Peter John Lambert

Measuring Remote Work Using a Large Language Model (LLM)

KEY MESSAGES

- **Large Language Models (LLMs) can dramatically improve upon traditional text-based measurement tools used by economists**
- **We fit, test and train the "Work-from-Home Algorithmic Measure" (WHAM) model to detect new online job postings offering remote/hybrid arrangements. The WHAM model has near-human accuracy. We deploy this model at scale, processing hundreds of millions of job ads collected across five countries and thousands of cities**
- **The share of new ads offering remote/hybrid jobs increased four-fold in the US and more than five-fold in the UK, Australia, Canada, and New Zealand, between 2019 and 2023. These data and more are available for researchers at wfhmap.com**
- **The "remote work gap" across cities, occupations, and high/low salary workers continues to widen, and the hare of advertised remote/hybrid work is highly skewed towards white-collar workers and cities which are hubs for government, business, technology, and higher education**
- **LLMs offer massive potential for empirical research using text data, but one should adhere to best practices and understand the "do's and don'ts" of these technologies. Generative AI offers immense promise, with some significant limitations**

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. There are few modern precedents for such an abrupt, large-scale shift in working arrangements.

Peter John Lambert

is a PhD student at the London School of Economics and Political Science.

Most previous efforts to quantify and characterize this shift rely on surveys of workers and employers or assessments of remote work feasibility by occupation. In our paper, "Remote Work across Jobs, Companies, and Space" by Hansen et al. (2023), we use the information contained in job vacancy postings, which are readily available and have massive geographic coverage.

We analyze the full text of hundreds of millions of job postings in five English-speaking countries. In doing so, we apply a state-of-the-art Large Language Model (LLM) to analyze the text and determine whether the job allows remote/hybrid work. We fit, test, and refine this LLM using 30,000 classifications generated by human readings. We also identify each job vacancy's city, employer, industry, occupation, and other attributes.

Our approach to studying the remote work phenomenon has several noteworthy strengths:

1. Our data cover all vacancies posted online by job boards, employer websites, and vacancy aggregators across five countries. Coverage on this scale is infeasible with survey methods.
2. Postings typically describe the job and its attributes in detail, as suggested by a median posting length of 347 words. It also reflects a legal right and represents a future-looking organizational commitment rather than temporary arrangements.
3. We develop the WHAM model (our own LLM) that reads and classifies postings in an automated manner. The model achieves a 99 percent accuracy rate in flagging jobs that allow for remote/hybrid work, significantly outperforming other methods for text-based measurement.
4. 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 monthly in hundreds of occupations, thousands of cities, tens of thousands of employers, and city-by-occupation and employer-by-occupation cells. We continuously update and post many of these statistics at wfhmap.com.

The remainder of this article is split into three sections. In the next section, I discuss our research paper's data and measurement approach. I also provide some detail on our approach's performance compared to widely used methods in text-based measurement. The third section documents several patterns in the diffusion of advertised remote/hybrid jobs. Lastly, I discuss the potential for text-based measurement using LLMs. I share some "do's and don'ts" when using these technologies and discuss the potential benefits and drawbacks of the new wave of Generative AI for empirical text-based-measurement in economics.

DATA AND MEASUREMENT

Data

We examine over 250 million online vacancy postings collected by Lightcast (formerly Emsi Burning Glass), an employment analytics and labor market information firm. Lightcast scrapes postings from over fifty thousand online sources, including vacancy aggregators, government job boards, and employer websites. Lightcast claims to cover a "near-universe" of online postings in our five countries during the period covered by our analysis.

For each online vacancy posting in our dataset, we can access a plain text document scraped from the job listing. We also observe the posting date, employer name, occupation, location of the employer, industry, and more. We consider postings listed from January 2014 to February 2023.

The resulting dataset covers hundreds of millions of online vacancy postings in five countries, spanning 5.2 million employers and nearly 40 thousand cities.

For our baseline results, we re-weight the postings in each country-month cell to match the US occupational distribution of new online vacancy postings in 2019.

Measurement

The measurement problem we face is determining whether each job posting allows a new hire to work remotely, understood here to encompass both fully remote and hybrid positions. 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 hybrid working arrangements, we use a threshold of at least one day per week for our positive classification. This approach effectively measures an employer's willingness to offer flexibility in work-location.

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 at 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.

We found that a "keywords" approach was immediately problematic, due to high prevalence of (i) negation, (ii) context-dependent language, and (iii) wide array of language used to refer to remote work arrangements. To overcome this, we instead relied on a large-language model (LLM) which we call the "Work-from-Home Algorithmic Measurement," or WHAM model.

We build our WHAM model using the following steps:

1. *Partition the set of all text documents using coarse keyword measures:* In order to inform a sampling strategy of which text extracts to send to human auditors, we first partitioned the set of all documents. To do this, we relied on keyword search methods–which can be implemented with low cost. We constructed a set of very broad keywords, such as "remote," "job," "work," and so on.
2. *Collect 30,000 human labels:* We asked humans on the Amazon Mechanical Turk platform to classify whether a passage of text constituted an offer of remote/hybrid work arrangements. We used a sample of 10,000 text passages and asked three auditors to evaluate each passage. This forms the basis for our training data and provides a set of labels to evaluate model performance.
3. *Take an existing pre-trained LLM:* We took the DistilBERT language model, which comes pre-trained on the complete English-language Wikipedia and thousands of unpublished books. This model has shown in industrial applications to already have a very high grade of performance at understanding the rich context-dependencies between words in a sequence.
4. *Further pre-training the LLM:* We further pre-trained this model by exposing it to millions of passages from online job vacancies in our corpus. This ensures the resulting model understands context-dependencies between words in the context of job advertisements.
5. *Fine-tune the LLM to predict remote/hybrid work:* We next deployed the fully pre-trained model on the task of predicting whether a passage of text constitutes an explicit offer of remote work. We did this by embedding a final prediction layer in the neural network structure of the model.

These steps result in our WHAM model, which we use to predict remote/hybrid arrangements across the full set of job ads. We show in the next section that this model produces a 99 percent accuracy rate—relative to human auditors—greatly outperforming other text-measurement technologies. It even shows a five-fold higher accuracy rate compared to GPT-3.

Evaluating Performance

To evaluate the performance of our WHAM model, we remove a portion of our human-labelled text passages from the training stage and evaluate performance on this held-out sample. As well as measuring the overall performance of WHAM, we also assess performance of a variety of other measurement technologies.

We first take a dictionary of keywords used in the literature to measure remote work arrangements (Adrajan et al. 2021), and classify remote work based on the presence of these terms ("Dictionary"). We next augment this dictionary with a negation adjust-

Table 1

WHAM Outperforms Other Classification Methods

	(1)	(2)	(3)
Prediction technology:	Error rate	Precision	F1 score
Dictionary	0.14	0.15	0.25
Dictionary w/ negation	0.07	0.28	0.40
Logistic regression	0.07	0.26	0.40
Logistic regression w/ negation	0.05	0.36	0.50
GPT-3	0.05	0.36	0.52
WHAM (Baseline)	0.01	0.75	0.85

Note: This table reports classification performance metrics, which we calculate using a hold-out sample of human-classified text sequences. "Error rate" is the overall rate of misclassifications (relative to humans). "Precision" is the ratio of true-positive classifications to the sum of true positives and false positives. "F1 score" is the harmonic mean of Precision and "Recall", where Recall is the fraction of true positives divided by the sum of true positives and false negatives – i.e., the denominator is the true number of positives, according to human classifications.

Source: Author's own calculation.

ment, whereby the keyword match is only taken as a positive classification in the absence of nearby negation terms (our set of negation terms comes from the VADER sentiment analysis dictionary). Next, we implement a Logistic regression approach, following the methodology used in Adams-Prassl (2020). We also extend this to include a negated implementation. Finally, we implement a zero-shot classification method using GPT-3.

Table 1 shows the performance of the above prediction technologies. We see that our baseline WHAM model delivers the highest accuracy, with an error rate of just 1 percent relative to human predictions. This is a fourteen-fold improvement relative to the Dictionary of keywords approach. The WHAM model also outperforms our GPT-3 implementation, which has an error rate of 5 percent. The performance gains of our WHAM model are even more impressive in terms of the F1 score[1], which assigns more weight in the performance evaluation to the class of positive values.

The key difference between our approach and others is that WHAM considers surrounding words, which may change the meaning of the text. To illustrate this, we show in Figure 2 some examples where the dictionary leads to spurious classifications (see below). We also illustrate how the attention mechanism of WHAM understands the context surrounding each passage, overcoming the limitations of the dictionary/keyword measurement.

In sum, our approach to measuring remote/hybrid work arrangements has substantial performance improvements relative to widely used algorithms in the economics literature. A key contribution of the paper is to provide a concrete case study, and document in detail the relative performance improvements in this context.

[1] The F1 score is a metric used to evaluate the performance of binary classification models, which are models that distinguish between two classes or categories. It is a measure that combines both precision and recall, giving equal weight to both. Precision is the fraction of true positive predictions out of all positive predictions, while recall is the fraction of true positive predictions out of all actual positive instances. The F1 score is calculated using the harmonic mean of precision and recall.

PATTERNS IN ADVERTISED REMOTE WORK

Advertised Remote Work Diffusion across Countries

How did the share of advertised hybrid and fully remote work differ across countries prior to, during and after the pandemic? Figure 1 shows the monthly time series of the share of advertised remote/hybrid work for the US, the 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 US during 2019. Three high-level facts emerge:

- *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 percent. While this immediate increase occurred across all five countries, the level-change was most pronounced in countries with a more severe initial Covid outbreak (US, UK and Canada)
- *Sustained growth thereafter.* Since the large spike in March-April 2020, there has been sustained growth in the percentage of advertised remote work. In level-terms, this growth has been most pronounced in the UK (here Covid lockdowns lingered longest and were most severe relative to the other countries in the sample). We also see evidence of higher growth rates in Australia and New Zealand as their pandemic experience worsened during 2021. 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 is that our measure of new job vacancies lags the stock of employees working from home,

possibly because employers were slow to accept this as a permanent practice.

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

Remote Work across Jobs

Figure 2 shows the share of advertised remote work by broad occupation groups (based on two-digit SOC 2010 classifications). The differences across broad occupation groups varies greatly. In 2019, we see that just one-in-twenty job ads in "Computer and Mathematical" occupations explicitly offered remote work arrangements in their postings, whereas in 2022 this share raises to a more one-third of new ads offering 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 specialized equipment or customer interactions.

Remote Work across Cities and the "Remote Work Gap"

We next turn to more granular monthly time series for selected "US ciU.S.es," shown in Figure 3. As well as illustrating the granularity of our data, several interesting features emerge from these time series:

– Cities from the North-East and West regions (e.g., San Francisco (SF), Boston, New York (NYC)) all experience similar increases at the outset of the pandemic but have very different growth levels subsequently. By 2023, these differential growth rates result in very dispersed levels.
– We 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 far less volatility. Savannah and Miami Beach appear to have partially reverted to pre-pandemic shares of advertised remote work.

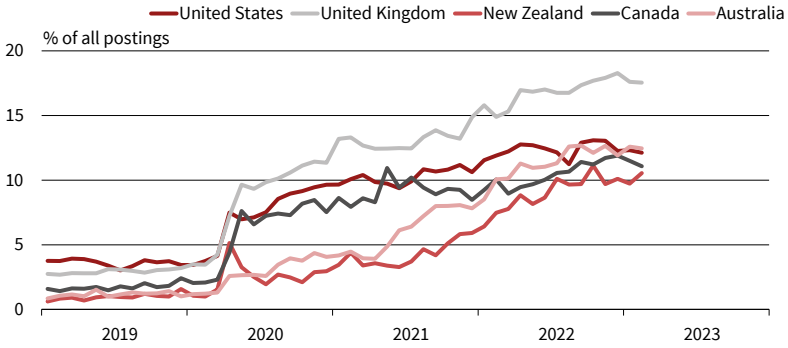
Other Patterns and Trends

In our research paper, as well as in the data available at wfhmap.com, we document several other facts about the discussion of advertised remote/hybrid work in online job vacancies postings. These include:

– Wide dispersion in occupation-level shares of advertised remote/hybrid work.
– A lot of within-occupation heterogeneity, even for occupations with very high shares of overall remote work.
– Measure of task-feasibility at the occupation level can vary a great deal from actual advertised remote work, for example due to varying worker and firm preferences.
– Pre-pandemic share of remote/hybrid is a strong predictor of 2022-23 remote work share for occupations, but a worse predictor for cities.
– This suggests confounding city-level factors are salient, such as institutions, pandemic experience, transport and internet infrastructure, and cultural norms.

Figure 1

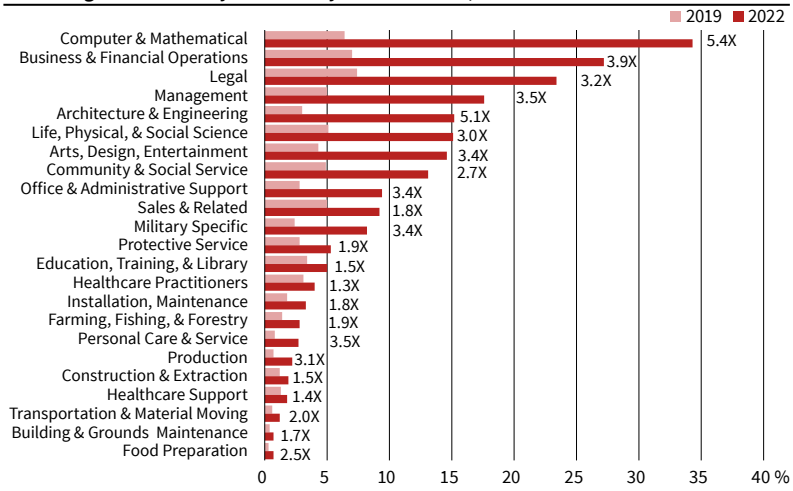
Vacancy Postings That Explicitly Offer Hybrid or Fully Remote Work Rose Sharply in All Five Countries from 2020



Note: This figure shows the percent of vacancy postings that say the job allows one or more remote workdays per week, encompassing both hybrid and fully-remote working arrangements). We compute these monthly, country-level shares as the weighted mean of the own-country occupation-level shares, with weights given by the US vacancy distribution in 2019. Our occupation-level granularity is roughly equivalent to six-digit SOC codes.
Source: Authors' calculation.

Figure 2

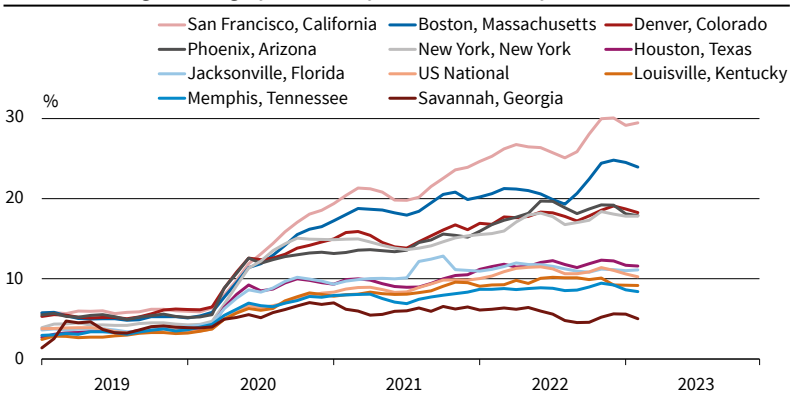
Professional, Scientific and Computer-Related Occupations Have the Highest Shares of Postings That Offer Hybrid or Fully-Remote Work, US Data



Note: Each bar reports the percent 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).
Source: Authors' calculation.

Figure 3

Share of Postings Offering Hybrid or Fully Remote Work Vary across US cities



Note: We calculate the monthly share of all new job vacancy postings which explicitly advertise remote working arrangements (i.e. both hybrid and fully-remote), by selected cities. Prior to aggregation at the monthly level, we employ a jackknife filter to remove a small number of outlier days. This figure shows the 3-month moving average. Cities chosen above are selected examples to illustrate the wide cross-city spread.
Source: Authors' calculation.

– Our measure of advertised hybrid/remote work from new job postings correlates strongly with the American Community Survey (ACS)'s measure of the proportion of employed who "mostly work from home."

In sum, our research paper provides a measurement approach which leverages a huge corpus of text and provides near-human classification accuracy at scale. We use this to document patterns in advertised remote work at a fine spatial granularity and do so with monthly real-time updates. These data can be accessed by researchers at wfhmap.com.

SOME DO'S AND DON'TS OF USING LARGE LANGUAGE MODELS

Our paper shows that LLMs offer huge potential for economists seeking to measure information from text data sources. If properly implemented, these models can deliver near-human accuracy at huge scale. With text data already a mainstay of empirical analysis, these technological advancements offer huge opportunity to researchers. Here are five quick do's and don'ts which other researchers might find helpful:

– Do: Invest in high quality "ground truth" measures of the feature of interest. In our case, we used humans to label a sizable set of text extracts. Any model will only ever be as good as your initial training data. Platforms like Amazon Mechanical Turk (AMT) are hugely useful and cost effective for extracting these labels. When using these platforms, screen auditors carefully. It helps to pay an efficiency wage premium to ensure quality work. It's also useful to have at least some of the labels processed by multiple auditors, to add an intensive margin to the training data in the case of disagreement.
– Don't: Refrain from working with very lengthy documents. In our application, we split job ads roughly into paragraphs. This increased the number of documents to process but offers two important benefits. First, it reduces the cognitive cost of humans conducting audits. Second, it reduces the potential for over-fitting, ensuring the language model identifies the correct linguistic features.
– Do: Ensure the training data is well balanced, especially when the feature of interest is very unbalanced. In our case, there were vastly more negative (not WFH) text extracts. Even a single job ad which offers remote work typically mentions this in a single paragraph. A good sampling strategy will over-weight documents likely to contain the feature of interest, while still allowing for many random draws from the full population to enter the training data.
– Don't always think you need the latest-and-greatest tools! For a great many applications, classification based on a set of key terms will work brilliantly. For other use-cases, a trained classifier using word-vectors as inputs will also work great. No matter the technology employed, always test performance on labelled data. Applications that work well with keywords are typically cases where attrition bias is stable both over time and cross-sectionally.
– Do: Consider fine-tuning the LLM. If a large language model is warranted, it is very helpful to fine-tune the model for your specific classification task (e.g., by adding a prediction layer at the end of a neural network). The alternative is to collect generic vector embeddings of passages, and then fit a prediction algorithm using these vectors as inputs. Fine-tuning the model will help the huge number of parameters in these models work towards your specific measurement question.

GENERATIVE AI AND TEXT-BASED MEASUREMENT IN ECONOMICS

Perhaps the most transformational breakthrough in LLMs is the recent mainstream adoption of "Generative AI" tools such as OpenAI's ChatGPT. These technologies will have far reaching and profound impacts, not least of which will be on empirical research using text. Nonetheless, there are some important limitations which users ought to be aware of.

Chat Bots Are Zero-shot Measurement Technologies

As a measurement technology, the currently available set of Generative AI tools is inherently "zero-shot," meaning that the output provided by the model is the final measurement, with no opportunity for further refinement based on feedback.

This is due to their extensive size and reliance on specialized computational resources, and because the models themselves are proprietary technology. Consequently, researchers must rely on web or API-based interfaces to interact with these models, which restricts their ability to further optimize the model for performance in a specific context.

In our work, we found that GPT-3 was approximately five times less accurate than our WHAM model. This is despite our model relying on 44 million parameters in comparison to the 175 billion parameters powering GPT-3.

The superior performance of our model is almost wholly attributed to the fine-tuning process, whereby a significant proportion of the model's parameters were optimized for the specific task of predicting offers of remote/hybrid work.

It remains uncertain whether the development of increasingly larger and more refined models will eventually render fine-tuning obsolete. For more bespoke measurement exercises, the value of fine-tuning is likely to remain a key reason for sticking with publicly available LLMs instead of using generative AI for direct measurement.

Training Data: AI vs Humans

Even if the Generative AI tools exhibit superior measurement performance, the cost of implementing this at scale is another reason to favor deploying earlier generation LLMs. One way to utilize these technologies in a cost-effective way is to use them to develop the training data on which a smaller more cost-effective model is trained.

The evidence on whether this is a good idea is mixed. We found that humans performed better at a binary classification exercise when we exposed them to small text-extracts. More generally, the larger the text extract, or the more classification categories presented to humans, the less reliable they become (as measured by disagreement rates). A recent paper by Galard et al (2023) found that ChatGPT outperformed human auditors when processing five separate categories.

In some sense, with a large enough set of well-intentioned auditors, humans can never be collectively "wrong." After all, we are typically measuring a feature that has salience through human interpretation. If no human recognized that a document offered remote work, well, did it?

Philosophy aside, the practical question is whether, on a given budget, a small sample of human audits will be as informationally useful to training a model as a potentially larger set of labels extracted from a generative AI. For limited budgets, longer documents, and many features of interest, this is likely to be true. Finally, consider that a model trained on any set of labels will be constrained by the quality of these labels, so if the Generative AI lacks accuracy, the final model will too.

REFERENCES

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

Adrjan, P., G. Ciminelli, A. Judes, M. Koelle, C. Schwellnus and T. Sinclair (2021), "Will It Stay or Will It Go? Analysing Developments in Telework during COVID-19 Using Online Job Postings Data", *OECD Productivity Working Paper* 30.

Gilardi, F., M. Alizadeh and M. Kubli (2023), "ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks", *arXiv*:2303.15056.

Hansen, S., P. J. Lambert, N. Bloom, S. J. Davis, R. Sadun and B. Taska (2023), "Remote Work across Jobs, Companies, and Space", *NBER Working Paper* 31007.