work2vec: Using Language Models to Understand Wage Premia

Sarah H. Bana[*]

December 7, 2024

Abstract

Hedonic regressions have long helped economists understand how job characteristics contribute to earnings, but measurement challenges have limited which attributes could be analyzed systematically. Using a new dataset linking salary information from Greenwich.HR to job posting data from Burning Glass Technologies, I apply natural language processing techniques to decompose how different job characteristics contribute to earnings. The resulting model explains 83 percent of salary variation—a 19 percent improvement over traditional occupation-location controls. Using an attribution method called integrated gradients, I identify which words most strongly predict salaries. I then develop an entity extraction model to categorize posting content into activities, amenities, education, experience, firm names, general and technical job skills, hours, job titles, and location. The analysis reveals that job activities dominate both in frequency and earnings relevance. While skills and job titles have been used as proxies for tasks, directly measuring the activities described in job postings provides better insight into wage determination. This represents the first decomposition to quantify how such a wide range of workplace characteristics—rarely captured in administrative data—shapes earnings.

[*] Chapman University and Stanford Digital Economy Lab. Email: sarah.bana@gmail.com. I would like to thank Max Feng for his very valuable research and data engineering support, and Catherine Weinberger for her thoughtful feedback. Thank you to Cary Sparrow at Greenwich.HR for providing the data and details around it. I appreciate the comments and suggestions from seminar participants at the Stanford Institute for Human-Centered Artificial Intelligence, Boston University, Emory University, Simon Fraser University, the University of Minnesota, the Brynjolfsson Lab, MIT Initiative on the Digital Economy, University of California Santa Barbara Applied Lunch, Chapman University, the University of Nevada at Reno, the University of Queensland, Federal Reserve Bank of Philadelphia, Federal Reserve Board, Federal Reserve Bank of Minneapolis – Opportunity & Inclusive Growth Institute, University of Nebraska, Lincoln, Bureau of Labor Statistics Data Science Users Group, and the Harvard Kennedy School Growth Lab, along with countless friends and family members. Special thanks for the funding from the Stanford Institute for Human-Centered Artificial Intelligence's Google Cloud Credit Grant and Chapman University Dean's Fund that enables this research. This is a companion paper to work entitled "work2vec: Learning a Representation of Jobs" with Erik Brynjolfsson, Daniel Rock, and Sebastian Steffen. They have been amazing collaborators and provided valuable guidance throughout this work. Throughout the course of this project, I was told that ″job2vec″ was trademarked. Accordingly, we have adjusted to using ″work2vec.″ All errors are my own.

1 Introduction

Hedonic wage regressions have long helped economists understand the determinants of earnings and the sources of earnings inequality (Rosen, 1986). With improved measurement capabilities, researchers have steadily expanded the set of attributes that explain earnings variation, from skills and tasks to location effects and job titles (Deming and Kahn, 2018; Autor and Handel, 2013; Card et al., 2023; Marinescu and Wolthoff, 2020). The relative contribution of each factor to earnings remains unclear, as each of these characteristics has been studied using different, specialized datasets that do not overlap. The emergence of large-scale job posting data creates an opportunity to analyze these various factors within a single dataset.

While job postings offer rich descriptions of positions, previous research has typically reduced this information to fixed effects —categorical variables for job titles, binary indicators for skill requirements, and other discrete measures. The actual text of these postings might contain additional predictive information about wages, particularly in their descriptions of work activities. Natural language processing techniques allow us to systematically analyze these textual descriptions, extracting and quantifying the full set of job attributes that may influence wages.

To demonstrate that text content matters for earnings, I construct a novel dataset linking recruiter-inputed salary information from Greenwich.HR to the text of job postings from Burning Glass Technologies. This linkage allows me to reframe wage prediction as a supervised learning problem. The input, job posting text, is converted to embeddings using BERT, a state-of-the-art natural language processing model (Devlin et al., 2018). The resulting model explains 83 percent of salary variation—a 13 percentage point improvement over a baseline of occupation by location fixed effects. This substantial increase in predictive power suggests that important wage-relevant information is embedded in the text itself, beyond what can be captured by standard categorical variables. For example, the model distinguishes between contextual uses of words: ′models′ carries different implications for earnings when describing someone who ′deploys machine learning models′ versus who ′models exceptional customer service.′

The model′s 83 percent explanatory power might appear modest compared to previous work. Marinescu and Wolthoff (2020) find a coefficient of determination (R^2) of almost 90%, looking at the explanatory power of job titles using posted wages on CareerBuilder. However, this high explanatory power stems from in-sample prediction with job title fixed effects, where each unique title serves as its own predictor. When I evaluate job titles' predictive power out-of-sample—a more demanding test of their ability to explain earnings variation—their performance drops substantially, suggesting that previous findings may reflect overfitting rather than true predictive power.

With evidence that job posting text contains meaningful wage information beyond job titles alone, I turn to understanding which words and phrases most strongly predict wages. To do this, I apply an explainability method called integrated gradients, which attributes importance to each word in a posting (Sundararajan et al., 2017).

The output of this method is an importance score for every token in a posting. I describe the most positive and negative attributions, which have interpretable magnitudes in terms of their contribution to log earnings. The most positive words are associated with job titles, such as "dentist," "psychiatrist," "superintendent," "director," and "engineer." On the other hand, the most negative words are associated with experience, activities, and hours, such as "intern/internship," "cleaning," "entry," "hour," and "aide." Geographic patterns also emerge, with coastal cities commanding wage premiums. These initial patterns suggest that job postings encode meaningful information about wage determination through multiple channels, which motivates a more systematic categorization of the text.

To systematically analyze these patterns, I train an entity extraction model using Gemini 1.5 to categorize words into ten attributes of jobs. These categories, drawn from theoretical and empirical work on wage determination, include activities (Acemoglu and Autor, 2011), amenities (Rosen, 1986; Sorkin, 2018), experience and education levels (Mincer, 1974), firm names (Bonhomme et al., 2023), hours (Mincer, 1974), job titles (Marinescu and Wolthoff, 2020), locations (Card et al., 2023), and skills (Deming and Kahn, 2018). By combining the integrated gradient and entity extraction models, I find that about half (49 percent) of tokens in a posting are related to these constructs, with nearly two- thirds (63 percent) of these tokens having salary implications.

Activities performed at work dominate both in frequency (18.59 percent of tokens) and earnings relevance (21.33 percent), suggesting that understanding exactly what workers do on the job is fundamental to earnings determination. Traditional human capital measures—education and experience—have outsized earnings effects relative to their mention in postings, though their combined influence (3 + 4 percent) remains a third of the influence of activities. General job skills like 'problem-solving ' and 'communication' appear frequently but have smaller earnings implications, likely because employers value these skills regardless of whether they explicitly list them. These patterns suggest that while job characteristics affect earnings through multiple channels, the actual work activities explain more variation than traditionally measured attributes.

Understanding how the labor market prices different job characteristics has been a central focus of labor economics, with hedonic regression techniques providing a framework for estimating these equilibrium values (Mincer, 1974; Rosen, 1986; Heckman et al., 2006; Weinberger, 2014; Deming, 2017). Because both sides of the market are heterogeneous, the equilibrium prices provide information to both firms and workers.

This work builds on three pioneering papers that highlighted the wage heterogeneity within occupations and demonstrated that additional characteristics like tasks, skills demanded, and job titles can explain this variation. Autor and Handel (2013) use a survey that collects new data on the job activities of a representative sample of 1,333 U.S. workers, demonstrating that within-occupation measures of tasks have significant and economically meaningful predictive power for earnings. The drawback of this approach is its lack of scale: to identify rare characteristics, the sample must be substantial. To identify differences over time, the survey must be conducted repeatedly. To the best of my knowledge, the survey has not been conducted since its initial fielding.

Papers that followed used data from online job boards, which enabled analyses to be done in closer to real-time while avoiding costly surveys. Deming and Kahn (2018) showed that skill requirements affect average wages of professionals across MSAs, explaining up to 94 percent of the variation in average wages in MSA-occupation cells. However, their analysis focused on average wages, despite substantial variation within occupations, and was understandably limited to professional job advertisements, as online job postings during that period (2010-2015) leaned heavily towards professional occupations. Marinescu and Wolthoff (2020) found a coefficient of determination (R^2) of almost 90 percent when examining the explanatory power of job titles using posted wages on Career Builder. While this number is remarkably high, it is limited to the sample of under 20 percent of postings that posted wages. Given that postings with and without wages systematically differ, this may be difficult to extrapolate to the general population. My work extends this research by introducing a new dataset with salaries derived from the metadata of job postings. I also demonstrate that job titles have little out-of-sample predictive power because the number of unique job titles is very high.

Previous research has analyzed job characteristics through fixed effects—using categorical variables for titles—or through Burning Glass Technologies' skill taxonomy. However, these approaches reflect data constraints rather than theoretical foundations. The task-based approach of Acemoglu and Autor (2011) emphasizes tasks as the fundamental unit of production. Without systematic measurement of tasks, researchers relied on proxies—Lin (2011) and Autor et al. (2024) use the emergence of new job titles to identify changes in task content. Similarly, the availability of a skill taxonomy has led to substantial research focus on skills. Natural language processing now allows direct measurement and comparison of these characteristics in a single dataset, enabling the first systematic comparison of how titles, skills, and tasks contribute to earnings. The analysis reveals that tasks, not skills or titles, are the primary driver of earnings variation. The actual work activities—while more difficult to systematically classify—explain more earnings variation than any other job attribute.

While skills are not the primary driver of earnings, understanding how the market values different job attributes—from skills and amenities to locations—helps market participants make better decisions. Workers can better target their training investments, firms can optimize their job design and compensation, and policymakers can develop more effective workforce development programs.

The remainder of the paper proceeds as follows. Section 2 describes the data and measurement approach. Section 3 presents the salary prediction model and examines the predictive power of job titles. Section 4 analyzes which words and phrases most strongly predict salaries and develops an entity extraction model to decompose these effects. Section 5 concludes.

2 Data

The data comes from two distinct data vendors, Greenwich.HR and Burning Glass Technologies. In this section, I describe the elements of the data used for each portion of the analysis.

2.1 Greenwich.HR Data

Greenwich.HR (GHR) is a labor market intelligence firm that provides real-time labor market data to application developers, analysts and consultants. GHR consolidates job postings from millions of different sources. A major advantage of the GHR data is that they have collected pay data for over 70 percent of job postings collected in recent months. Though the exact method by which GHR collects this data is proprietary, I outline the approach in general terms to lend credence to the estimates.

While many postings do not contain information on wages, it is common practice for job posting platforms to solicit salary data from the recruiter posting the job. For example, in Figure A1 Panel A and B, it can be seen on one popular platform, Indeed, that recruiters are encouraged to fill in either the exact rate, the range, a starting salary, or a maximum salary. This screenshot is for illustrative purposes only, as the platforms and methods for integrating data used by GHR are proprietary. Panel A suggests that this incentivizes applicants. In Panel C, a similar screen is included for LinkedIn.

This information can be found on the applicant side when searching for postings. Visualized in Figure A1 Panel D, a postings' salary band can be inferred by whether it appears in the search results when changing the pay threshold. These images are intentionally taken from different platforms to demonstrate the ubiquity of this practice. For the main analysis in this study, including the salary distribution presented in Figure 1, I use the median of the wage band provided by GHR for each posting. This approach provides a representative point estimate for each job's compensation while accounting for the range often given in salary information.

This pay data provides a major asset for analysis. However, like many new datasets, there are limitations. First, GHR did not collect the raw job text until 2020. Second, GHR sought to be a comprehensive source of U.S. jobs only beginning in March 2019. Prior to this time period, the focus was on public firms and certain sectors. The first limitation can be overcome by connecting GHR postings to Burning Glass Technologies, which does collect the full text of the posting. The second limitation precludes time series analyses on the changing wage premia over time going back. Because the COVID19 pandemic occurred in 2020, likely changing the premia associated with certain skills and the distribution of postings, this work focuses on cross-sectional variation in wages from April 2019 to December 2019.

GHR contains 34,181,463 job postings for the period April 2019 to December 2019 (9 months). Of these postings, 17,077,904 contain posted salaries (50.0 percent). The posting distribution is displayed in Figure 1. As evidenced by the jagged lines in the density distribution, recruiter-inputted salaries do bunch at round numbers.

A crucial aspect of this analysis is that the salary information is derived from the metadata of job postings, rather than from the visible content of the postings themselves. Throughout this manuscript, I refer to this as the ”recruiter-inputted salary,” distinguishing it from the more commonly discussed concept of ”posted salaries.” This distinction is important because recruiter-inputted salaries may provide a more comprehensive and less biased view of the salary landscape than visibly posted salaries.

Posted salaries have been a subject of recent research, yielding two key consensus statements. First, posted salaries are highly selected (Batra et al., 2023), meaning they may not represent the full spectrum of salaries in the job market. Second, the prevalence of posted salaries is significantly influenced by legislation, such as the Colorado Equal Pay for Equal Work Act (Arnold et al., 2022), which mandates salary disclosure in job postings. This legislative trend, which occurred after the sample period of this study, has led to increased transparency but also potential strategic behavior by firms in salary posting. While these developments are not directly addressed in this work due to the timing of the sample, they underscore the value of analyzing recruiter-inputted salaries, which may be less affected by such selection issues and legislative changes.

To assess the representativeness of the GHR salary data, I conducted a detailed comparison with the Current Population Survey (CPS) earnings distribution. While there are notable differences between the two distributions, particularly in the median and upper percentiles, these can be largely attributed to differences in job composition and reporting methods between new job postings and existing employment. The GHR data tends to reflect lower salaries, which is consistent with the overrepresentation of high-turnover and lower-skilled jobs in online postings. A full analysis of this comparison, including a discussion of potential reasons for these differences and their implications for this study, is provided in Appendix B. Despite these differences, the GHR data offers a unique and valuable perspective on salary information in the job market, particularly for new positions and emerging trends.

2.2 Burning Glass Technologies Data

Burning Glass Technologies (BGT) is an analytics software company that strives to provide real-time labor market information to higher education institutions, firms and municipalities. The product used in this analysis is the job postings data, collected from over 40,000 online job boards and company websites. These postings are deduplicated in a proprietary manner and the job title and employer name are cleaned.

For the analysis described, the key attribute of the data employed is the raw job text. This raw text has been seldom used in prior research, and contains virtually all the information that the applicant will see. The job text frequently contains information about the firm, the role, and the application procedure, though this is not systematic.

For illustrative purposes, the raw job posting text of two sample postings from October 2019 are displayed in Figure 3. Both postings use different terms to convey similar information. For example, in the first posting, responsibilities are outlined in the "Key Responsibilities" section, while these same thoughts are outlined in the second posting under the heading, "What would you do? The Specifics." Postings also differ in length, and some postings have some information about benefits and how to apply.

The GHR data contains recruiter-inputted salary, while the BGT data contains the job text. I link these two datasets using the firm name, job titles and date of the posting. This is a fuzzy match, especially because the two datasets are cleaned differently.

The matching process between the GHR and BGT datasets, while necessary for combining salary information with job text, introduces risks of both Type I and Type II errors due to the fuzzy nature of the match. There's a risk of false positives (Type I errors) where incorrect matches are made, and conversely, a risk of false negatives (Type II errors) where true matches are missed. To mitigate Type I errors, which could introduce bias in the relationship between the salary and its text, the matching criteria are intentionally restrictive. This approach prioritizes precision over recall, accepting a higher rate of unmatched true pairs to minimize false matches. While this results in a smaller matched sample, it ensures greater confidence in the validity of the matches made. The potential bias from the smaller sample is mitigated by the fact that vendor-specific data cleaning processes are presumably independent of the relationship between job text and salary, preserving the representativeness of the observed correlations. Specific details on the matching process are described in Appendix A.1.

The resulting sample consists of 1,953,572 postings, with 1,069,759 (54.8 percent) containing salary information. These salary-inclusive postings serve as the input for the models described in the next section.

3 Salary Prediction Model

3.1 Salary Prediction Model Structure

Traditional approaches to analyzing job posting text, such as word counting or n-gram methods, face significant limitations. These methods often result in sparse matrices and fail to capture the nuanced meanings of words in different contexts. For instance, the word "python" could refer to a programming language or a reptile, depending on its context.

To address these challenges, I employ BERT (Bidirectional Encoder Representations from Transformers), a powerful contextual language encoding framework introduced by Devlin et al. (2018). BERT processes text by breaking it into tokens—individual words or parts of words. While simple words like 'the' or 'and' are single tokens, longer or specialized terms are often split into meaningful subwords (e.g., 'healthcare' becomes 'health' and 'care'). As a rule of thumb, 100 tokens correspond to approximately 75 words in English text. BERT's key advantage lies in its ability to generate context-dependent word embeddings for these tokens, allowing it to distinguish between different uses of the same word based on surrounding text.

In our application, BERT transforms each token in a job posting into a 768-dimensional vector that encapsulates its meaning in context. This allows our model to capture subtle differences in language use that are crucial for understanding job descriptions and their relationship to salaries. We use pre-trained BERT embeddings, which have been shown to perform well across a wide range of natural language processing tasks. These embeddings, trained on general-purpose text corpora, are well-suited for analyzing job postings, as both tend to use language accessible to a general audience.

Due to BERT's token limit of 512, we only process the first 512 tokens (approximately 384 words) of each job posting.[1]

Our model architecture, visualized in Figure 4 and detailed in Table 1, consists of several key components:

- BERT Embedding Layer: Transforms the input text into context-aware embeddings.
- Convolutional Neural Network (CNN): Summarizes the embeddings, effectively capturing phrases and local patterns.
- Global Max Pooling: Extracts the most salient features from the CNN output.
- Fully Connected Layers: Further process the extracted features to predict the salary.

[1] Longer postings seem to have more information about the application process and not actually about the job itself.

Greater discussion on the layers of the model and the hyperparameters are described in the Appendix. This structure allows us to leverage the rich textual information in job postings to predict salaries, going beyond simple keyword matching to understand the full context of job requirements and responsibilities.

3.2 Model Evaluation

To assess the effectiveness of our NLP model in predicting salaries from job posting text, we compare its performance against relevant benchmarks commonly used in labor economics. This evaluation is crucial for testing our central hypothesis: that the full text of job postings contains substantially more salary-relevant information than traditional categorical variables used to classify jobs.

We evaluate our model and the benchmarks using two standard metrics: the coefficient of determination (R^2) and Root Mean Square Error (RMSE). These metrics capture both explanatory power and prediction accuracy.

To ensure the robustness and generalizability of our results, we focus on out-of-sample performance. While our model is trained on 855,477 job postings from April 2019 to December 2019, all evaluation metrics reported here are based on a separate "test" set of 214,281 postings not used in the training process. This approach is essential for assessing how well our models generalize to new, unseen data and for avoiding overfitting—a common pitfall where models perform well on training data but fail to generalize to new observations.

Table 2 presents the out-of-sample performance of various models, progressing from established approaches in the labor economics literature to newer text methods. This comparison allows us to quantify the additional information captured by different approaches to analyzing the full text of job postings.

Additional details on the models used for benchmarks are available in the Appendix.

3.2.1 Models Based on Traditional Measures

As unit of analysis is the job posting, as opposed to at the individual (as wage regressions are commonly done), there is no demographic information available. I begin with models that account for fundamental salary differences across occupations, which describe the kind of work done on the job. Groshen and Levine (2002) state that detailed occupations explain two to three times the variation that demographic, education, and broad (1-digit) occupation controls. Education requirements will be captured by the posting, so it is mainly demographic information that is missing from this analysis.

The occupation code, location, and skills used as predictors in this section are structured data available from Burning Glass Technologies (BGT). It's important to note that BGT infers the occupation and skills required from each posting. We consider three primary models:

1. Occupation Fixed Effects: The Roy (1951) model emphasizes that the distribution of earnings is a function of selection into occupations. As such, a minimum standard is to allow for occupations to have different average salaries. A model using six-digit occupation codes yields an R^2 of 0.590.[2] This is significantly higher than typical individual-level earnings regressions but leaves substantial variation unexplained.

2. Occupation by MSA Fixed Effects: Jobs vary in both their requirements and salaries across place. A fully interactive model, with separate fixed effects for each occupation by MSA, would capture the variation discussed in Hershbein and Kahn (2016), by allowing for different local labor markets to have different skill requirements (and therefore, different wages) for different occupations. Incorporating location through Metropolitan Statistical Areas (MSAs) increases the R^2 to 0.695.

3. Skill Clusters: Previous work has demonstrated that skills articulated in job postings have significant predictive power for wages. Studies such as Acemoglu et al. (2022) and Deming and Kahn (2018) have utilized skill data from BGT to characterize differences within

[2] Of the postings in the test sample, almost 95% are tagged with a six-digit occupation label. Postings missing such a label are categorized separately for this analysis.

occupations across firms or MSAs. Building on this research, we incorporate BGT's detailed skill clusters into our model. These 648 granular skill categories provide specific categorizations such as ”Cybersecurity” and ”Java” within Information Technology, or ”Vehicle Repair and Maintenance” within Maintenance, Repair, and Installation. Incorporating these detailed skill clusters increases the model's R^2 to 0.765, demonstrating the additional explanatory power of fine-grained skill information beyond occupation and location.

3.2.2 Text Based Models

Moving beyond structured data, we explore models that analyze the full text of job postings:

4. TF-IDF Model: As a step towards full text analysis, we implement a tf-idf (term frequency-inverse document frequency) model. This approach represents job postings as vectors of word importance, capturing some textual information but lacking context understanding. The TF-IDF model achieves similar performance to the occupation by MSA model, explaining 69 percent of the variation in salaries. Notably, the TF-IDF model captures this substantial amount of variation without requiring explicit occupation classification. This demonstrates that free-form text data can be as valuable as information gathered through costly surveys or administrative data collection methods. The performance of this model not only showcases the potential of text analysis but also highlights the cost-effectiveness of leveraging existing job posting text for labor market insights. At the same time, it underscores the limitations of context-insensitive approaches, setting the stage for more advanced NLP techniques.

5. BERT-based NLP Model: Our most advanced model, described in detail in Section 3, uses BERT to analyze the full text of job postings. This model substantially outperforms all others, achieving an R^2 of 0.825. This means that the context dependent text can explain over 80% of the variation in metadata salaries.

The RMSE comparisons further highlight the progression in model performance. While the

RMSE decreases only slightly between models (1) and (4), our NLP model achieves a 39.3% reduction in RMSE compared to the occupation by MSA specification.

These results demonstrate that the full text of job postings, when analyzed using advanced NLP techniques, contains substantially more salary-relevant information than can be captured by traditional categorical variables or even detailed skill classifications.

3.3 Relative Role of Job Titles

The analysis above has focused occupations, skills, and other attributes found in the text. Given occupation codes are not readily available on job postings, one may ask whether job titles may sufficiently predict salary. Marinescu and Wolthoff (2020) examine the relative role of job titles in explaining salary. To do this, they look at postings on one website, CareerBuilder, in two metropolitan areas (Chicago and Washington DC) between January and March 2011. It is worth mentioning that at the time of their sample period, online job postings were often only representative of the professional labor market (Deming and Kahn, 2018).

As readily acknowledged in Marinescu and Wolthoff (2020), the number of job titles is highly skewed to the right. The 20 percent sample of posted wages — which comprises of 11,708 observations — has 4,669 job titles after cleaning the data.[3] In my sample, described in detail in Section 2, the ratio of job titles to observations is even greater. There can be many explanations for this: first, the data is collected from a later time period, in a larger number of geographies, across a greater number of job boards.[4] Each of these can explain an increase in the number of titles.

I replicate the Marinescu and Wolthoff (2020) cleaning protocol from their publicly available Stata code, and add a few additional adjustments in the spirit of their data cleaning. I then evaluate the performance of job titles in two ways:

[3] In the text, they describe job titles being free-form and cleaning the data to remove punctuation, employer names, or job locations, focusing on the first four words of the job title. The number of job titles focusing on the first four words is slightly reduced to 4,553.

[4] It is worth mentioning that additionally, because of the data collection approach used by Burning Glass Technologies, scraping errors occur in my sample at a far greater rate than in the Marinescu and Wolthoff (2020) sample (their data was collected directly from CareerBuilder).

1. Consistent with Marinescu and Wolthoff (2020), I perform in-sample prediction using all the job titles in a fixed effects regression. Effectively, there is no train or test dataset in their context and so the model is both estimated and evaluated on the same job titles.

2. I estimate the performance of job titles using the train/test split described in Section 3.2. This approach removes information coming from the job titles that do not exist in the train dataset.

The concern that arises with the first method, in-sample prediction, is the risk of overfitting. Overfitting occurs when a model, tailored closely to a specific dataset, captures not just the general trends, but also the noise particular to that dataset. As a result, while the model may produce very accurate predictions for the training data, its performance can significantly degrade when applied to new, unseen data. This may be especially likely when dealing with a large number of predictors, such as the vast number of job titles. By contrast, the train/test split method aims to mitigate this risk, providing a more robust assessment of the model's predictive performance.

In Table 3, the odd numbered columns demonstrate the predictive performance in-sample, as described in (1) above, while the even numbered columns estimate the predictive performance out-of-sample as in (2). The first two columns use all the words in a job title. The next two columns limit the job titles to their first four words, and finally the last two columns limit the job titles to their first three words. The predictive performance differs wildly across the odd and even numbered columns. While the job titles explain 91.7 percent of the variation in sample, the corresponding regression out of sample explains only 2.3 percent of the variation. This stark contrast in predictive performance, with job titles accounting for 91.7 percent of the variation in-sample but only 2.3 percent out-of-sample, underscores the model's tendency to overfit. Such a significant drop-off in explanatory power when moving from in-sample to out-of-sample analysis highlights the challenges in generalizing the model to new data.

These empirical findings can be corroborated with additional research that suggests firms strategically title workers. As described in Cohen et al. (2023), firms avoid overtime payments by titling their workers ”managers,” which makes them exempt from overtime requirements from the Fair Labor Standards Act. This strategic titling is not limited to managerial positions; for instance, Marriott's use of "guest environmental expert" for what appears to be a standard housekeeping role further illustrates this practice.

However, it is important to note that these findings do not universally apply to all datasets that use job titles. In particular, standardized sources such as the O*NET Alternate Titles list may provide a more reliable foundation for analysis because they use multiple inputs of title information. Some sources require job titles to reach a critical mass before inclusion, which helps to mitigate the issues of overfitting and strategic titling observed in more free-form datasets.

4 Words and Phrases Associated with Salaries

Having developed a model that predicts salaries from job posting text, we now turn to identifying which words and phrases are most strongly associated with higher or lower salaries. This task is challenging due to the complexity of our NLP model, which contains hundreds of millions of parameters. However, recent advances in explainable AI offer techniques for generating post hoc explanations of individual predictions made by such models (**?**). One such technique is called integrated gradients, which assigns an importance score to each input feature. Integrated gradients computes partial derivatives of the model output with respect to each input feature (Sundararajan et al., 2017). In the context of this paper, an input feature is a token (word or part of a word). Such features are not constrained to be the same within a posting – concretely, the word "help" might be positively related to salary when used in one context, while negatively related to salary in another context. This is consistent with the nuance employed when describing work in job postings.

As described in the data section, postings are capped at 512 tokens. This amounts to approximately 96 million tokens over the 214,281 postings in the test set. The added complication is that words can have different meanings - the word senior can be used to describe the experience level of a position, or used to describe the clientele of an assisted living facility. The task, therefore, of decomposing words into categories is quite complicated. I will start with a simple descriptive analysis.

Before proceeding, it is important to reiterate that the output of the integrated gradients method is post hoc explanations, and these do not have a causal interpretation. However, the importance scores can provide insight into what characteristics might be valued in jobs.

4.1 Descriptive Analysis

LImiting the sample of tokens to those that contain letters, and aggregating tokens across all postings in the test set, produces 24,297 unique tokens. About half of these tokens are used less than one hundred times. The following figures use tokens that are mentioned at least one hundred times in the test set.

The weighted integrated gradient value for a given token over the test set is the sum of the integrated gradients divided by the count of that token. Returning to our example, the word "help" might be used positively and negatively within postings and across postings. The attributions are summed over all postings in the test set and divided by the frequency. This value has a straightforward interpretation: To use a concrete example from Figure 5, the inclusion of the word "intern" in a posting is, on average, associated with a 0.004 percent decrease in salary. This may be considered small, however, this is the effect of a *single* word, holding all else equal in the posting.

The most positive and negative words are displayed in Figure 5. Of the fifteen most positive words, most are words related to job titles, such as "dentist," "practitioner," "professor," and "senior." There is one location related word in the most positive words too – "francisco," part of San Francisco, where the median household income is $119,136, close to double the median household income in the United States as of 2020.

The most negative words, in Figure 5b, are very different. These denote entry level or junior positions, such as "intern/internship," "assistant," and "entry." There are two words related to "leasing," and "cleaning," tasks that are considered low-wage work. The list also includes titles such as "aide," "hostess," "bartender," "porter," "postdoctoral," and "scribe." The list also speaks to "childhood," likely referencing early childhood educators, who are one of the lowest-paid occupations.

Finally, the negative attributions list includes “Microsoft,” likely referring to the skill of Microsoft Office Suite.[5]

Figure 6 categorizes selected tokens into four groups: locations, activities, technology skills, and titles. These selections, while not exhaustive, reveal insightful patterns. Coastal cities and states typically have higher attributions, contrasting with lower attributions for Midwestern and Appalachian locations. In terms of activities, words like “manages,” “evaluate,” “repair,” and “educate” are positively attributed, while “clean,” “plans,” “assembles,” and “patrolling” have negative attributions.

The relationship between activities (or tasks) and wages has been well-theorized (Acemoglu and Autor, 2011), with the task-based framework becoming dominant in analyzing the interplay between technology and wages. However, much of the leading research in this area, such as Acemoglu and Restrepo (2022), infers activity differences indirectly through changes in industrial composition of occupations. A notable exception is Autor and Handel (2013), which uses a representative sample of 1,333 workers (928 for within-occupation analysis) from the Princeton Data Improvement Initiative (PDII) to directly measure worker activities.

This study builds upon and extends this line of research in two key ways. First, it directly measures worker activities without relying on costly surveys, instead leveraging the rich text data from millions of job postings. Second, it relates these activities to wages on an unprecedented scale, providing a more comprehensive and nuanced understanding of the activity-wage relationship. This approach not only corroborates existing theories but also offers new insights into the fine-grained relationships between specific activities and compensation across a broad spectrum of the labor market.

The technology skills are almost all positive: “cloud,” “sql,” “java,” and “python,” for example, with the exception of the aforementioned, “microsoft,” which has a strong negative attribution. This speaks to the hedonic and not causal nature of this model. Undoubtedly, a worker with additional skills will be more valuable, yet the selection into listing Microsoft

[5] While Microsoft also refers to the firm, the number of uses of Microsoft far outweights the postings for Microsoft firm postings.

Office Suite on a posting demonstrates something about the role.

To the extent that no such quantification of technology skills valuations has been developed, however, it may, in the future and with validation, be used to help make marginal human capital investment decisions.

Finally, Figure 6(d) looks at various title related tokens. “Retail,” “driver,” and “representative,” are all associated with earnings decreases, while “nurse,” “physician,” and “administrator,” are associated with earnings increases.

4.2 Decomposition through Entity Extraction Model

There are many determinants of earnings. Unfortunately, only few of these determinants are collected at a time in commonly used datasets: the publicly available Current Population Survey contains earnings, education, experience, occupation, and location, but leaves out firm names and skills that affect earnings. Despite being the gold standard for measurement of earnings, administrative datasets in the United States rarely have occupation. Administrative data measures firm effects well, but assumes that worker effects are constant over time, and fails to decompose the attributes of workers.

Job postings have these characteristics and more. Above, I demonstrate that each token in each posting can be attributed a value. The next step is putting the text of the postings into categories. I select the following categories based on both theoretical and empirical work in economics that identifies strong relationships between these constructs and earnings:

- **Activities**: While skills are crucial determinants of wages, the specific activities or tasks that workers perform also play a significant role in earnings determination. In the Ricardian model of the labor market, described in Acemoglu and Autor (2011), workers apply their skill endowments to tasks in exchange for wages. This model posits that the allocation of skills to tasks, rather than skills alone, determines wages. Empirical support for the importance of activities comes from Autor and Handel (2013), who conducted a survey to collect new data on the job activities of a representative sample of 1,333 U.S. workers. They demonstrate that within-occupation measures of task content have significant and economically meaningful predictive power for earnings. However, their approach, while insightful, faces limitations in scale: identifying rare characteristics requires a substantial sample size, and capturing changes over time necessitates repeated surveys. This study addresses these limitations by leveraging job posting data and natural language processing techniques, facilitating scalable analysis.

- **Amenities/Disamenities**: Compensating differentials have a long tradition in labor economics, reflecting the idea that workers may accept lower wages in exchange for desirable job attributes or demand higher wages to compensate for undesirable ones. Rosen (1986) presents a theory of equalizing wage differences, identifying several empirically measurable job attributes that contribute to compensating wage differentials. These include: (1) onerous working conditions, (2) intercity and interregional wage differences associated with climate, crime, pollution, and crowding, (3) special work-time scheduling, and (4) composition of pay packages including vacations, pensions, and other fringe benefits. While some of these attributes may overlap with categories previously discussed (e.g., climate differences could be considered part of location-related attributes, and scheduling might fall under the hours category), Rosen's framework highlights additional factors, particularly fringe benefits, that contribute to wage determination. Building on this, Sockin (2021) describes six broad categories of amenities described in Glassdoor reviews, which encompass both fringe benefits and various aspects of working conditions that may be detailed in job postings.

- **Education**: Education is a fundamental determinant of earnings, with a long-standing theoretical and empirical basis in economics. Mincer (1974), in his seminal work, develops a parsimonious model of wages using education, experience, and hours of work. While Lemieux (2006) highlights that the parsimonous model may not fit the data as well following the dramatic changes in the wage structure, traditional human capital, measured by years of education, is an important component.

- **Experience**: Experience, like education, is a crucial factor in determining earnings, with its effects evolving over an individual's career. The Mincer equation, mentioned earlier, also included experience as a key variable. However, subsequent research has refined our understanding of its impact. Murphy and Welch (1990) and Lemieux (2006) highlight the importance of using a more flexible specification than the original quadratic form to accurately capture the relationship between experience and earnings.

- **Firm Name**: A related literature in labor economics focuses on high wage firms (AKM 1999). Many papers have found substantial dispersion in firm wage premia, even conditional on worker effects. The current work aims to measure the firm effect in postings.
- **General Skills**: Deming and Kahn (2018) find strong evidence that general skills, such as cognitive skills and social skills, explain substantive variation in salaries at the occupation-MSA level. Leadership skills also command a wage premium (Kuhn and Weinberger, 2005). Deming and Kahn suggest that the impact of these skills on pay might be even larger at the individual worker level. This study extends their findings by examining the impact of general skills on wages at the individual job posting level, providing a more granular analysis of skill-based wage differentials than previously possible.
- **Hours**: Working hours are a critical component in determining overall earnings, reflecting both the quantity of labor supplied and potential wage premiums for different schedules. The third element of the Mincer equation is hours, underscoring its fundamental role in wage determination. Beyond the straightforward relationship between time worked and compensation, variations in hours can capture important aspects of job quality. For example, a night-shift role may be less desirable, and be compensated with greater pay.
- **Job Titles**: Marinescu and Wolthoff (2020) find strong predictive power of job titles. As described in Section 3.3, these might not be predictive alone, but combined with other elements of a role, may explain variation in wages.
- **Location**: There is large literature in urban economics highlighting differences in wages across location. More recent work finds that the pay of better-educated workers rises

more in larger or higher mean-wage CZs. Card et al. (2023) uses the Longitudinal Employer-Household Dynamics (LEHD) data to find that half of the variation in mean wages is attributable to place effects, when controlling for workplace pay premiums and person effects. This current work extends these findings by examining variation both within and across locations, while simultaneously considering a comprehensive set of job attributes.

- **Technical Skills**: The wage premium associated with technical skills, especially computer use, has a strong empirical basis (Krueger, 1993). Though the exact nature of this premium has been debated in economics literature (DiNardo and Pischke, 1997; Autor et al., 1998), its significance in wage determination remains widely acknowledged. This study finds evidence that the premium likely reflects both returns to skills and unobserved characteristics of positions or firms. The Microsoft example in Section 4.1, where adding a skill results in a negative coefficient, clearly demonstrates that in some instances, the premium captures unobservable job attributes rather than the skill itself. Nevertheless, the evolution of technology and the increasing specificity of technical requirements call for ongoing examination of how technical skills contribute to earnings differentials.

These categories are treated as mutually exclusive in the entity extraction exercise. However, certain phrases can be ambiguous. For example, "proven track record of successful product launches" could be characterized as experience, general skills or technical skills.

I provide Gemini 1.5 with the prompt shown in Appendix C, which describes each entity category and includes certain illustrative edge-cases. Drawing on annotations from human labelers who performed the entity extraction task, I explain that ″Technical Skills″ encompasses specific software abilities and technical phrases such as ″experience with REST APIs″—clarifying that even when phrased as experience requirements, technical competencies should be classified as skills. Similarly, the prompt makes clear that a statement like ′no selling experience required′ is an attribute about experience, even though it is expressed as an absence of experience. This exemplar-based approach, known as few-shot prompting, helps guide the model's entity extraction. While finetuning could optimize performance for this specific task, I use few-shot prompting since Gemini 1.5 cannot currently be finetuned.

While the LLM effectively identifies the entities, precise positional information is needed. Therefore, I use regular expressions to find their exact textual position within the posting. This positional data is essential because it allows for exact matching with the values derived from the integrated gradients exercise.

The model operates at the token level, which is necessary because a single sentence often contains multiple entity types. Consider how the model classifies this sentence: "Description: As a logistics flow team member, you are responsible for unloading, moving and stocking freight, which can range from a wide variety of items across all departments in a Target store." The model correctly classifies 'logistics flow team member' as the job title, 'unloading, moving and stocking freight' as activities, and 'Target' as the firm name. However, most tokens are actually common words like articles and prepositions. In fact, 24 out of 36 tokens (66.67 percent) in this typical sentence do not map to any entity category.

Table 4 shows the results of the analysis at the token level for 1000 postings (results are currently being scaled up to the entire test set – over 200,000 postings). Column (1) shows the mean token share attributable to each entity, while Column (2) shows the share of integrated gradient (IG) values. The integrated gradient share represents each entity's proportion of the total magnitude of salary relevance, where both numerator and denominator use the sum of absolute values of the gradients in the posting.

Approximately 49 percent of the text in a posting falls into the above categories. The median token is unclassified (final row in the table). To make comparisons across classified text, Columns (3) and (4) repeat the shares excluding unclassified text.

The values can be interpreted as follows: 18.59 percent of the tokens in a job posting describe the activities performed at work, and activities performed at work explain 21.33 percent of the variation in salaries. Activities are, by far, the most common mention in a job posting, and the greatest contribution to salary. This is consistent with the Ricardian model of the labor market characterized by Acemoglu and Autor (2011), which emphasizes the crucial distinction between workers' skills and job tasks. Skills themselves do not directly produce output; rather, they are applied to tasks to generate production. Prior work has often focused on skills, treating education as a proxy, because education is easier to measure in traditional datasets. However, my analysis of job postings emphasizes that understanding wages requires measuring not just the skills workers possess, but also the specific tasks to which these skills are applied. This distinction becomes particularly relevant when workers with similar skill sets can perform different bundles of tasks, leading to wage variation that cannot be explained by skill measures alone. This prominence of activity/task descriptions aligns with previous work by Autor and Handel (2013), which within the framework of a Roy model, finds returns to tasks within occupation.

Amenities and disamenities are a relatively small share of both the text (2.18 percent) and predicted wages (1.36 percent). This may not be surprising because many of the amenities that are highlighted in Sockin (2021) are respect/abuse, leadership, managers, and coworkers, which may not effectively be described in a posting. Additionally, the focus on activities likely encompasses some elements of what are traditionally considered amenities in the literature. For example, Lavetti and Schmutte (2017) use average fatality risk in three digit occupation by industry cells. If the activities at work sufficiently characterize the occupations, then the variation in fatality risk they are capturing would be characterized in this paper as activities and not amenities.

The next two categories, education and experience, are the traditional measures of human capital. Both education and experience have small token shares, but larger integrated gradients shares. However, even their combined influence on wage determination (2.95 + 4.05 percent) is less than half that of activities (21.33 percent).

This contrasts with the behavior of firm names. While firm names are similarly a small part of the posting (2.33 percent), they have little weight in terms of predicting earnings (1.72 percent). This may be a function of the model not correctly capturing firm fixed effects. Despite the large sample of job postings (one million), it is likely the case that the model is underpowered to identify systematic firm wage differences of the kind documented in studies of worker movements across employers (Song et al., 2018). This is a major limitation of the model, but may be mitigated with more postings within firm. This is left for future work.

General job skills, for example, such as “welcoming and helpful attitude towards guests and other team members,” “excellent communication skills,” and “proven problem solving skills” consist of 9.59 percent of text in postings. However, they explain a lower share – 8.18 percent of the variation in salaries. These skills may not be strongly associated with wage differentials, likely because their mention in job postings does not meaningfully distinguish between jobs—employers likely value problem-solving and communication skills, whether or not they explicitly list them in the posting. This is in contrast to statements about education, experience, and technical skills, to be described later, which all have greater IG shares than token shares.

The next category is hours, which represents a small contribution to wages (2.49 percent). This category captures both employment status (part-time vs. full-time) and scheduling information (shifts, flexibility). However, since wages in this data are annualized to account for differences in hours worked, the modest predictive power of hours-related text likely reflects compensation adjustments for non-standard schedules rather than variation in total hours worked.

Job titles are, by far, the most informative content type as measured by the ratio of integrated gradient share to token share: while comprising only 3.28 percent of tokens, they explain 12.69 percent of salary-relevant content. This substantial predictive power still ranks second to activities among classified elements, suggesting that understanding what workers actually do provides more insight into wage determination than job titles alone. As described in Section 3.3, job titles have predictive power alone (e.g. Column 6 of Table 3, their unique contribution when accounting for other factors is considerably smaller, highlighting the importance of considering the full set of job characteristics.

Locations constitute 2.08 percent of tokens but affects 3.43 percent of salary variation, exhibiting the same pattern as education where mentions are sparse but earnings effects are substantial. The integrated gradients approach allows us to isolate location’s contribution to earnings while simultaneously accounting for other job characteristics—from firm identity to required skills to work activities. This ability to separate geographic effects from other job attributes is particularly valuable given Card et al. (2023)’s finding that firm heterogeneity is crucial for understanding geographic earnings premiums. The model identifies systematic patterns, some displayed in Figure 6(a), with coastal cities and California associated with higher earnings, while states like Florida, Kansas, Kentucky, Missouri and cities like Tucson command lower earnings.

Technical skills constitute 4.61 percent of tokens and explain 4.99 percent of earnings variation. While previous research has used Burning Glass Technologies' skill taxonomy to study technical skills, this approach allows the labor market itself to identify valuable technical skills through their relationship with earnings. This is particularly important given Deming and Noray (2020)'s finding that technical skills rapidly become obsolete. The integrated gradients analysis, with some specific examples in Figure 6c reveals which specific technical competencies currently command earnings premiums, without relying on potentially outdated skill taxonomies.

There are few empirical counterparts for this decomposition exercise. Deming and Kahn (2018) find that demand for cognitive and social skills accounts for five percent of the variation in firm pay, after controlling for occupation, industry, education and experience requirements. Though statistically different, this is roughly of the same magnitude as the general job skills identified in this analysis (8.18 percent), where social and cognitive skills would be categorized. However, their analysis relies on BLS OES earnings, which are only available at the occupation-MSA level, limiting their ability to capture within-occupation variation.

While other work, such as Card et al. (2023), identifies person effects in earnings, it does not differentiate the specific elements that drive these effects. This analysis breaks down job postings into specific, measurable attributes, providing concrete insights about what determines earnings. These estimates can inform worker investment decisions and help educators design curricula that align with labor market returns.

5 Conclusion

This paper develops the first natural language processing model to predict the salary of job postings using the text. With new data on salaries from the metadata of job postings, the inputs and outputs are well-defined. This lends itself to the task of supervised machine learning, where the task is to derive the function that relates text to salaries. Because text in job postings is written in commonplace language, I use the technique called transfer learning – applying knowledge gained from solving one problem to apply to this problem of salary prediction. In practice, this means that the first layer of my salary prediction model is pre-trained word embeddings from the BERT model, trained on English language Wikipedia and the Book Corpus.

My model substantially exceeds performance by any conventional baseline – a 39 percent decrease in RMSE and a 19 percent (13 percentage point) increase in R^2 compared to models with occupation by MSA fixed effects. This demonstrates that variation important for earnings can be found in the text of online job postings.

Having established that job posting text predicts earnings, I use integrated gradients, an explainability method, to identify which words drive these predictions. The analysis reveals that various tokens—describing locations, activities, technology skills, and titles—play meaningful roles in determining salary. To systematically analyze these patterns, I develop an entity extraction model that categorizes phrases into ten theoretically and empirically motivated attributes of jobs. Combining integrated gradients with entity extraction provides the first comprehensive decomposition of how different job characteristics contribute to earnings.

The analysis reveals that work activities are the primary driver of earnings variation, explaining 21.33 percent of salary-relevant content. This finding validates task-based models of the labor market. While skills and education remain important, their influence appears secondary to the actual work being performed. Moreover, the ability to directly measure tasks, rather than relying on proxies like job titles, provide additional insights into how work is valued. The results suggest that previous research may have overemphasized easily measured attributes like skills at the expense of understanding the core activities that drive production and earnings.

This decomposition of job postings creates a framework for improving labor market transparency. While financial markets have developed sophisticated tools for valuing assets, the labor market has lacked systematic ways to value job attributes and human capital investments. These estimates provide a foundation for better decision-making by workers, firms, and policymakers. Moreover, by directly measuring tasks and how they're valued, this approach offers new ways to understand how technological change affects work and wages.

This analysis has a number of important limitations. The cross-sectional nature of the data prevents us from studying temporal variation in attribute returns—for instance, how the value of non-wage amenities might fluctuate with labor market conditions, as during the COVID-19 pandemic. Additionally, the current sample size constrains our ability to precisely estimate firm effects. Perhaps most importantly, the model explains only half of the variation in tokens and less than two-thirds of the variation in salary, suggesting either methodological limitations or the existence of important factors not captured in job posting text.

Promising directions for future research emerge. As technology evolves, particularly with advances in large language models, tracking changes in returns to specific technical skills and work activities will be valuable. These estimates reflect equilibrium valuations in the labor market, and making these valuations more transparent might affect their magnitude—similar to how financial market prices adjust to new information. Such price discovery, however, is precisely the goal. This work provides a foundation for better understanding how the labor market values different aspects of jobs, enabling more informed decisions by workers, firms, and policymakers.

Figures and Tables

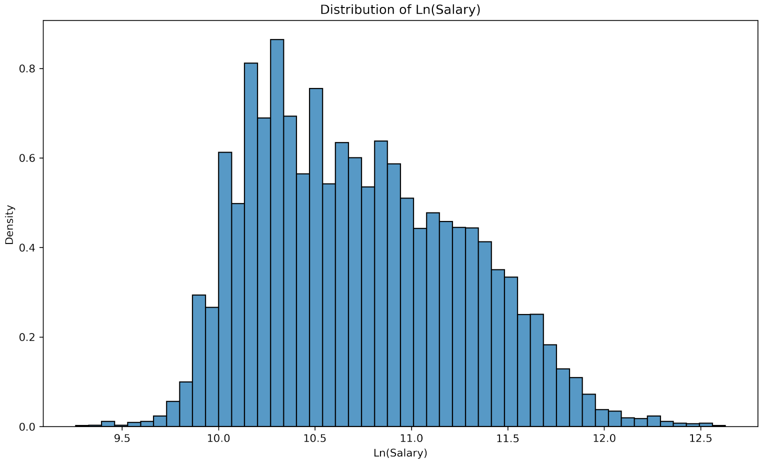


Figure 1: GHR Salary Distribution from April 2019 to December 2019

Notes: This figure describes the posted salary distribution of the 34,181,463 Greenwich.HR job postings with salary metadata posted between April 2019 and December 2019. The mean of the distribution is 53295.82.

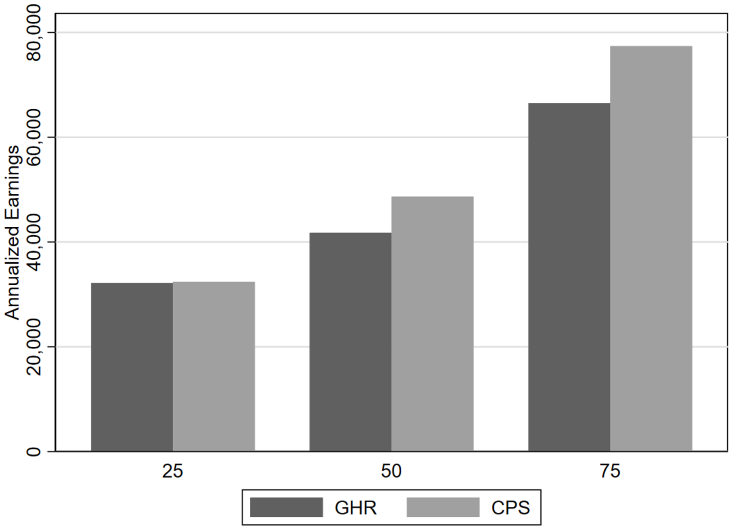
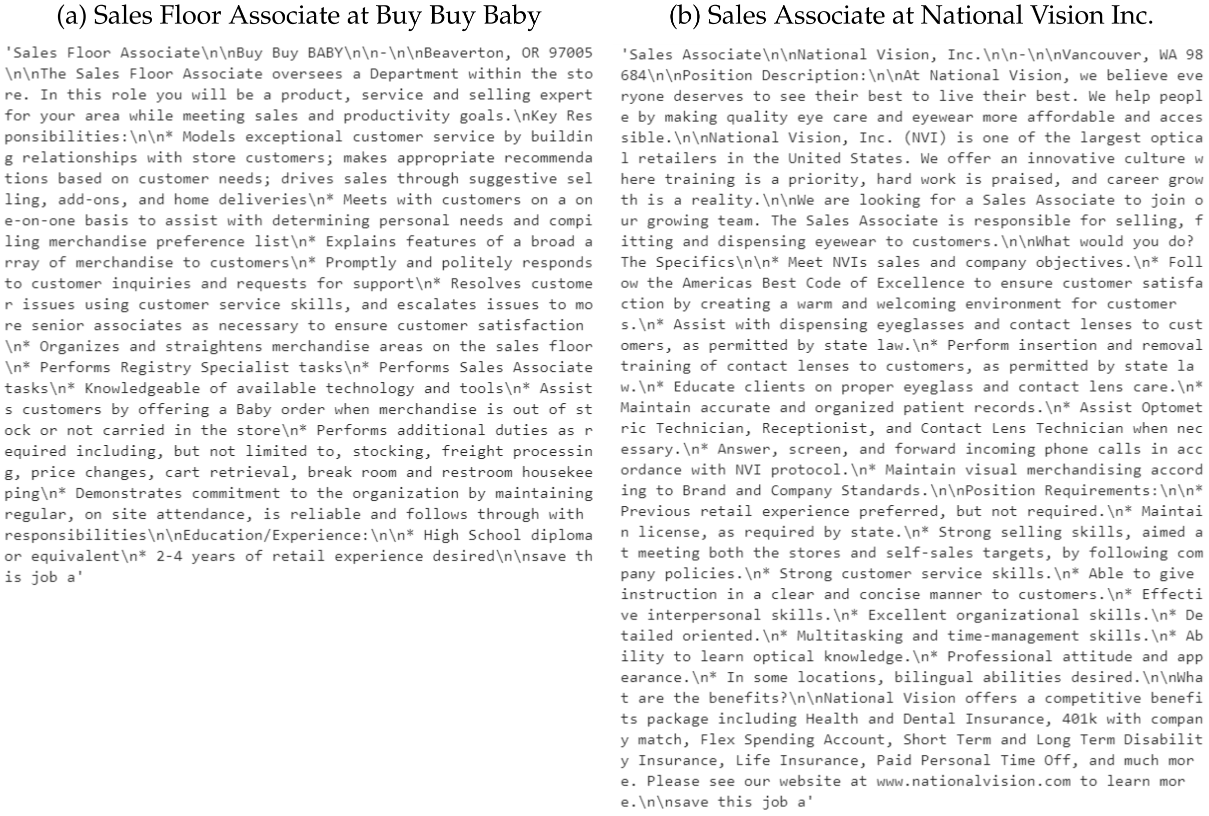


Figure 2: CPS - GHR Comparison

Notes: The Current Population Survey (CPS) values for quartiles of weekly earnings come from the Bureau of Labor Statistics' Usual Weekly Earnings of Wage and Salary Workers News Release Fourth Quarter 2019, available at `https://www.bls.gov/news.release/archives/wkyeng_01172020.htm`. CPS earnings are annualized by multiplying by 52. Data represent earnings before taxes and other deductions and include any overtime pay, commissions, or tips usually received (at the main job in the case of multiple jobholders). Greenwich.HR (GHR) salaries come from the full set of 37 million postings with salary metadata available.

Figure 3: Sample Job Postings in the Portland - Vancouver - Hillsboro Metropolitan Statistical Area



(a) Sales Floor Associate at Buy Buy Baby

```
'Sales Floor Associate\n\nBuy Buy BABY\n\n-\n\nBeaverton, OR 97005
\n\nThe Sales Floor Associate oversees a Department within the sto
re. In this role you will be a product, service and selling expert
for your area while meeting sales and productivity goals.\nKey Res
ponsibilities:\n\n* Models exceptional customer service by buildin
g relationships with store customers; makes appropriate recommenda
tions based on customer needs; drives sales through suggestive sel
ling, add-ons, and home deliveries\n* Meets with customers on a on
e-on-one basis to assist with determining personal needs and compi
ling merchandise preference list\n* Explains features of a broad a
rray of merchandise to customers\n* Promptly and politely responds
to customer inquiries and requests for support\n* Resolves custome
r issues using customer service skills, and escalates issues to mo
re senior associates as necessary to ensure customer satisfaction
\n* Organizes and straightens merchandise areas on the sales floor
\n* Performs Registry Specialist tasks\n* Performs Sales Associate
tasks\n* Knowledgeable of available technology and tools\n* Assist
s customers by offering a Baby order when merchandise is out of st
ock or not carried in the store\n* Performs additional duties as r
equired including, but not limited to, stocking, freight processin
g, price changes, cart retrieval, break room and restroom housekee
ping\n* Demonstrates commitment to the organization by maintaining
regular, on site attendance, is reliable and follows through with
responsibilities\n\nEducation/Experience:\n\n* High School diploma
or equivalent\n* 2-4 years of retail experience desired\n\nsave th
is job a'
```

(b) Sales Associate at National Vision Inc.

```
'Sales Associate\n\nNational Vision, Inc.\n\n-\n\nVancouver, WA 98
684\n\nPosition Description:\n\nAt National Vision, we believe eve
ryone deserves to see their best to live their best. We help peopl
e by making quality eye care and eyewear more affordable and acces
sible.\n\nNational Vision, Inc. (NVI) is one of the largest optica
l retailers in the United States. We offer an innovative culture w
here training is a priority, hard work is praised, and career grow
th is a reality.\n\nWe are looking for a Sales Associate to join o
ur growing team. The Sales Associate is responsible for selling, f
itting and dispensing eyewear to customers.\n\nWhat would you do?
The Specifics\n\n* Meet NVIs sales and company objectives.\n* Foll
ow the Americas Best Code of Excellence to ensure customer satisfa
ction by creating a warm and welcoming environment for customer
s.\n* Assist with dispensing eyeglasses and contact lenses to cust
omers, as permitted by state law.\n* Perform insertion and removal
training of contact lenses to customers, as permitted by state la
w.\n* Educate clients on proper eyeglass and contact lens care.\n*
Maintain accurate and organized patient records.\n* Assist Optomet
ric Technician, Receptionist, and Contact Lens Technician when nec
essary.\n* Answer, screen, and forward incoming phone calls in acc
ordance with NVI protocol.\n* Maintain visual merchandising accord
ing to Brand and Company Standards.\n\nPosition Requirements:\n\n*
Previous retail experience preferred, but not required.\n* Maintai
n license, as required by state.\n* Strong selling skills, aimed a
t meeting both the stores and self-sales targets, by following com
pany policies.\n* Strong customer service skills.\n* Able to give
instruction in a clear and concise manner to customers.\n* Effecti
ve interpersonal skills.\n* Excellent organizational skills.\n* De
tailed oriented.\n* Multitasking and time-management skills.\n* Ab
ility to learn optical knowledge.\n* Professional attitude and app
earance.\n* In some locations, bilingual abilities desired.\n\nWha
t are the benefits?\n\nNational Vision offers a competitive benefi
ts package including Health and Dental Insurance, 401k with compan
y match, Flex Spending Account, Short Term and Long Term Disabilit
y Insurance, Life Insurance, Paid Personal Time Off, and much mor
e. Please see our website at www.nationalvision.com to learn mor
e.\n\nsave this job a'
```

Notes: The job text of two sample postings in raw form from Burning Glass Technologies.

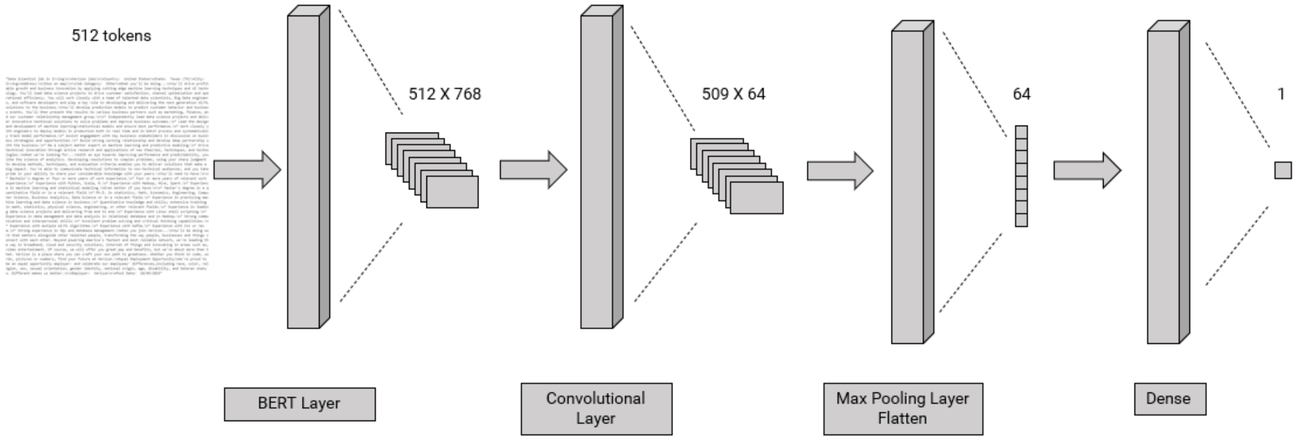
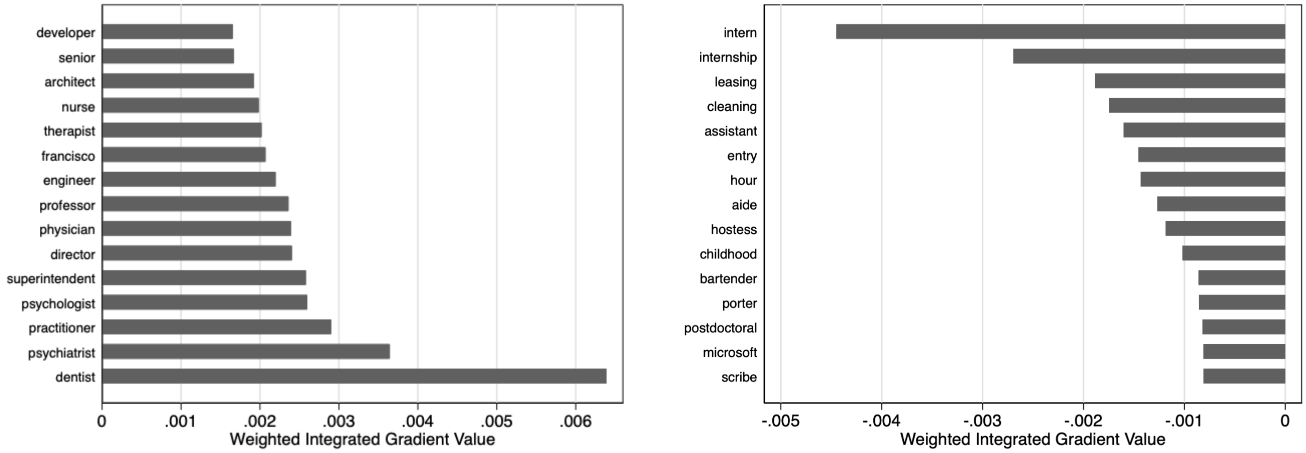


Figure 4: Model Structure

Notes: This figure displays the model structure. A job posting of 512 tokens is the input. BERT embeddings take these 512 dimensions and assign each a 768 dimensional vector depending on context. The next layer is a one dimensional convolutional layer. It is ultimately identifying n-grams that are predictive. The resulting matrix is 509 x 64. The next layer is a global max pooling layer, which captures the most relevant features from a sentence. This layer is flattened and turned into a 64 dimensional vector, which eventually predicts one dimensional ln(salary). The parameters are also laid out numerically in Table 1.

Figure 5: Words with the Most Positive and Negative Attributions

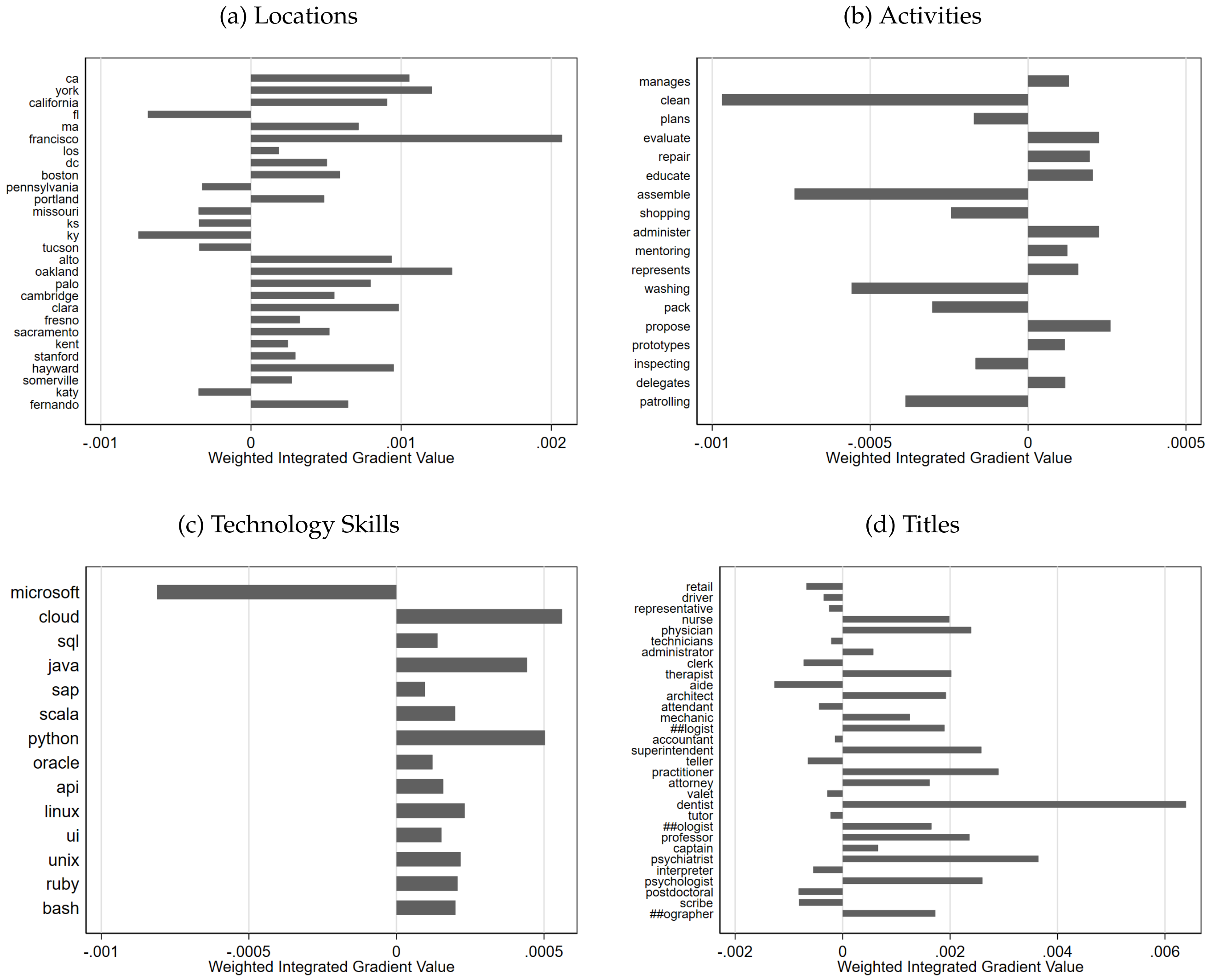


(a) Positive Attributions

(b) Negative Attributions

Notes: This figure takes the average value of the integrated gradients across postings and displays the most positively and negatively attributed tokens. The model takes ln(median annual earnings) as the dependent variable, and therefore the integrated gradients can be read as percents. For example, the inclusion of the word "intern" in a posting is, onaverage, associated with a 0.004 percent decrease in salary.

Figure 6: Select Tokens in Postings and Their Integrated Gradients



(a) Locations

(b) Activities

(c) Technology Skills

(d) Titles

Notes: The four figures show select tokens and their average integrated gradient values across the test set.

Table 1: Model Architecture

Layer Type	Dimensions	Number of Parameters
Input Layer	(X, 512)	
BERT Layer	(X, 512, 768)	109482240
Convolutional Layer	(X, 509, 64)	196672
Global Max Pooling Layer	(X, 64)	
Flatten	(X,64)	
Batch Normalization	(X, 64)	256
Dense Layer	(X, 64)	4160
Dense Layer	(X, 1)	65
Total params: 109,683,393		
Trainable params: 201,025		
Non-trainable params: 109,482,368		

Notes: This table describes the architecture of the natural language processing model used to predict salaries. In this table, X denotes the number of postings fed into the model. The input is 512 tokens of a job postings from October 2019. These 512 tokens are fed into a BERT embedding layer, where each token is given a 768 dimensional vector that is context dependent. At this point, each posting has 512 x 768 dimensions – likely too many inputs to a single salary value, so the next layers are focused on condensing dimensionality. The first step is a convolutional layer, which takes 512 x 768 dimensions, and reduces it to 509 x 64. The next layer, a global max pooling layer, takes the maximum values from this 509 x 64 matrix, which can be perceived as the most salient features, and condenses it to just 64 dimensions. The following two layers flatten and normalize these layers. Eventually, these 64 dimensions are condensed to a single dimension - the natural log of salary.

Table 2: Out-of-Sample Performance Metrics

	(1)	(2)	(3)	(4)	(5)
	Occupation	Occupation	Occupation, MSAs,	TF-IDF	BERT
	FEs	x MSA FEs	& Skill		
			Clusters		
R^2	0.590	0.695	0.765	0.689	0.825
RMSE	0.330	0.317	0.249	0.287	0.200
Occupations	785	785	785		
Locations	-	807	807		
Skill Categories			648		

Notes. This table summarizes the performance of the natural language processing model, in Column (5), to a number of relevant baselines. Relevant metrics are R^2 (coefficient of variation) and Root Mean Square Error (RMSE). The entire test set (214,281 observations) is used in every model. The outcome is ln(salary). Column (1) includes six digit Standard Occupation Classification (SOC) fixed effects. Column (2) interacts these occupation fixed effects with MSA fixed effects. Column (3) estimates a random forest regressor model with Burning Glass Technologies Skill Cluster categories. Column (4) is a TF-IDF model. Column (5) is the model described extensively in Section 3.

Table 3: MW Approach Applied to Current Data

	(1)	(2)	(3)	(4)	(5)	(6)
	Full Job Title		First Four Words		First Three Words	
R^2	0.917	0.023	0.904	0.201	0.876	0.401
Prediction	Full	Test	Full	Test	Full	Test
Number of Titles	485600		401177		296333	

Notes: This table reports the predictive performance (R^2) of job titles in explaining log salary variation. Regressions estimated are fixed effects regressions, where job titles are dummy variables. Odd-numbered columns estimate regressions where both train and test job titles are predictors, with the R^2 measured on the entire sample. Even-numbered columns repeat the exercise using only train job titles as predictors, with the R^2 measured only on the test sample. Columns (1)-(2) use all words in job titles, columns (3)-(4) use only the first four words, and columns (5)-(6) use only the first three words. Job titles are cleaned following the protocol in Marinescu and Wolthoff (2020) with additional adjustments in the same spirit. Number of titles is the number of titles in the full sample.

Table 4: Category Shares in Posting Text

	(1) Token Share	(2) IG Share	(3) Token Share (Adjusted)	(4) IG Share (Adjusted)
Activities	18.59	21.33	30.97	29.41
	(0.612)	(0.645)	(0.862)	(0.821)
Amenities/Disamenities	2.175	1.359	4.559	2.201
	(0.141)	(0.117)	(0.289)	(0.190)
Education	1.418	2.951	3.077	4.840
	(0.0705)	(0.142)	(0.149)	(0.242)
Experience	2.968	4.050	6.209	6.330
	(0.124)	(0.185)	(0.231)	(0.269)
Firm Names	2.330	1.719	6.409	3.259
	(0.0710)	(0.0783)	(0.261)	(0.181)
General Job Skills	9.594	8.184	19.56	12.83
	(0.316)	(0.332)	(0.555)	(0.489)
Hours	2.284	2.492	5.253	4.305
	(0.116)	(0.128)	(0.268)	(0.221)
Job Titles	3.277	12.69	8.597	22.51
	(0.0929)	(0.241)	(0.285)	(0.510)
Location	2.077	3.432	5.335	6.033
	(0.0728)	(0.0997)	(0.208)	(0.197)
Technical Skills and Tools	4.610	4.993	10.02	8.281
	(0.186)	(0.224)	(0.390)	(0.372)
Unclassified Text	50.67	36.80		
	(0.642)	(0.601)		

Notes: Column (1) is the mean token share attributable to each entity. Column (2) is the share of integrated gradient values. The integrated gradient share represents each entity's proportion of the total magnitude of salary relevance, where both numerator and denominator use the sum of absolute values of the gradients in the posting. Unclassified text is text that is not put into one of the categories in the model. To make comparisons across classified text, Columns (3) and (4) repeat the shares excluding unclassified text. The standard errors of the means are reported in parentheses. This analysis involves a random sample of 1000 postings, of which 996 postings in the test set have been properly processed.

References

Acemoglu, D. and D. Autor (2011): "Chapter 12 - Skills, Tasks and Technologies: Implications for Employment and Earnings," Elsevier, vol. 4 of *Handbook of Labor Economics*, 1043 – 1171.

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

Acemoglu, D. and P. Restrepo (2022): "Tasks, Automation, and the Rise in U.S. Wage Inequality," *Econometrica*, 90, 1973–2016.

Arnold, D., S. Quach, and B. Taska (2022): "The Impact of Pay Transparency in Job Postings on the Labor Market," Available at SSRN: `https://ssrn.com/abstract=4186234`.

Autor, D., C. Chin, A. Salomons, and B. Seegmiller (2024): "New Frontiers: The Origins and Content of New Work, 1940–2018*," *The Quarterly Journal of Economics*, 139, 1399–1465.

Autor, D. H. and M. J. Handel (2013): "Putting Tasks to the Test: Human Capital, Job Tasks, and Wages," *Journal of Labor Economics*, 31, S59–S96.

Autor, D. H., L. F. Katz, and A. B. Krueger (1998): "Computing Inequality: Have Computers Changed the Labor Market?*," *The Quarterly Journal of Economics*, 113, 1169–1213.

Batra, H., A. Michaud, and S. Mongey (2023): "Online Job Posts Contain Very Little Wage Information," Working Paper 31984, National Bureau of Economic Research.

Blair, P. Q. and D. J. Deming (2020): "Structural Increases in Demand for Skill after the Great Recession," *AEA Papers and Proceedings*, 110, 362–65.

Bonhomme, S., K. Holzheu, T. Lamadon, E. Manresa, M. Mogstad, and B. Setzler (2023): "How Much Should We Trust Estimates of Firm Effects and Worker Sorting?" *Journal of Labor Economics*, 41, 291–322.

Card, D., J. Rothstein, and M. Yi (2023): "Location, Location, Location," Working Paper 31587, National Bureau of Economic Research.

Cohen, L., U. Gurun, and N. B. Ozel (2023): "Too Many Managers: The Strategic Use of Titles to Avoid Overtime Payments," Working Paper 30826, National Bureau of Economic Research.

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

Deming, D. J. (2017): "The Growing Importance of Social Skills in the Labor Market*," *The Quarterly Journal of Economics*, 132, 1593–1640.

Deming, D. J. and K. Noray (2020): "Earnings Dynamics, Changing Job Skills, and STEM Careers*," *The Quarterly Journal of Economics*, 135, 1965–2005.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (2018): "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*.

DiNardo, J. E. and J.-S. Pischke (1997): "The Returns to Computer Use Revisited: Have Pencils Changed the Wage Structure Too?*," *The Quarterly Journal of Economics*, 112, 291–303.

GROSHEN, E. L. AND D. I. LEVINE (2002): "Do Rising Returns to Skills Affect Employer Wage Structures?" .

HECKMAN, J. J., L. J. LOCHNER, AND P. E. TODD (2006): "Chapter 7 Earnings Functions, Rates of Return and Treatment Effects: The Mincer Equation and Beyond," Elsevier, vol. 1 of *Handbook of the Economics of Education*, 307–458.

HERSHBEIN, B. AND L. B. KAHN (2016): "Do Recessions Accelerate Routine-Biased Technological Change? Evidence from Vacancy Postings," Working Paper 22762, National Bureau of Economic Research.

KRUEGER, A. B. (1993): "How Computers Have Changed the Wage Structure: Evidence from Microdata, 1984-1989," *The Quarterly Journal of Economics*, 108, 33–60.

KUHN, P. AND C. WEINBERGER (2005): "Leadership Skills and Wages," *Journal of Labor Economics*, 23, 395–436.

LAVETTI, K. AND I. M. SCHMUTTE (2017): "Estimating Compensating Wage Differentials with Endogenous Job Mobility," Working Paper 29, Labor Dynamics Institute, Cornell University.

LEMIEUX, T. (2006): *The "Mincer Equation" Thirty Years After Schooling, Experience, and Earnings*, Boston, MA: Springer US, 127–145.

LIN, J. (2011): "Technological Adaptation, Cities, and New Work," *The Review of Economics and Statistics*, 93, 554–574.

MARINESCU, I. AND R. WOLTHOFF (2020): "Opening the Black Box of the Matching Function: The Power of Words," *Journal of Labor Economics*, 38, 535–568.

MINCER, J. (1974): *Schooling, experience, and earnings*, National Bureau of Economic Research.

MURPHY, K. M. AND F. WELCH (1990): "Empirical Age-Earnings Profiles," *Journal of Labor Economics*, 8, 202–229.

ROSEN, S. (1986): "Chapter 12 The theory of equalizing differences," Elsevier, vol. 1 of *Handbook of Labor Economics*, 641–692.

ROY, A. D. (1951): "Some thoughts on the distribution of earnings," *Oxford economic papers*, 3, 135–146.

SOCKIN, J. (2021): "Show Me the Amenity: Are Higher-Paying Firms Better All Around?" Tech. Rep. 9842, CESifo, available at SSRN: `https://ssrn.com/abstract=3957002` or `http://dx.doi.org/10.2139/ssrn.3957002`.

SONG, J., D. J. PRICE, F. GUVENEN, N. BLOOM, AND T. VON WACHTER (2018): "Firming Up Inequality*," *The Quarterly Journal of Economics*, 134, 1–50.

SORKIN, I. (2018): "Ranking firms using revealed preference," *The quarterly journal of economics*, 133, 1331–1393.

SUNDARARAJAN, M., A. TALY, AND Q. YAN (2017): "Axiomatic Attribution for Deep Networks," in *Proceedings of the 34th International Conference on Machine Learning*, ed. by D. Precup and Y. W. Teh, PMLR, vol. 70 of *Proceedings of Machine Learning Research*, 3319–3328.

WEINBERGER, C. J. (2014): "The Increasing Complementarity between Cognitive and Social Skills," *The Review of Economics and Statistics*, 96, 849–861.

Appendix A Technical Details

A.1 Matching BGT and GHR Postings

The matching process aimed to identify corresponding job postings between Burning Glass Technologies (BGT) and Greenwich.HR (GHR) datasets. Both datasets contained job postings with fields including unique posting IDs, job dates, titles, employers, and location information, with GHR additionally providing salary data. The matching was performed on a week-by-week basis throughout 2019 to manage computational requirements and ensure temporal alignment of postings.

The matching methodology began with data preparation steps to standardize the comparison fields. A key challenge was that the two datasets used different approaches to recording geographic information. While both contained city, state, and county fields, these were often inconsistent in format and granularity between the datasets. To address this, we implemented geographic blocking using geohash encoding of the latitude and longitude coordinates with precision level 5. This approach converted geographic coordinates into a standardized format where similar geohash values represent nearby locations, allowing us to match jobs within the same geographic area even when the text-based location fields differed. Records with invalid geohash encodings (encoded as ’00000’) were removed to ensure location accuracy.

In addition to geographic standardization, company names and job titles were standardized by converting text to lowercase and replacing non-alphanumeric characters with underscores to ensure consistent comparison. We implemented the matching process using the Python recordlinkage library. The matching strategy employed the geohash-based blocking to reduce the comparison space, ensuring that only records within the same geographic area were compared. This blocking step was crucial for both computational efficiency and match accuracy, as it prevented false matches between jobs that might have similar titles and company names but were clearly in different locations.

For each potential pair within the same geographic block, we calculated similarity scores using the Jaro-Winkler string similarity method for both company names and job titles, producing scores between 0 and 1, where 1 indicates an exact match. To identify high-quality matches, we established strict thresholds requiring both company and title similarity scores to be at least 0.85 (85% similar). After applying these thresholds, we removed duplicate matches by keeping only the highest-scoring match for each job posting. This conservative matching approach prioritized match quality over quantity, ensuring high confidence in the identified pairs while acknowledging that many job postings remained unmatched due to variations in how companies and titles were recorded across the two datasets.

A.2 Data Cleaning

This version of the paper uses Burning Glass Technologies (BGT) data prior to the April-May 2020 revision. The text fields have some vestiges of the website that was scraped including the term "Save" (referencing the ability to save the job on the website), "Rating," "Review," and "Help." We remove these terms along with the Career Builder estimated salary, which occasionally shows up in some postings.

We take the first 512 tokens of the posting (though results are robust to starting randomly in a posting and taking 512 contiguous tokens).

A.3 Model and Implementation Details

The model is trained on Google Cloud Platform, using Huggingface Transformers. The procedure fine-tunes a pre-trained model in Tensorflow with Keras. The layer types and dimensions are in Table 1, with the following additional details:

- The BERT layer is BERT-base-uncased.
- The convolutional layer is one dimensional and has 64 filters (this is the dimensionality of the output space), and kernel size is 4 (to approximate phrases).
- The dense layers use a rectified linear unit activation function and a normal kernel initializer.

- The optimizer is Adam with a default learning rate.

The outcome for the model is ln(salary), where salary is the median recruiter-inputted salary. The objective function was to minimize Root Mean Squared Error (RMSE).

Appendix B Assessing the Representativeness of Greenwich.HR Data

B.1 Comparison with Current Population Survey (CPS) Data

To the best of my knowledge, there is no source of nationally representative posted salaries to compare GHR data to determine potential selection issues. The best alternative is comparing the salary distribution to the distribution of weekly earnings in the Current Population Survey. The Current Population Survey (CPS) collects earnings from one-fourth of the monthly sample, limited to wage and salary workers. The closest comparison is usual weekly earnings, representing data before taxes and other deductions, and including any overtime pay, commission or tips usually received.

I use the fourth quarter in 2019's CPS release for this comparison, graphically depicted in Figure 2. The 25th percentile of CPS weekly earnings is \$623, which at 52 weeks a year is \$32396. This is quite close to the 25th percentile of GHR salaries, at \$32175.19. The median CPS weekly value is \$936, which is an annual value of \$48,672. This is much lower than the GHR median of \$41,750. This pattern continues, with the 75th percentile of CPS earnings is \$77376 annually, while the GHR percentile is \$66501.

There can be several reasons to expect the posting distribution and the actual salary distribution to differ. The two broad categories of reasons are (1) differences in job composition and (2) differences in reporting of pay.

The posting distribution represents new jobs, and therefore, industries and occupations that have higher turnover are likely to be overrepresented. For example, according to the BLS Job Openings and Labor Turnover Survey (JOLTS), the government sector has relatively low turnover, while the private sector has higher turnover. Within the private sector, there are also notable differences: leisure and hospitality is a high turnover industry, while durable good manufacturing is low turnover. Moreover, there are notable differences within occupations. In one extreme example, seasonal work has tremendous turnover, with large fractions of Lifeguards, ski patrol, and other recreational protective service workers being rehired at the beginning of every season. Given that higher turnover jobs are more likely to be lower wage, this is consistent with the overall directional difference between the posting distribution and the CPS distribution.

Differences in job composition between the posting distribution and the actual salary distribution can also be a function of how workers are hired. First, not all jobs are posted online. Previous research on online job postings has emphasized that as online job postings have become more common, firms and jobs added more recently are lower skilled (Blair and Deming, 2020). Moreover, not all jobs are posted and some postings may still represent more than one vacancy, despite the best attempts to deduplicate. To the best of my knowledge, there is no credible estimate of the fraction of jobs that are not posted, although ongoing work by researchers at the Bureau of Labor Statistics seeks to answer this question.

Though the job composition is likely different, the CPS and GHR are also measuring different underlying concepts. The CPS usual weekly wage includes expected overtime, commission and tips. These are not included in the GHR data.

The distributions are clearly different; however, it is difficult to assess whether this is a cause for concern. Future analyses will test robustness to various assumptions about the distribution.

B.1.1 Analysis of Salary Presence in GHR Postings

Another approach to assessing the representativeness of GHR salaries is to measure how much other observable characteristics can explain whether the salary exists. Using a 20 percent random subsample of postings from April 2019 to December 2019, I regress a binary for whether the salary is missing on six digit Standard Occupation Classification (SOC) code fixed effects. If occupations that are higher wage are less likely to be well represented with salary data, then occupation fixed effects should explain considerable variation in whether the salary is missing.

Instead, I find that the pseudo R^2 on a probit regression with occupation fixed effects is only 0.0134. That is, which postings have salaries in the data cannot be explained by the occupations of those postings. This is, by no means, conclusive evidence that salaries from metadata are random. However, it does suggest that the process by which salaries appear in metadata differs from what might be expected for posted salaries.

Appendix C Prompt for Gemini Entity Extraction

Below is a job posting:

```
<jobposting>
{}
</jobposting>
```

I would like you to tag every phrase in the job posting that is related to one of the following entities: Location, General Job Skills, Technical Skills and Tools, Activities Performed at Work, Education, Experience, Amenities/Disamenities, Job Titles, Firm Name, Full-Time/Part-Time. Additional details are provided below. Also provide the placement of the phrase in the posting.

Tag each phrase separately in CSV format, each row should contain the entity name and the phrase.

Entity definitions are:

Location: should be the location of the job

- Include: Zip codes that sometimes follow the city or state (should be 5 digits) should also be included. Any time you see the location mentioned in the posting, please tag it.
- Exclude: Location of the client if travel is involved.

General Job Skills: Broad, transferable abilities applicable across various jobs and industries.

- Include: Interpersonal skills (e.g., "strong communication skills", "team player")
- Include: Cognitive abilities (e.g., "problem-solving skills", "analytical thinking")
- Include: Personal attributes (e.g., "detail-oriented", "self-motivated")
- Include: Transferable professional skills (e.g., "project management", "leadership abilities")
- Include: Relationship-building skills (e.g., "track record of building and maintaining client relationships")
- Include: adjectives describing the skills. For example "exceptional interpersonal verbal and written communication skills". Err on the side of including the whole phrase: "business intelligence to serve as a true consultative partner"

- Note: Skills in this category should be broadly applicable across different roles and industries

Technical Skills and Tools: Specific technical knowledge, proficiencies, tools, methodologies, or certifications required for the job. This includes both industry-specific and general technical skills.

- Computer programming languages, software applications, technical frameworks, APIs, version control systems (e.g., "proficiency in Python", "experience with REST APIs", "knowledge of Git")
- Healthcare technical skills and medical terminology
- Hand tools and equipment operation skills
- Word processing and office software proficiencies
- Foreign language skills
- Certifications and licenses:
  - Technical certifications (e.g., "AWS Certified Solutions Architect", "PMP certification")
  - Medical certifications (e.g., "X-ray license", "CPR certification")
  - Driver's license, including phrases like "must possess and have an acceptable driving record from the DMV"
  - Food handler's card
- Social media skills and digital marketing tools
- Industry-specific technical knowledge (e.g., "understanding of HIPAA regulations", "familiarity with Agile methodologies")
- Specific technical processes or methodologies (e.g., "expertise in A/B testing", "proficiency in financial modeling")
- Specific tools or technologies, even if mentioned with experience (e.g., "5 years of Java experience", "background in using CRM systems")
- Track records of repeated technical or accomplishments (e.g., "proven track record of successful product launches", "history of implementing efficient supply chain processes")

Note:

- Prioritize this category for any specific, learnable technical skill, tool, or certification, even if it's mentioned in the context of experience.
- Include both highly specialized skills (like programming languages) and more general technical skills (like driving or food handling) that require specific training or official approval.
- When in doubt, err on the side of including skills in this category if they involve specific tools, technologies, or skills that need special training to perform.

Activities Performed at Work: What they will be doing at work **Education:** Degree information

including anything related to majors. Include "preferred or required or recommended or the full phrase related to education"

- Include: "ba/bs in a related field", "High school diploma preferred or equivalent" – or equivalent should be included.

Experience: Time spent in a role, industry, or specific accomplishments that improve over time.

- Include: Specific durations in roles or industries (e.g., "5 years in marketing", "decade of experience in healthcare")

- Include: Phrases indicating level of experience (e.g., “entry-level”, “senior-level”, “seasoned professional”)
- Include: Track records of repeated accomplishments (e.g., “proven track record of successful product launches”, “history of exceeding sales targets”) or explicitly mentioning lacking these (e.g. “no selling training required!”)
- Include: Phrases implying accumulated knowledge over time (e.g., “extensive background in financial services”)
- Exclude: Specific technical skills or tools, even if mentioned with a time element (these should go under Technical Skills and Tools)
- Exclude: General interpersonal or soft skills (these should go under General Job Skills)

Amenities/Disamenities

- Include: Retirement plan, holidays, healthcare, other free stuff
- Other example: “Use of a company truck will also be provided, pending clearance with a clean driving record.” “Employees become eligible for company benefits after 60 days of employment”
- Exclude: Competitive pay. However, if it says “competitive pay and benefits package” or something similar, you can go ahead and highlight competitive pay as well
- Include: Travel required for the role.

Job Titles: Title and any abbreviation of title should be highlighted throughout the posting

- Notes: Even if it is not the correct title as in it has been truncated or has a typo in it, it should be listed as a title. If the title is used anywhere in the posting (for example a client success associate might be called an associate later in the posting, the title should be tagged in all cases).
- Include: The full job title: “The senior manager of pricing strategy” “associate”

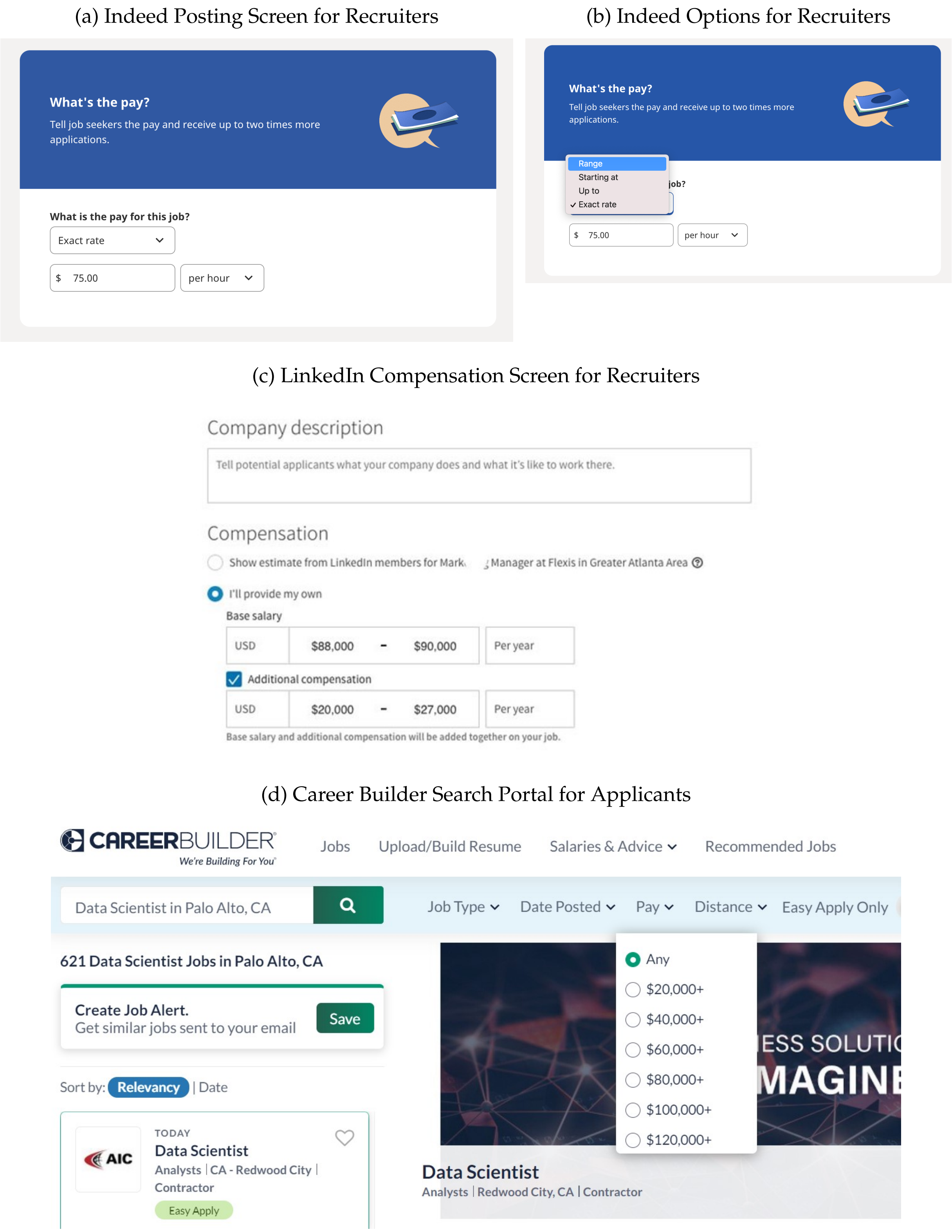
Firm Name: Firm name and potentially division if it is used in conjunction with firm name “Chapman University Argyros School of Business and Economics invites...” would include all eight words as firm name. Any other names for the firms, or abbreviations, should also be tagged.

- Include: Sometimes a posting will say “stillwater billings clinic, a billings clinic affiliate organization” - both of these characteristics should be tagged as firm name.

Full-Time/Part-Time: If it says full-time/part-time that should just be highlighted. This can also include information about the shift (morning shift, night shift)

- Include: Anything about hours or scheduling. If a role is temporary, then that information should also be included.
- Exclude: Some of the postings at the top have when they were posted, for example, “5 hours ago time estimated” – these should be excluded.

Figure A1: Screenshots of Job Board User Interfaces for Recruiters To Input Salaries



(a) Indeed Posting Screen for Recruiters

What's the pay?

Tell job seekers the pay and receive up to two times more applications.

What is the pay for this job?

Exact rate

$ 75.00 per hour

(b) Indeed Options for Recruiters

What's the pay?

Tell job seekers the pay and receive up to two times more applications.

$ 75.00 per hour

(c) LinkedIn Compensation Screen for Recruiters

Company description

Tell potential applicants what your company does and what it's like to work there.

Compensation

○ Show estimate from LinkedIn members for Mark Manager at Flexis in Greater Atlanta Area

● I'll provide my own

Base salary

USD $88,000 – $90,000 Per year

☑ Additional compensation

USD $20,000 – $27,000 Per year

Base salary and additional compensation will be added together on your job.

(d) Career Builder Search Portal for Applicants

Data Scientist in Palo Alto, CA

621 Data Scientist Jobs in Palo Alto, CA

Create Job Alert.
Get similar jobs sent to your email
Save

Sort by: Relevancy | Date

TODAY
Data Scientist
Analysts | CA - Redwood City |
Contractor
Easy Apply

● Any
○ $20,000+
○ $40,000+
○ $60,000+
○ $80,000+
○ $100,000+
○ $120,000+

Data Scientist
Analysts | Redwood City, CA | Contractor

Notes: This figure demonstrates recruiter side of job posting platforms, which provide the opportunity for recruiters to input salaries. In Panel A, a recruiter is asked the pay for the job. They are incentivized by the statement, "Tell job seekers the pay and receive up to two times more applications." In Panel B, options are displayed.