 10,000 human labels and utilizing a large language model. We compute the share of all new vacancy postings from the universe of postings from the United States. We reweight data in each month so that it matches the 2019 distribution of occupations.

Figure 3: Share of Remote Work job ads using "Dictionary" measurement deviates substantially from our WHAM measurement, using U.S. Data



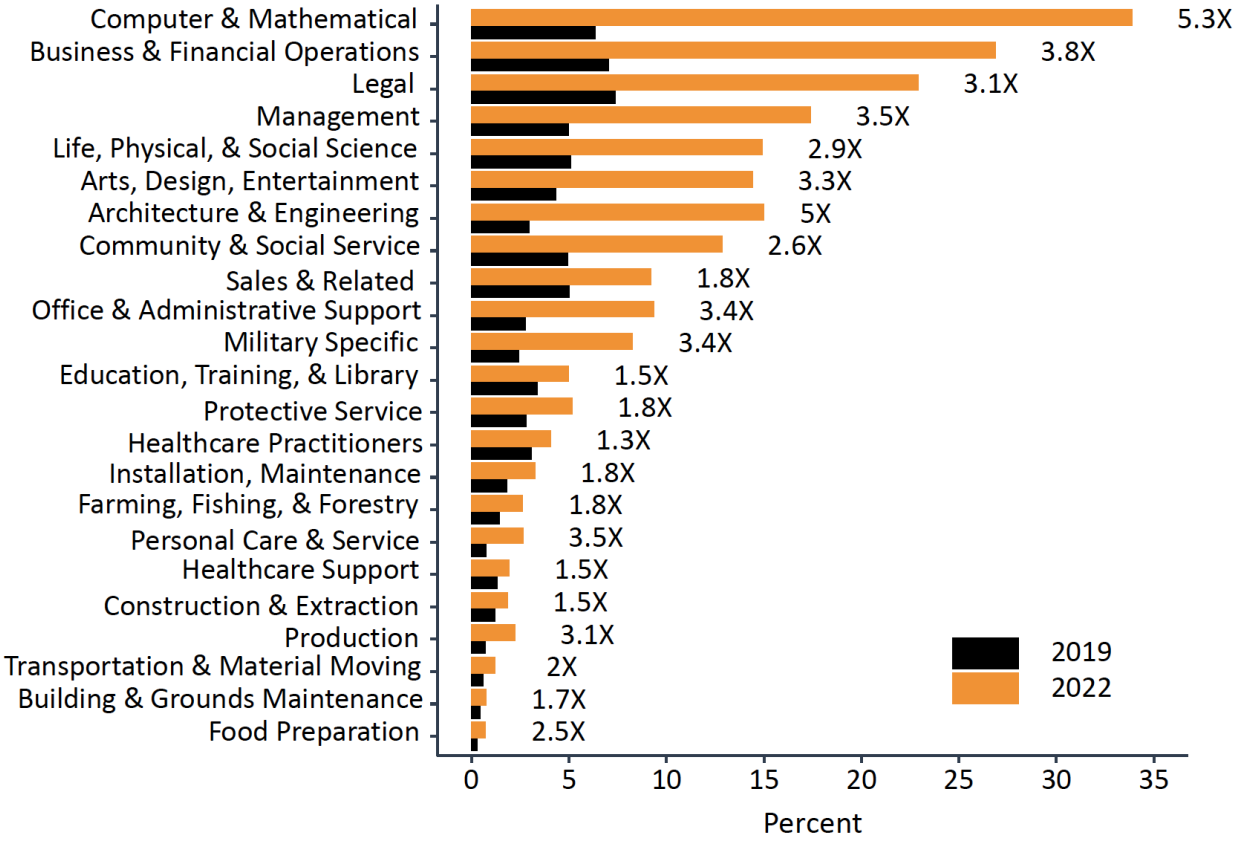
Note: This figure shows the share of vacancy postings in each 2-digit SOC occupation group which are classified as offering remote work arrangements. The y-axis shows the growth rate from 2019 to 2022, based on measurement using a "Dictionary" measurement approach. The set of keywords in our dictionary is based on those used in Adrjan et al. (2021). The x-axis does the same calculation, based on our "WHAM" algorithimic measurement approach, which is trained on over 10,000 human labels. Our change calculations comes from DHS (add ref), which calculates change = (x_post - x_pre) / 0.5 * (x_post + x_pre). The blue-dashed line shows a 45 degree line.

Figure 4: The Share of Vacancy Postings that Explicitly Offer Remote Work Rose Sharply in All Five Countries from 2020



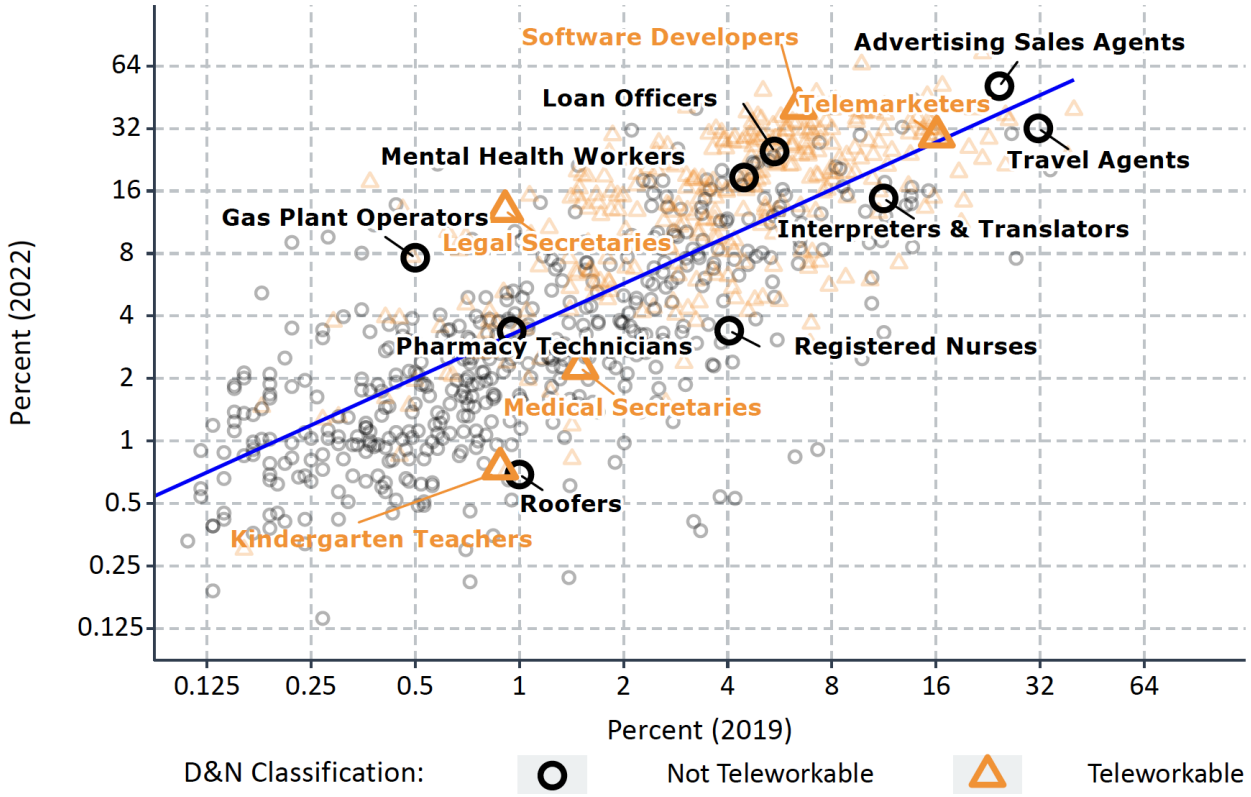
Note: This figure shows the share of vacancy postings that say the job allows one or more remote workdays per week. We compute these monthly, country-level shares as the weighted mean of the own-country occupation-level shares, with weights given by the U.S. vacancy distribution in 2019. Our occupation-level granularity is roughly equivalent to six-digit SOC codes. See Appendix XXX for the corresponding raw series and series based on alternative weighting schemes.

Figure 5: Remote-Work Posting Shares Are Highest in Professional, Scientific and Computer-Related Occupations, U.S. Data



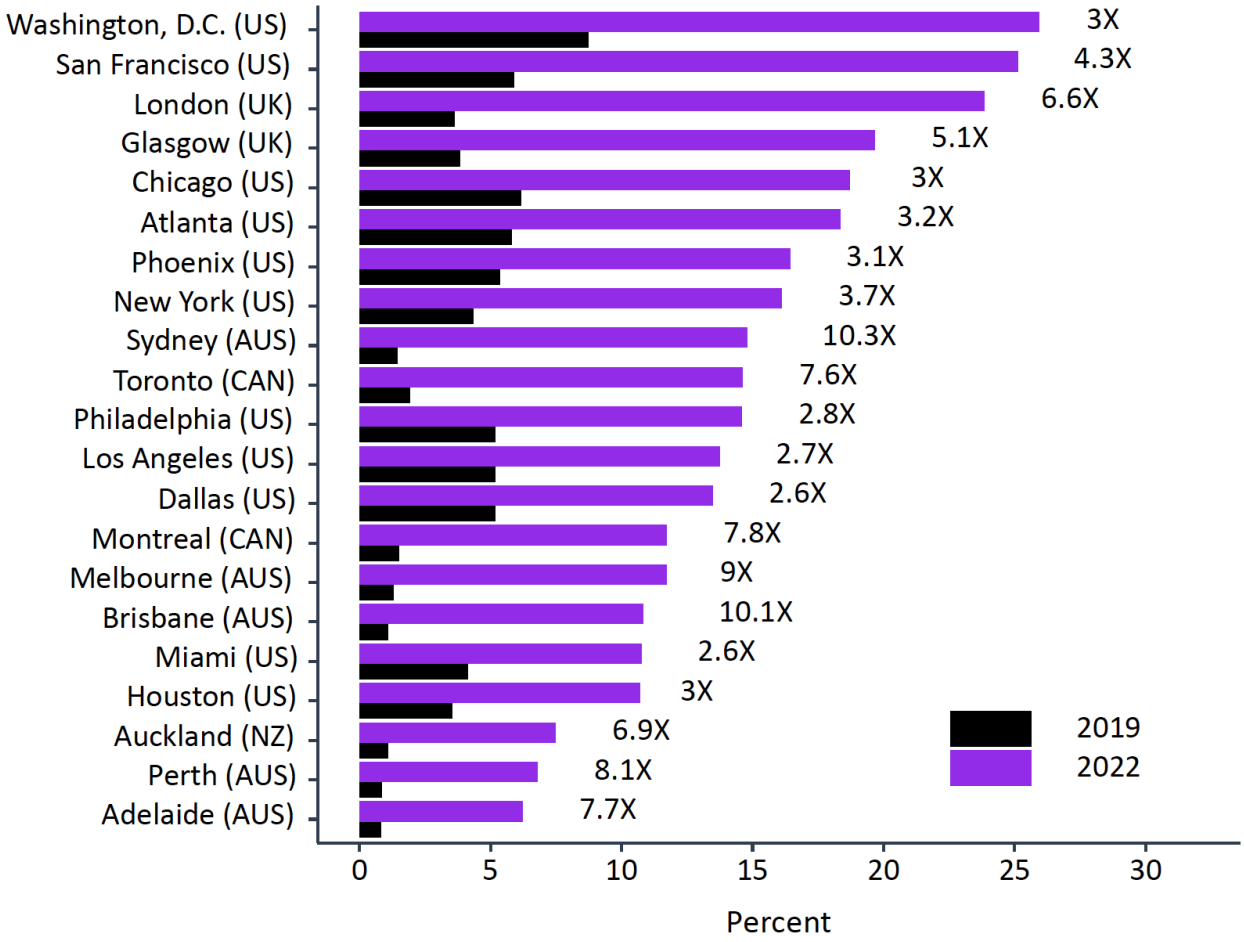
Note: Each bar reports the share of vacancy postings that say the job allows one or more remote workdays per week in the indicated period and occupation group (two-digit SOC). The "2022" values reflect data from December 2021 to November 2022.

Figure 6: Remote-Work Posting Shares Rose In Almost Every Occupation, U.S. Data



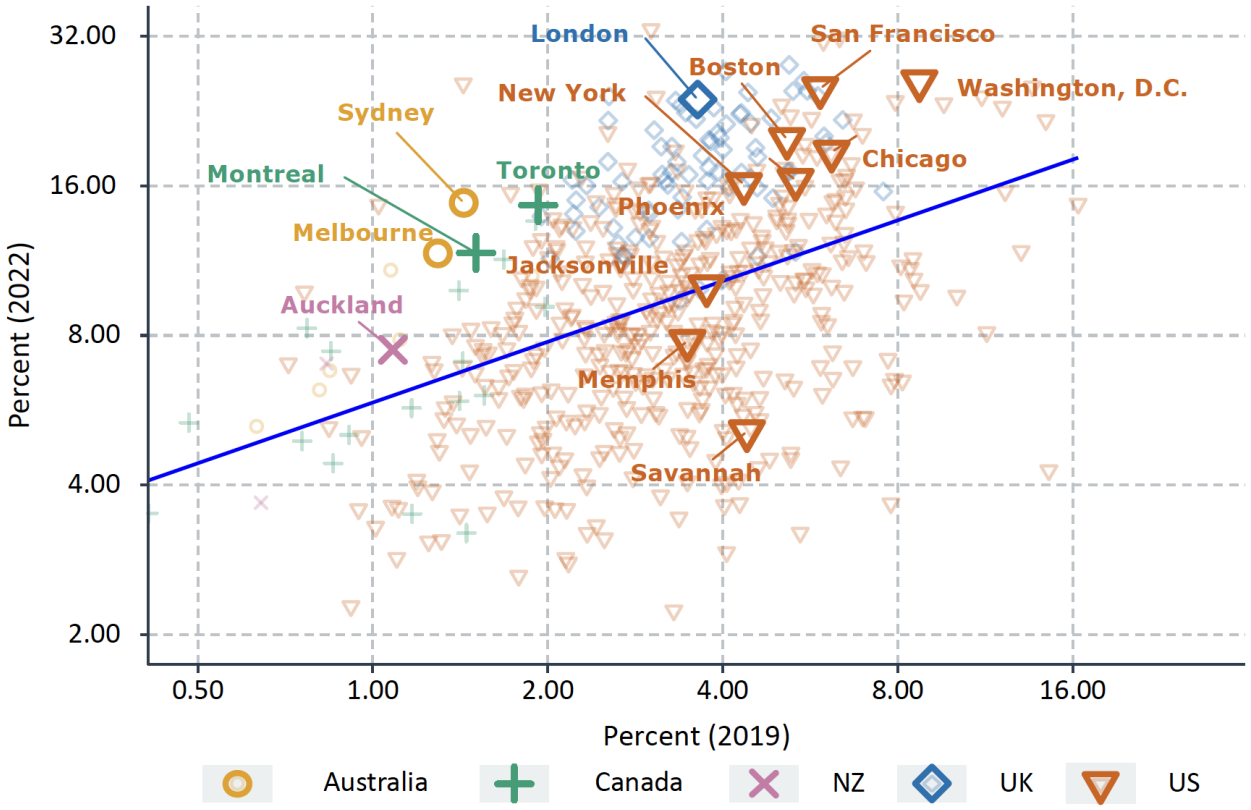
Note: This figure plots the percent of postings that say the job allows one or more remote workdays per week for 875 occupations in 2019 and 2022 (December 2021 to November 2022). We define occupations by ONET codes, dropping those with fewer than 250 posts in 2019. The line shows the unweighted OLS fit: $log(y) = 1.22 + 0.75log(x)$, which has an R^2 value of 0.63.

Figure 7: Remote-Work Posting Shares Vary Widely across Major Cities



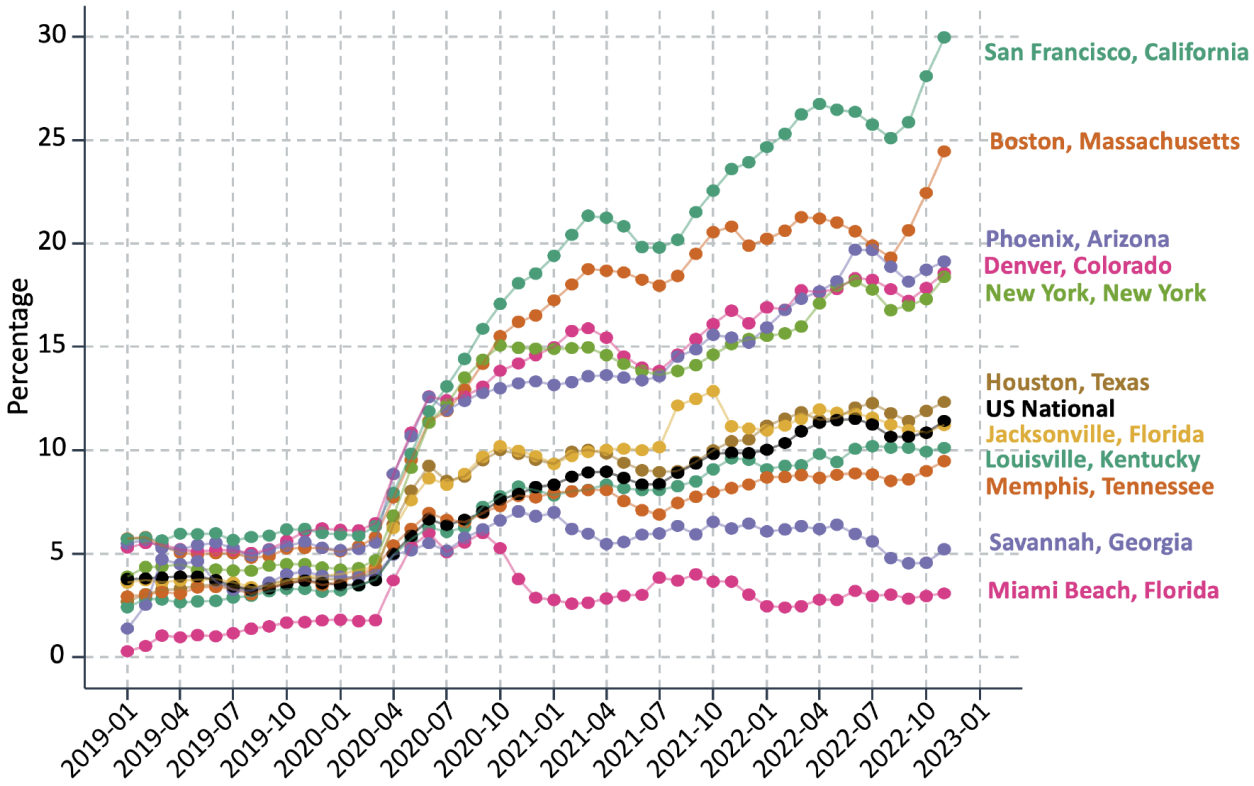
Note: This figure reports the percent of postings that say the job allows one or more remote workdays per week for selected cities in 2019 and 2022 (December 2021 to November 2022). "City" refers to the location of the establishment or firm that is hiring.

Figure 8: Remote-Work Postings Grew at Different Rates across Cities since the Pandemic



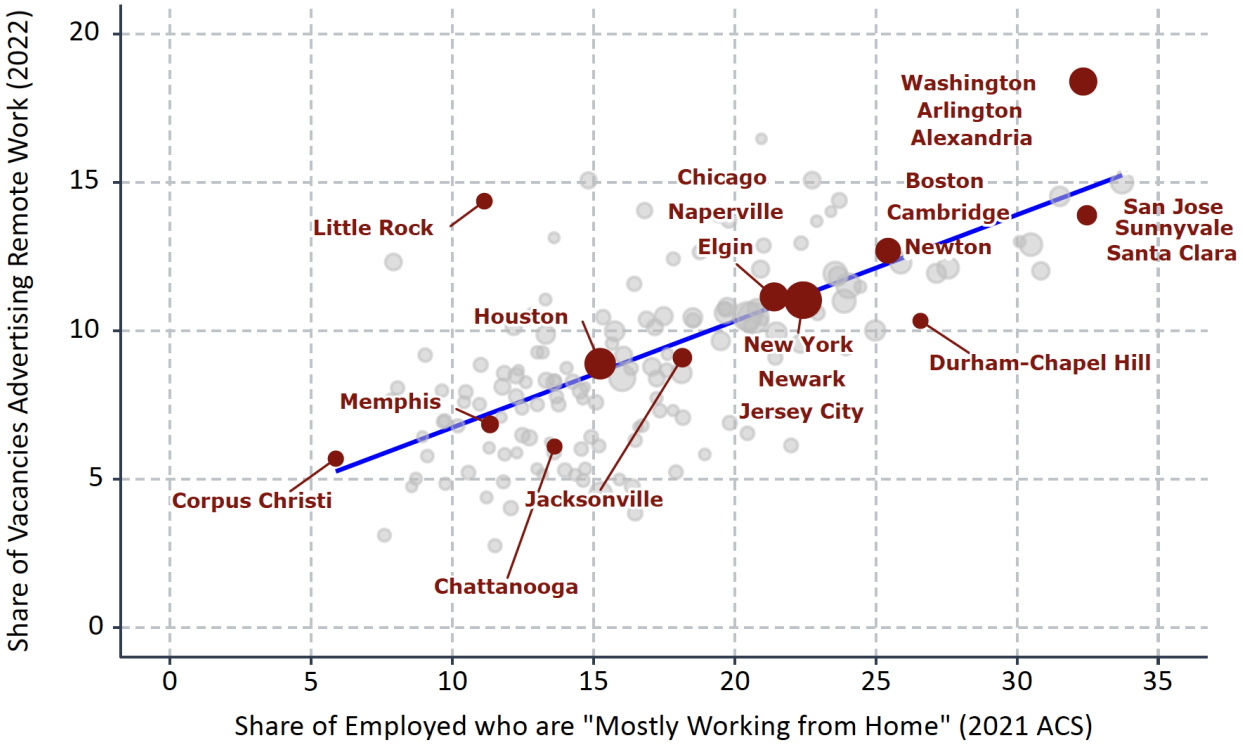
Note: This figure plots the city-level percent of postings that say the job allows one or more remote workdays per week in 2019 and 2022 (December 2021 to November 2022). "City" refers to the location of the establishment or firm that is hiring. The line shows the unweighted OLS fit: $log(y) = 1.77 + 0.42log(x)$, which has an R^2 value of 0.19.

Figure 9: Remote-Work Posting Shares Vary Across US Cities



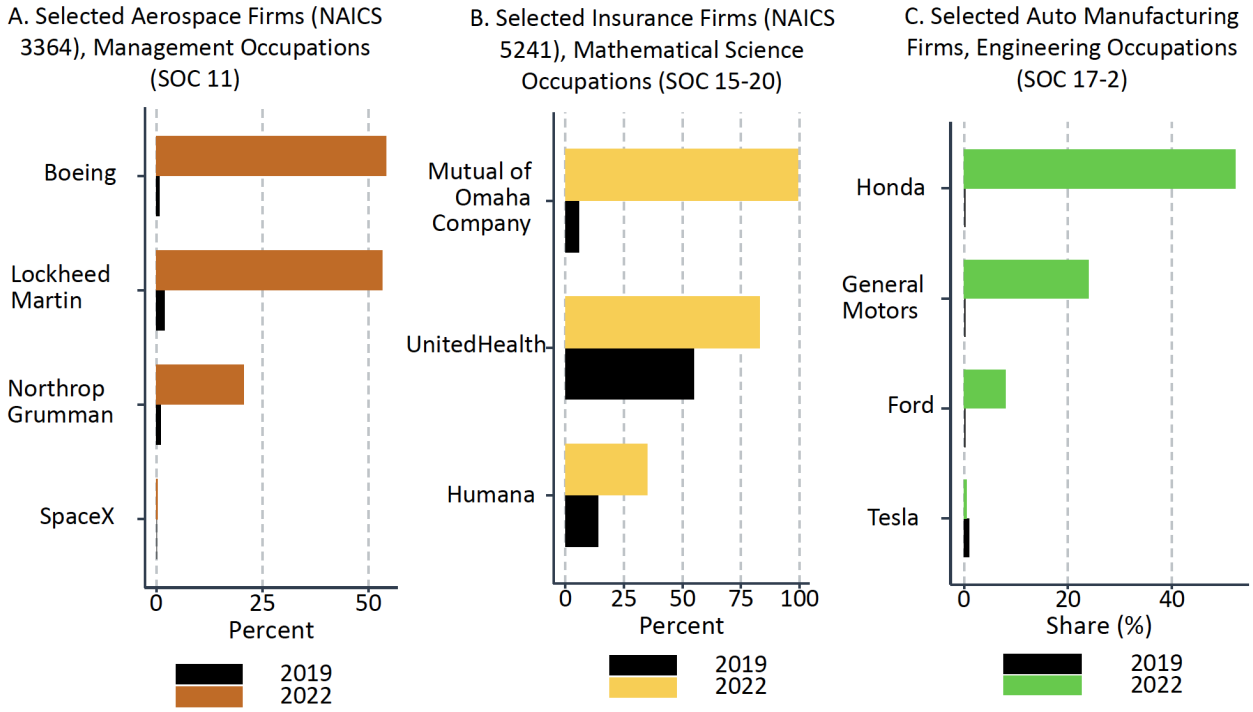
Note: We calculate the monthly share of all new job vacancy postings which explicitly advertise remote working arrangements, by selected cities. Prior to aggregation at the monthly level, we employ a jackknife filter to remove a small number of outlier days (see Appendix XXX: Data for further details). This figure shows the 3-month moving average. Cities chosen above are selected examples to illustrate the wide cross-city spread.

Figure 10: Remote-Work Posting Shares Compared to the Share of Employed Who Are "Mostly Working from Home", U.S. Metropolitan Statistical Areas



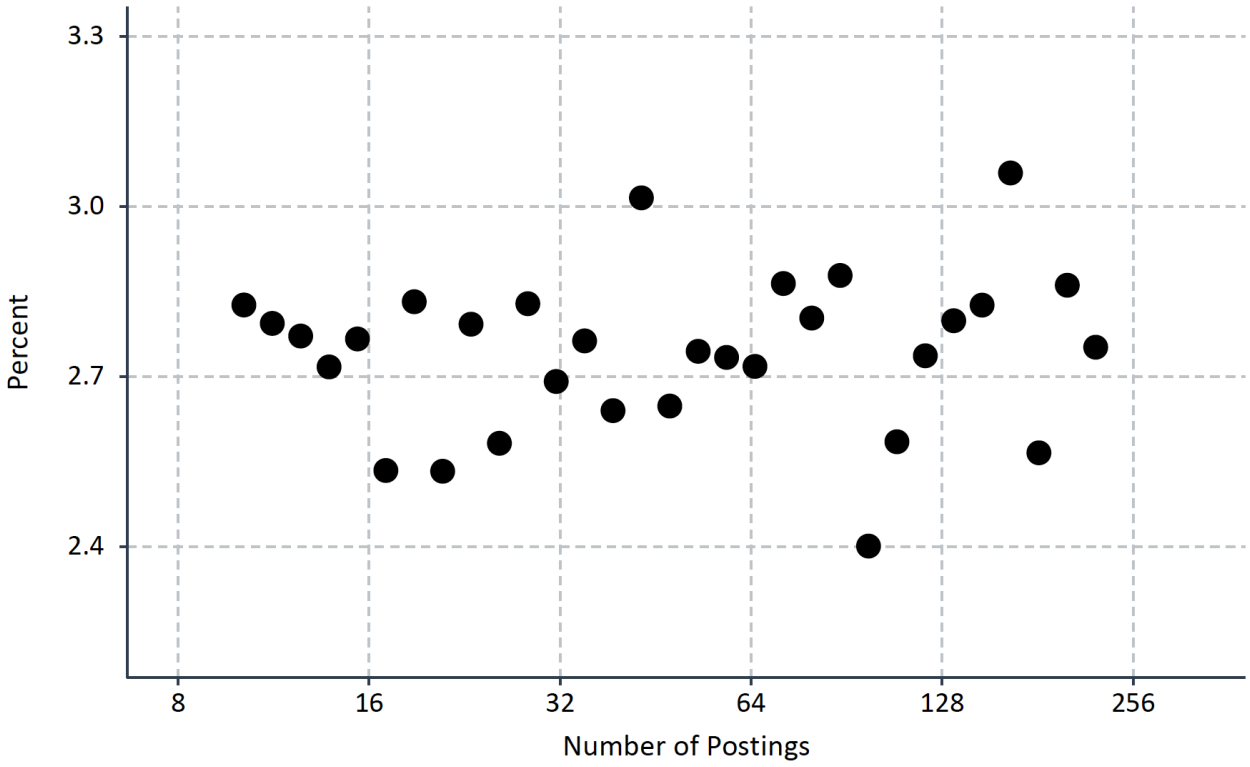
Note: The vertical scale is the percent of postings in 2022 (December 2021 to November 2022) that say the job allows one or more remote workdays per week. The horizontal scale is the percent of employees who say they 'mostly worked from home' in 2021 in the American Communities Survey (ACS), using ACS sampling weights. The line shows the unweighted OLS fit: $log(y) = 3.16 + 0.36log(x)$, which has an R^2 value of 0.55. The regression includes one observation that is outside the plotted axes.

Figure 11: The Prevalence of Postings that Allow Remote Work Varies Greatly, Even among Same-Industry Firms Recruiting in the Same Occupational Category



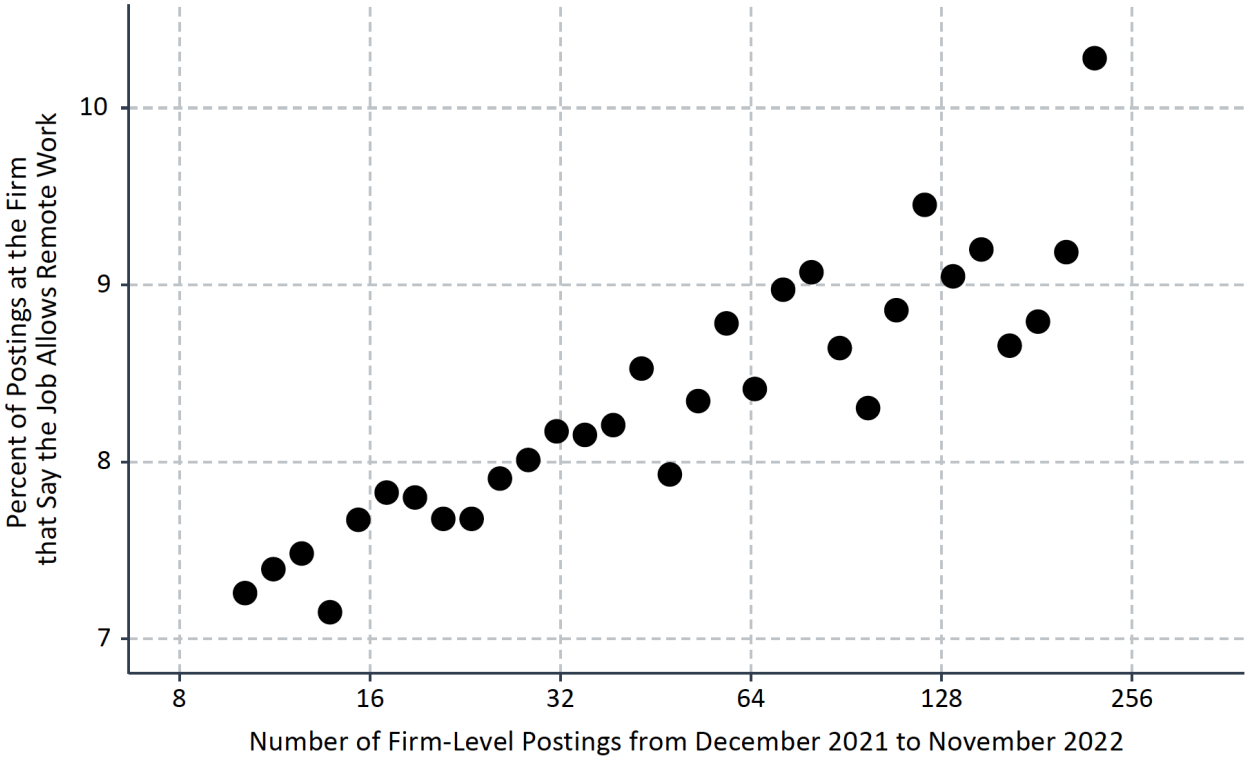
Note: For each firm, year and indicated occupations, we report the percent of U.S. postings that say the job allows one or more remote workdays per week. Data for "2022" cover the period from December 2021 to November 2022.

Figure 12: No clear relationship between an employer's number of new job vacancy postings and employer's percent of WFH vacancies in 2019



Note: Underlying data comes from employers, which summarised in a binscatter above. The y-axis shows our measurement of the share of all new job vacancy postings which explicitly advertise remote working arrangements, calculated for each employer with 10 or more vacancies during 2019. The x-axis shows the number of new vacancy postings for each employer, during the period 2021 Q3 to 2022 Q2, inclusive. The large circles depict a bin-scatter across 30 bins. Standard errors are present, but not discernible.

Figure 13: Employers with High Remote-Work Posting Shares Also Have More New Job Openings, U.S. Data for 2022



Note: To construct this chart, we consider all employers that posted 10 or more U.S. job vacancies in 2022 (December 2021 to November 2022) and sort these employer-level observations into 30 bins defined by their total number of U.S. vacancy postings in 2022. For each bin, we then average over employers to compute the reported bin-level means.

ONLINE APPENDIX

6 Data Appendix

In this Appendix we provide further commentary on the corpus of online job vacancy postings.

6.1 Data Provider

Our corpus of online job vacancy postings is provided by the labour market and analytics company 'Lightcast' (formerly Emsi Burning Glass). Lightcast has been scraping online job vacancy postings in the USA since 2007, and has continued to expand to other countries.

6.2 Web Sources

Each job vacancy posting is scraped by Lightcast from the internet. Specifically, the company scrapes over 50,000 web sources. These sources include private online job vacancy aggregators (e.g. Indeed.com, Monster.com), public online job boards (e.g. New York City Department of Labour's 'JobZone'), and employers' own recruitment web pages (e.g. careers.microsoft.com, usajobs.gov). Lightcast actively audits their list of web sources to ensure data from new websites is on-boarded in a timely manor[34]. One of the main competitive advantages of Lightcast's data product is the breadth of their sources. These data are often referred to in the literature as the 'near universe' of online job vacancy postings.

6.3 What's in the job vacancy posting data?

[inline]should we include a description of the work done by Kelsey to validate firm level ads? Once an online job vacancy posting is scraped, Lightcast processes this data to produce three categories of information: (i) plain text, (ii) meta data, and (iii) structured data. A description of each of these categories follows presently:

6.3.1 Plain Text

The plain text of each job ad contains the full textual description of the job, as written by employers. To construct this, Lightcast takes the HTML file scraped from a given website and does two further processing steps. First, it parses out portions of the HTML file which do not contain information about the vacancy (e.g. removing website headers, footers, and

[34] One reason we eschew analysis of the count of postings and instead focus on shares is that the underlying donor pool of online sources is constantly changing.

side-menu bars). Second, Lightcast takes this portion of HTML which (ideally) contains only information about the job vacancy, and turns it HTML into plain-text.

6.3.2 Meta Data

Each vacancy posting also contains a number of meta-data items. These are immutable properties of each web scraped vacancy. The most important of these is the date the page was scraped. Another important piece of meta-data is the URL from which the posting was scraped.

6.3.3 Structured Data

The most commonly used data product that Lightcast creates is the set of structured data. This dataset contains one row for each job vacancy posting, and a large number of additional information such as the job title, occupation, salary, educational requirements, location, and employer name. These variables are extracted using Lightcast's own proprietary algorithms. These fields differ from meta data because they may contain missing values and/or measurement error due to imperfect algorithmic extraction.

6.4 Errors and Missing Information

Overall, the data product is a highly informative and accurate product. We view the incidence of errors as very minute, but acknowledge that any dataset with hundreds of millions of observations scraped from over 50,000 sources will never be perfect. Both the structured data and the plain text data require a number of pre-processing steps and the use of algorithmic feature extraction, which in a very small number of cases produce errors (e.g. misclassification of occupations, truncation of plain text, presence of erroneous text). In this subsection we highlight some of the errors we have encountered, and discuss the strategies we employed to ensure our results remain robust to such issues.

6.4.1 Missing Values

A specific value (e.g. the educational requirement for a job) might be missing for at least two reasons: (i) the employer does not mention this explicitly in the text of the job ad, and (ii) the algorithm used to extract this feature from the text failed. The former issue is especially problematic in the context of educational requirements (e.g. we see that very few vacancies for Medical Doctors explicitly mention a requirement to have gone to medical school). This is because certain features of the job will likely be taken as given (for example,

specialized degrees for medical doctors). We also see that a large share of vacancy postings do not list the salary (this is almost entirely due to lack of information, and not poor feature extraction). One could employ imputation methods to address these missing values (see Bana (2022), who predict the salary with a very high degree of accuracy from the text). The main strategy employed in this paper was to only utilise covariates which contain fewer missing values, such as occupation classifications and location information.

6.4.2 Erroneous Plain Text

In a very small number of cases we observe that the plain text includes some parts of the website other than the job description. For example, the plain text from one job board in New Zealand included a number of vacancy posting text from ads that were being cross-promoted to the browser, essentially turning each document into a compilation of six job ads.[inline]what do we do in these cases?

6.4.3 Truncated Plain Text

In a small number of other cases, the plain text is truncated. For example, we found one employer who listed each jobs location using an interactive link which must be clicked to appear. Since the web scraper only parses static information, this portion of the job ad was missing from the plain text. We conducted extensive tests, and stress that in the vast majority of cases the plain text provides an accurate representation of the job vacancy posting.

6.5 Checking for Correlated Measurement Error

As discussed above, since our measurement of remote working relies on the underlying plain text, some measurement error is inevitable. One concern we took seriously is the possibility that this noise may be correlated within online sources. We discuss our approach to addressing this below.

6.5.1 Validation of Large Job Boards

In some instances the pre-processing of job posting websites includes additional erroneous text. When this occurs, it is very likely to be true for all job postings scraped from the same web source. To ensure our results are not overly sensitive to such issues, we first identify the twenty largest web sources for each country. We then create twenty versions of country-level time series of monthly remote work vacancy shares, leaving one job board out at a time.

This process revealed one problematic source from each of Canada, USA, NZ and UK. We found two problematic job boards in Canada. 6.1 reports the fraction of total job ads that were removed after dropping postings from these sources.

Table 6.1: Web Sources Dropped from Sample

Source:	Country:	Share:
A	USA	6.7%
B	UK	3.6%
C	NZ	28.9%
D	Canada	3.9%
E	Canada	3.5%

Note: We do not identify these job boards, to avoid any potential conflict with the commercial interests of these websites. Researchers should reach out to us if they would like to know the names of each source which we drop. The "share" column reports the fraction of all postings from Jan 2014 to July 2022, within the same country, that is removed from the raw data after we dropped the corresponding source.

6.5.2 Outlier Detection and the Jack-Knife Filter

When we present monthly time series data, we apply an algorithm which filters outlier days whose contribution to the over-all monthly share of vacancy postings offering remote work is at odds with other days in a given month. This filter has a very minimal impact on the results (e.g. we drop less than a quarter of one percent of job postings from the US based on this filter). The few outlier days we do filter out occur when a large number of vacancies get posted on a single day which are concentrated by employer/occupation/web source. Our extensive audits of the data reveal that outlier days are due to compositional discontinuities at the daily frequency, and not caused by measurement error in our algorithm. Our filter is based on the Jack-Knife resampling procedure, and works as follows:

- For a given calendar month M denote S_M as the share of vacancy postings which offer remote work
- For each day $t \in M$, compute the share of remote work postings *excluding* all postings on this focal day t from the calculation. Define this share as $S_{M\setminus\{t\}}$
- If the absolute level deviation between S_M and $S_{M\setminus\{t\}}$ is greater than 2 percentage points, or else if the absolute ratio of their natural logarithms is greater than 0.1, then we classify focal day t as an outlier

- Recalculate the share of vacancy postings for month M excluding all postings on outlier days

This filter alters the data minimally. For example, in the United States, it removes 0.2% of the total number of vacancy postings. The number of postings which are filtered is shown in the below table:

Table 6.2: Jack-Knife Time Series Filter

Country	Fraction of Postings Filtered (%)
NZ	7.74
Australia	0.78
Canada	1.43
United Kingdom	0.06
United States	0.21

6.6 Representativeness of Online Job Vacancy Postings

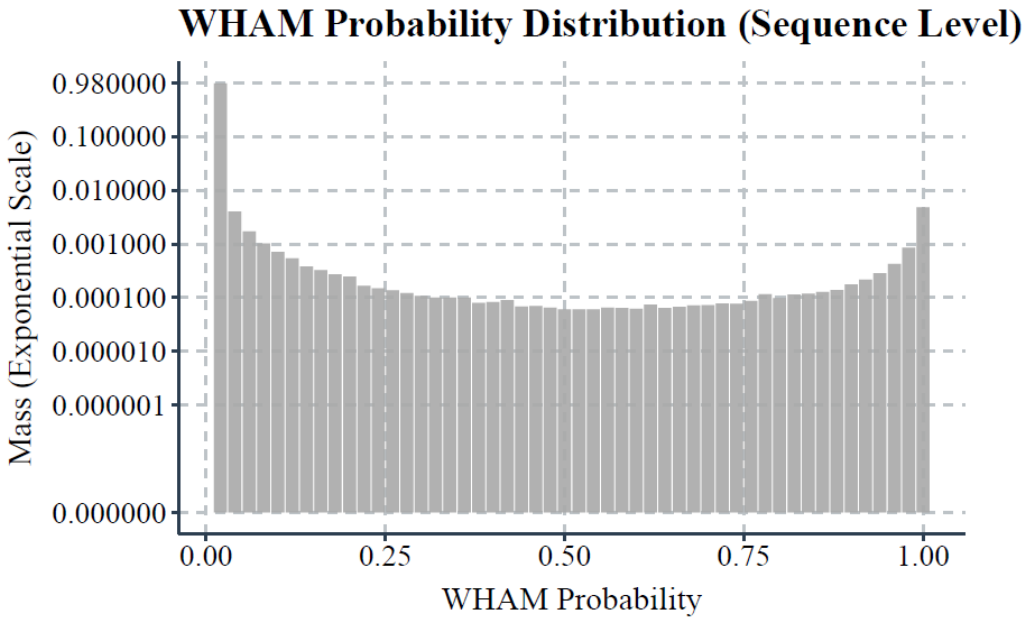
Lightcast frequently reviews the representativeness of the job vacancy postings it scrapes, to ensure the information renders an accurate picture of the overall labour market. Both our analysis and that of our data provider, as well as many other papers in the literature who utilise these data, all find a high degree of fidelity between the share of job vacancies across occupations and industries, and other official Government data products which measure similar phenomena.

In our baseline results, we also re-weight the data to reduce sensitivity to shifts in the overall composition of the labour market. The next section discusses this further, but we note that this provides additional robustness to concerns of representativeness

7 Supplementary Results

7.1 Additional results for Section 2

Figure 7.1: Most Sequences are Assigned a Predicted Probability by WHAM at Extreme Values



Note: We assign a predicted probability to each sequence in the full job posting dataset using our trained neural network model. This figure presents a histogram of the number of sequences that fall in different bins according to these predictions.

7.2 Additional results for Section 3

Table 7.3: Most Job Postings Either Have Zero or One Sequence that Explicitly Offers Remote Work

(1) Positive Sequences	(2) Number of Job Ads	(3) Share of Total (%)
0	40,006,052	90.42
1	2,682,844	6.06
2	989,084	2.23
3	365,970	0.83
> 3	201,523	0.46

Note: This table tabulates how many text sequences in each US job posting from 2021 are classified as offering remote work according to WHAM. A typical job ad is split into six sequences. Most postings (90.42%) have no positive sequences. Of the remaining fraction, most have only one positive sequence.

Table 7.4: Remote Work Classification Performance using Human Labels, by Algorithm (Weighted)

	(1) Error Rate $\frac{FP+FN}{TP+TN+FP+FN}$	(2) TP Rate $\frac{TP}{TP+FN}$	(3) TN Rate $\frac{TN}{TN+FP}$	(4) FP Rate $\frac{FP}{FP+TN}$	(5) FN Rate $\frac{FN}{FN+TP}$	(6) Precision $\frac{TP}{TP+FP}$	(7) F1 Score $2\frac{\text{Precision}*\text{TP Rate}}{\text{Precision}+\text{TP Rate}}$
All Zero	.03	.00	1.00	.00	1.00	.00	.00
Dictionary	.14	.81	.86	.14	.19	.15	.25
Dictionary w/ Negation	.07	.74	.94	.06	.26	.28	.40
Logistic Regression	.07	.81	.93	.07	.19	.26	.40
Logistic Regression w/ Negation	.05	.83	.95	.05	.17	.36	.50
GPT-3	.05	.93	.95	.05	.07	.36	.52
WHAM (Generic English)	.02	.96	.98	.02	.04	.66	.78
WHAM (Baseline)	.01	.97	.99	.01	.03	.75	.85

Note: This table reports the performance of alternative classification algorithms for detecting remote work in a held-out test set of 4,050 unique text sequences that is re-weighted to feature a label distribution more representative of the full Lightcast sample. We create a simulated dataset of 1000 * 4,050 = 4,050,000 observations, 3% (97%) of which are sampled with replacement from the set of positive (negative) test set examples. Each observation is assigned a true value of 0 or 1 according to the majority vote of human labelers, and a predicted value of 0 or 1 according to the particular algorithm. FP is the number of false positives, e.g. the number of observations with a true 0 and predicted 1. Similarly, FN is the number of false negatives, TP the number of true positives, and TN the number of true negatives. All performance measures are built from these four counts. A description of each algorithm is provided in Appendix 8.

7.3 Aggregate Time Series with Alternative Occupation Weights

When we present our aggregate time series (e.g. Figure 4) we adopt a weighting scheme to remove changes in the overall composition of online job vacancy postings. In this Appendix subsection, we expand on these baseline results by depicting the same time series under alternative re-weighting procedures. We first briefly discuss these schemes, and then present Table 7.5 which summarises the impact of these schemes in 2021. Finally, we provide the aggregate time series pertaining to each alternative weighting scheme. In sum, the main features of the time series discussed in Section 4 are robust to the choice of weighting scheme.

We present results from five alternative weighting schemes. In all cases, the weighting approach first calculates the share of vacancy postings in a given country, for a given month, in each six-digit occupation. We then take a weighted average across all occupations to arrive at the final aggregate share of remote work vacancy postings for that month in that country. The different weights used are:

1. Unweighted
   Here we do not apply any additional weights, which implicitly uses the contemporaneous share of online vacancy posts in a given month to define the relative importance of each occupation on the aggregate measure

2. 2019 Vacancy Weighted
   For each country, we calculate the share of online vacancy postings in each occupation in 2019, using our corpus of job ads. This share is then used as the weighting scheme across every month in our time series.

3. 2019 Employment Weighted (baseline specification)
   For each country, we calculate the share of *employment* in each occupation in 2019. We do this by using official national statistics (see Table **??** four our data sources). This share is then used as the weighting scheme across every month in our time series.

4. 2019 US Vacancy Weighted
   Same as '2019 Vacancy Weighted', except that we impose the weights calculated for the USA across all countries.

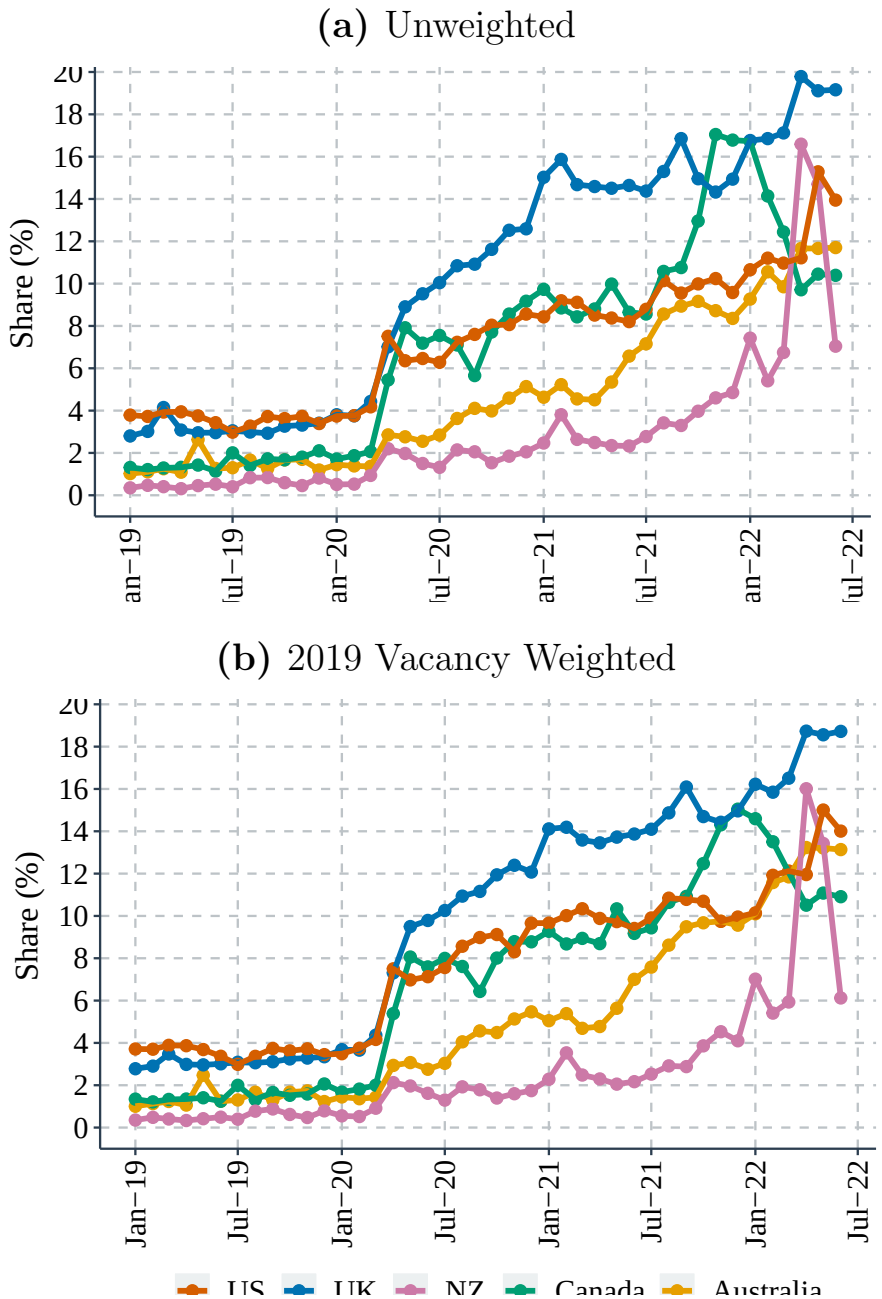
5. 2019 US Employment Weighted
   Same as '2019 Employment Weighted', except that we impose the weights calculated for the USA across all countries.

Table 7.5: 2021 Share of Remote Work Vacancy Postings, by Different Occupation-weighting Scheme

Occupation-weights:	Australia	Canada	NZ	UK	US
Unweighted	6.81	10.93	3.78	15	9.18
2019 Vacancy Weighted	7.27	10.66	4.04	14.34	10.08
2019 Employment Weighted	5.04	9.57	3.28	10.6	7.17
2019 US Vacancy Weighted	6.12	12.11	4.63	13.1	10.08
2019 US Employment Weighted	4.49	8.93	3.21	9.38	7.17

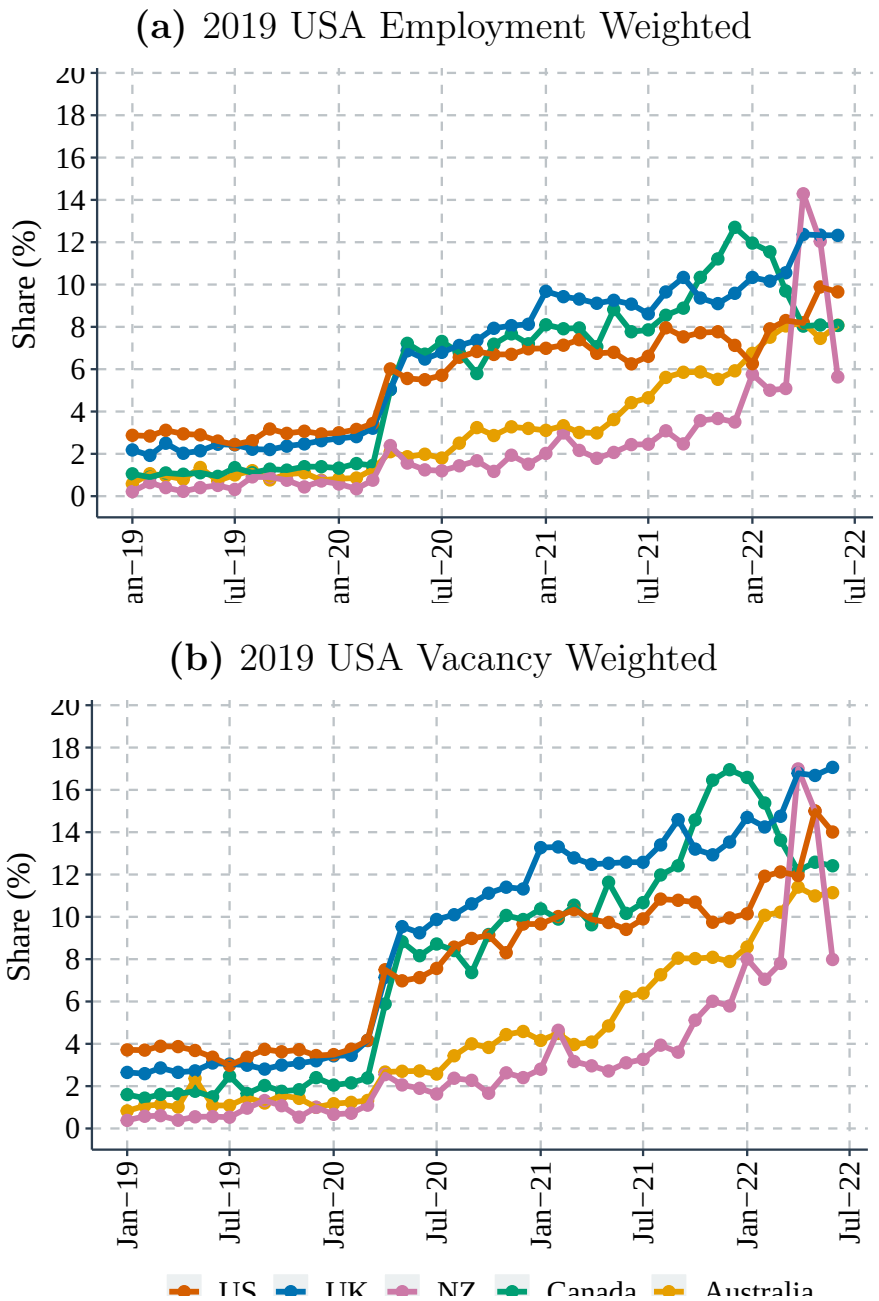
Note: This table presents the share of 2021 job vacancy postings which explicitly offer remote working. We calculate the monthly share of remote work vacancy postings in each country using alternative weighting schemes, and then take the simple average over these twelve months in 2021. * is our baseline specification used in the main body of the paper. For a discussion of these different weighting schemes, see Appendix 7.3.

Figure 7.2: WFH Adoption Across Countries, alternative weighting schemes



Note: This panel shows alternatives to Figure **??**, using different weighting schemes. Panel (a) applies no weights to the underlying series, i.e. it is based on the contemporaneous occupation mix across online vacancy postings. Panel (b) uses the 2019 share of vacancies in each occupation within countries, and applies these weights across all months. Remote work is measured from the text of job ads, using our 'WHAM algorithm' (see Section **??**).

Figure 7.3: WFH Adoption Across Countries, alternative weighting schemes (continued)



Note: This panel shows alternatives to Figure **??**, using different weighting schemes. Panel (a) uses the employment share across occupations in the USA in 2019, and applies these weights to all countries in all time periods. Panel (b) uses the share of online vacancies across occupations in the USA in 2019, and applies these weights to all countries in all time periods.. Remote work is measured from the text of job ads, using our 'WHAM algorithm' (see Section **??**).

8 Supplementary Information on Measurement

8.1 Dictionary Approach

Here we describe how we implemented the dictionary-based approach. As is clear in the body of the text, we caution against using such an approach in the context of measuring remote work from job vacancy postings as we found substantial bias. But for replication and transparency, we provide the list of keywords and a discussion of our negation-adjustment procedure here. This list was chosen following the proposed set of keywords from Adrjan et al. (2021). Our dictionary approach consists of the following steps:

1. **Preprocessing**: We lowercase all text, remove punctuation symbols (except for the hyphen and the apostrophe), remove numbers, and replace all white spaces with a single one.

2. **Tagging**: We search for the appearance of any of the keywords from our list in the text of the job postings. For keywords containing multiple words (e.g. work from home) we allow for any arbitrary combination of white spaces and hyphens separating the words that compose the dictionary keyword (e.g. work-from-home, work- from- home).

3. **Binary classification**: Any job posting that contains a match to any of the dictionary keywords is classified as positive.

Table 8.6

working remotely	working from home	work remotely
work from home	work at home	teleworking
telework	telecommuting	telecommute
smartworking	smart working	remote work teleworking
remote work	remote	remotely
homeoffice	home office	home based
homebased		

Note: Adrjan et al. (2021) + "remotely" + "homebased"

8.1.1 Negation adjustment

To account for the possibility that a job posting is explicitly mentioning keywords related to remote work to forbid it, we implemented a negation adjustment to our dictionary approach.

We will refer to this method as *Dictionary with negation*. This adjustment follows the strategy proposed by **?** to capture negation in the context of sentiment analysis. For every keyword match from the dictionary within a job posting we consider it to be negated if any of the following is true:

1. There is a negation term in any of the three words before the keyword match
2. "no" or "not" appear in the two words after the keyword match
3. A word that contains "n't" is the immediate word after the keyword match

If a job posting is negated we then change its binary label from positive to negative.

Table 8.7

aint	arent	cannot	cant	couldnt	darent	didnt	doesnt
ain't	aren't	can't	couldn't	daren't	didn't	doesn't	dont
hadnt	hasnt	havent	isnt	mightnt	mustnt	neither	don't
hadn't	hasn't	haven't	isn't	mightn't	mustn't	neednt	needn't
never	none	nope	nor	not	nothing	nowhere	oughtnt
shant	shouldnt	uhuh	wasnt	werent	oughtn't	shan't	shouldn't
uh-uh	wasn't	weren't	without	wont	wouldnt	won't	wouldn't
rarely	seldom	despite	no				

Note: Used by **?** from VADER Sentiment Analysis tool + "no"

8.2 Data collection for Human labeling

Given that most of the text of a job ad is not relevant for identifying remote work, we had to develop a special sampling procedure in order to collect examples of sequences which had a high probability of discussing remote work. We used four main strategies for generating the set of sequences labeled by humans. For each of these procedures we balanced the sample across year-quarter from 2014Q1 up until 2021Q3. The final set of sequences we sent to our human auditors was drawn uniformly from four different groups:

1. Sequences that contained a keyword from an ex-ante list of 213 keywords that may indicate remote work
2. Sequences that contained any of 5 generic terms: *remote**, *home*, *work*, *location*, *tele**

3. Sequences that contained terms that may cause confusion when identifying the presence of remote work (e.g. *home repairs*, *nursing home*, *remote construction*)

4. Random sequences that were not part of any of the 3 groups defined above

Groups 1-3 were based on our quite large investigations into the performance of dictionary methods. In sum, this sampling procedure to generate data for human reading can be thought of as a very naïve attempt at reinforcement learning, albeit in an ad hoc manor. Conference participants have suggested leveraging a more systematic reinforcement learning algorithm, whereby we systematically iterate and send edge cases back to humans for labelling. This could be a promising strategy in future applications.

8.3 Language model details

Our WHAM measure uses a distilled version of BERT (i.e. DistilBERT) developed by Sanh et al. (2020) and available through HuggingFace. DistilBERT has a transformer architecture with 6 layers and 66 million parameters and was pre-trained with the task of predicting missing words in a corpus of unpublished books and all English Wikipedia. We use the uncased version of the model.

Table 8.8 contains the set of parameters for the unsupervised pre-training on job ad sequences and the supervised training on human-labeled sequences. Both set of parameters were selected following guidelines from the authors of BERT (Devlin et al. 2019a). In order to select the optimal parameters for the training on human-labeled sequences we performed an exhaustive search over learning rates $\{2*10^{-5}, 3*10^{-5}, 5*10^{-5}\}$, epochs $\{2, 3, 5\}$ and batch sizes $\{16, 32\}$. We performed a 3-fold cross-validation procedure and selected the model with highest average F1 score across the data splits.

Table 8.8: Language model training parameters

	Unsupervised pre-training	Training on human labels
Learning rate	5e-5	5e-5
Epochs	3	2
Batch size	8	16
Max sequence size	512	512
Percentage of masked tokens	15%	Not relevant

8.4 Additional methods description

Below we present a detailed description of the remaining methods.

8.4.1 All Zero

As a benchmark, we implement a trivial method that classifies every sequence as not containing remote work.

8.4.2 Logistic Regression

Following the approach proposed by **?** we estimate a logistic regression with L_1 regularization (LASSO) on the document-term frequency matrix of our labelled sequences. In order to do this, we start by applying the same pre-processing steps from the dictionary approach to the job postings: i) lowercase text, ii) remove punctuation (except for the hyphen), iii) remove numbers, and iv) clean white spaces. After this step, we split the text into individual tokens and build the document-term frequency matrix by using the 5,000 most common tokens. For all the keywords in our dictionary (described in 8.1) that are not part of the 5,000 most common tokens, we add a column in the document-term matrix with its counts. Finally, we transform the matrix into its binary form; every entry above one is replaced with a one. We use a 5-fold cross-validation procedure to find the optimal regularisation parameter among multiple candidates. We choose the parameter that achieves the highest average $F1$ score across the five splits.

8.4.3 Logistic Regression with Negation

We follow an identical procedure to the one described for the Logistic Regression but we further extend the document-term matrix with one extra column per keyword in the dictionary that indicates that the keyword was negated (according to our negation adjustment described before).

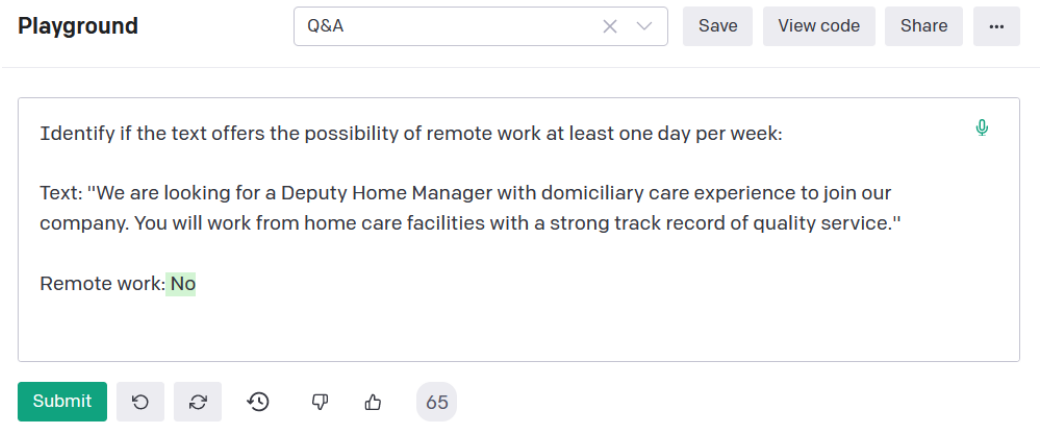
8.4.4 WHAM - Generic English

We develop a similar model to WHAM Baseline with the only difference that we skip the pre-training step on the corpus of job vacancy postings. We call this model WHAM (Generic English) since this is a model whose language representation entirely relies on the off-the-shelf generic training on Wikipedia and the Book Corpus done by the authors of the model.

8.4.5 GPT-3

We use OpenAI's GPT-3 model to generate predictions on the presence of remote work in our job postings. To do this, we craft a simple prompt that instructs the model to *"Identify if the text offers the possibility of remote work at least one day per week"* and ask the model to generate an answer. Figure 8.4 illustrates a particular example using OpenAI's Playground.

Figure 8.4: GPT-3 Example



Note:

In most cases, the text generated by GPT-3 is a *Yes/No* answer. Sometimes, however, the model generates longer answers (e.g. "temporarily due to covid"). In order to transform these answers into a binary prediction we do the following: i) lowercase the answer of GPT-3 and clean any additional white spaces and ii) if the answer contains "no" as part of its three first characters we assign a zero (no remote work) to the sequence, else we give it a one (remote work).

We test both the "text-davinci-02" model and the 'text-davinci-03" model using the same prompt and report performance of the former given its lower error rate with respect to our human labels.

8.5 Extended comparisons between methods

In the main paper, we show that there is substantial disagreement between the dictionary approach (using the set of terms in 8.6) and our approach to measuring remote work from job vacancy text. One additional test we can conduct is to evaluate the performance of

the dictionary on our labelled data. Table **??** contains these results. Table **??** assigns labels in test sequences depending on whether a term from the dictionary is present in the text. Table **??** considers an alternative in which the presence of a term only generates a positive classification if it is not surrounded by a negation term[35]. In both cases, the drop in accuracy is substantial: whereas WHAM achieves 98% accuracy on these sequences, dictionaries achieve 84% and negated dictionaries achieve 87%.

We also use our set of labelled data to evaluate the standard metrics of predictive performance, shown in Table **??**. This Table shows that when we compare our baseline 'WHAM' model to 'Dictionary' and 'Dictionary (negation-adjusted)', we see that the performance increase is substantial. Note that this test was done on data constructed as per Appendix 8.2, which explicitly biases the sample towards cases where dictionaries may have increased performance (indeed, 25% of this sample was populated based on exactly these dictionary criteria!). As such, we caution against this table as being the final arbiter of performance, and refer the reader to Table **??** for a comparison of WHAM to other methods across the full corpus of vacancy postings.

We also compare a number of alternative models in 7.4. We show the performance of 'BERT (dictionary examples)' which used the BERT sequence-embedding model but was trained using examples from the dictionary (as opposed to our baseline approach which uses human labelled text). Interestingly, we find that this approach has an almost identical performance to the dictionary itself. We believe this is because the model learnt the exact semantic structure of the dictionary classification process, and thus did not use its capacity for context-dependent inference. Put differently, without human labelled data a sequence-embedding model does not generalize beyond the dictionary used to generate the examples. Next, we estimate a logistic regression with L_1 regularization (LASSO) on the document-term frequency matrix of our labelled data. We call this method 'Logistic regression' and show that it is able to outperform both of our dictionaries. Then, we show the results for 'WHAM (no pre-training)', which applies all the same steps as our baseline model but does not conduct additional pre-training on the corpus of job vacancy postings. This performs only marginally worse than our baseline model, suggesting the off-the-shelf language model already parses the context-dependencies in English language job ads quite well.

8.6 Computational setup and associated costs

[35]To be precise, we consider a term negated if one of the following is present within a four-word window: 'can't', 'cant', 'don't', 'dont', 'isn't', 'isnt', 'no', 'not', 'unable', 'won't', 'wont'.

Table 8.9: Setup and costs

	Pretraining	Fine-Tuning	Full sample prediction
Computational setup	GCP Virtual Machine with 1 NVIDIA V100 GPU	GCP Virtual Machine with 1 NVIDIA V100 GPU	GCP Virtual Machine with 8 NVIDIA A100 GPUs
Total time (hours)	12	3	36
Job postings (per hour)	NA	NA	7,000,000
Cost per hour (USD)	$3	$3	$40
Total Cost (USD)	$36	$9	$1,440