# Job Ad Data: Mapping the NLx Corpus to O*NET Features*

Stephen Meisenbacher† Svetlozar Nestorov ‡ Peter Norlander§

September 13, 2025

**Abstract**

We introduce a monthly dataset of occupation, state, and industry level features of job ads from 2015 - 2025 built from a corpus of over 155 million online job ads provided by the National Labor Exchange (NLx) Research Hub. Designed to generate novel insight on the labor market for researchers and practitioners, we extract features of the O*NET taxonomy using standard dictionaries and purpose-built machine learning models. Our dataset includes over 100,000 unique features extracted from job ads, including detailed tasks, tools, technologies, wages, organizational structure, management practices, and more. We publish the Job Ad Analysis Toolkit, a collection of open-source machine learning models and dictionaries that we use to reliably and accurately extract over 2.6 billion data points from the NLx corpus. We demonstrate the potential for research by replicating several findings regarding labor market trends, and outline potential further uses for research and practice in education and workforce development.

***Keywords:*** Job Ads, NLP methods, ML, Embeddings, data transparency

*This work has been supported (in part) by Grant # 2111-34962 from the Russell Sage Foundation and a grant from Washington Center for Equitable Growth. We acknowledge and thank the National Labor Exchange Research Hub for data access. Grant funders and The National Labor Exchange (NLx) Data Trust bear no responsibility for the analyses or interpretations of the data presented here. The opinions expressed herein, including any implications for policy, are those of the authors and not of the NLx Data Trust members. We are grateful to Loyola University Chicago for support, particularly Dr. Kathleen Bobay, Ron Price, Jason Boyda, and Joe Koral for NLP and computational infrastructure assistance. We thank attendees of the National Labor Exchange Research Hub Connect conference for feedback. This publication uses the O*NET 29.1 database and classification system and the ESCO classification of the European Commission. We publish open-source code on our Github repository at https://github.com/Job-Ad-Research-at-QSB-LUC/JAAT and our open-source language models at https://huggingface.co/loyoladatamining.

†Technical University of Munich, School of Computation, Information and Technology

‡Loyola University Chicago, Quinlan College of Business

§Loyola University Chicago, Quinlan College of Business

# 1 Introduction

The availability of online job ad data has contributed to significant advances in research and practice. However, "data access restrictions " and the "lack of standardization across private and public data sources" have until now limited the use of this data (National Academy of Sciences, 2024, p. 117). The Occupational Information Network (O*NET) is the standard taxonomy of work that serves as a cornerstone for research and professional communities by providing an information architecture for the workplace (O *NET Development, 2025). However, measurement of tasks "needs substantial work," and O*NET's survey-based data collection method is not designed for longitudinal research or timely updates (National Academy of Sciences, 2024). We make two contributions toward advancing the study of the labor market that are in alignment with the National Academy of Sciences report and top recommendations from the November 2024 national Workforce Informatics Advisory Council (Hirsch and Hui, 2024).

First, we leverage the taxonomic structure of O*NET as a basis for feature extraction from job ad data and introduce a novel, large-scale description of workplace trends in the U.S. over the last 10 years and will make it available upon publication. We create aggregate occupation, industry, state, and month level statistics derived from job ad text from September 2015 - June 2025, using data provided by the National Labor Exchange (NLx) Research Hub. The NLx Research Hub's job ad corpus is "the most accurate and comprehensive collection of real, online job openings in the United States" and provides researchers and practitioners unparalleled access to real-time insight into the labor market. Our data is relevant to academic researchers as well as workforce development and education planning professionals in community colleges and higher education. Compared to the proof-of-concept aggregate dataset we build, we are unaware of any other dataset at present with as much structured information on as large a sample of job ads, and none that adopts O*NET's framework.

Second, we develop transparent, high-accuracy, efficient, open-source natural language processing (NLP) tools to accurately map language in job ads to standard O*NET architecture and features. We provide a new GitHub repository of these software tools (the Job Ad Analysis Toolkit (JAAT)). Domain-specific open-source machine learning (ML) models are more accurate than large language models (LLM), including in the job ad domain (Nguyen et al., 2024; Zhang et al., 2022), are more efficient and scalable than LLMs, and permit independent replication of results.

The layout of the paper is as follows. Section 2 introduces NLx and O*NET data and background information on the uses and limitations of the data in academia and practice. Section 3 summarizes data built and methods used. Section 4 illustrates potential uses of the data in practice, and in research in economics and management. Section 5 indicates directions for future work and concludes.

# 2 O*NET and Job Ad Data

This section provides an overview of job ad and O*NET data. We begin by describing the O*NET architecture for occupational information, and then summarize limitations of existing O*NET data. We then describe job ad data, uses, providers, and highlight problems related to access, standardization, and transparency. We motivate the need for accurate, structured, timely labor market data from job ads – and the tools to build such data according to standard structures – following methods that adhere to standards for replicability and transparency.

## 2.1 O*NET: The Occupational Information Network

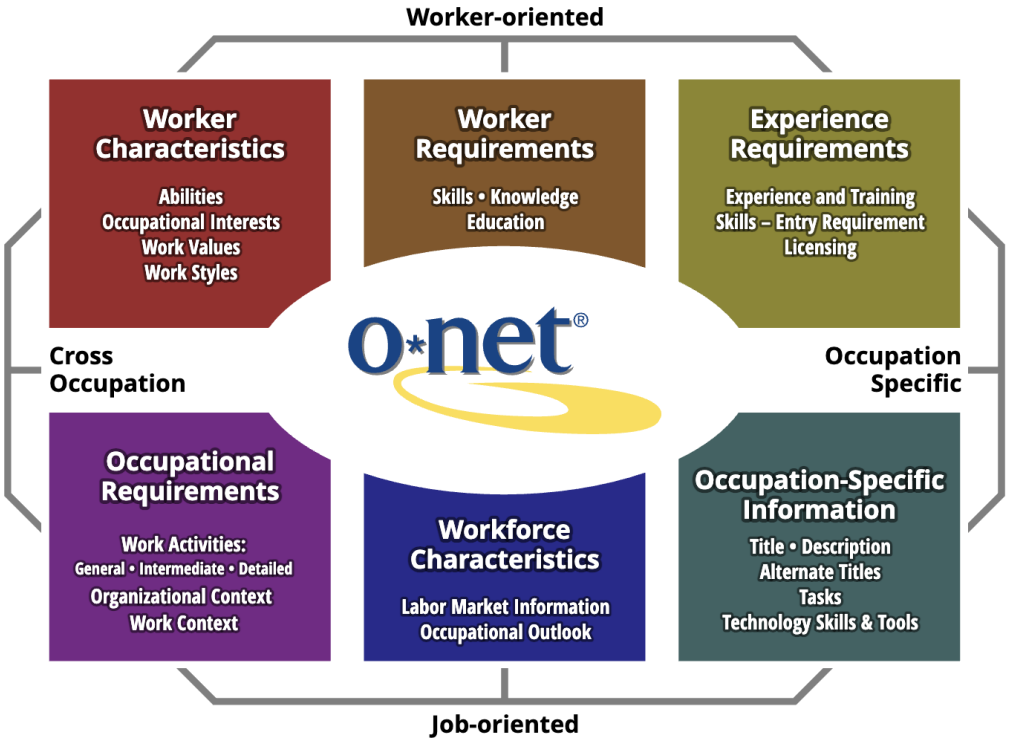
O*NET is a comprehensive database of occupational information tied to a content model that "identifies the most important types of information about work and integrates them into a theoretically and empirically sound system."[1] O*NET includes crosswalks and explicit relationships between 40 detailed tables of occupation, task, education, experience, tools, technologies, job titles, and more features of the workplace. Like other taxonomies in the sciences, O*NET is the product of efforts to develop, refine, and validate classification schemes for work, incorporating evolving individual and group judgments (Bowker and Star, 1999; Abend, 2023). O*NET's content model, displayed in Figure 1, includes "worker-oriented" (top), "job-oriented" (bottom), occupation-specific (right-side), and cross occupation (left-side) features.

Within the six major elements of the content reference model lies a hierarchical structure with increasing specificity. There are over 600 elements in the content model at five levels of depth. Each element may contain a great level of additional detail. For example, tasks (5.A.) contains a list of more than 20,000 task statements (5.A.1.) given unique codes that are linked within the O*NET database to Detailed (4.D.), Intermediate (4.E.), and Generalized Work Activities (4.A.). 5.E contains a list of over 8,000 job titles and alternative titles that are mapped to 2018 Standard Occupation Codes (SOC).

Based on survey responses, O*NET reports the Level, Importance, and Extent of specific elements within an occupation (see O*NET Scales). The Importance rating "indicates the degree of importance a particular descriptor is to the occupation. The possible ratings range from 'Not Important' (1) to 'Extremely Important' (5)." Importance covers only Tasks, Knowledge, Skills, Abilities, Work Activities, and Work Styles. Level "indicates the degree, or point along a continuum, to which a particular descriptor is required or needed to perform the occupation." Level is on a 0-7 scale, and covers Knowledge, Skills, Abilities, and Work Activities. Relevance "refers to the proportion of job incumbents who rated the provided task relevant to his/her job."

[1] See https://www.onetcenter.org/content.html, The O*NET Content Model, accessed May 27, 2025.

Figure 1: The O*NET Content Model




Note: We extract features following the structure of the O*NET Content Reference Model.

### 2.1.1 Uses of O*NET data

O*NET's measures of occupational task intensity and levels are frequently used by labor economists in what has been called the "task approach" (Autor, 2013). Work in this vein provides a richer view than traditional models of the interaction between worker skills, tasks on the job, and changes to technology or in trade (Acemoglu and Autor, 2011). Empirically, this entails dis-aggregating occupations and jobs into the tasks or bundles of tasks (high-level work activities in the O*NET structure) that comprise the job, and studying how changes in the economy impact workers who perform activities that are "routine", "non-routine", "physical", "cognitive", and "interpersonal" to research labor market trends (Deming, 2017).

Researchers often study the exposure of O*NET tasks or task bundles to a technological or other shock that is perceived to be changing the existing organization of work. One influential task-based approach follows Blinder (2009)'s study of the offshorability of jobs. Drawing from O*NET's measures of tasks, researchers calculate occupational exposure to a shock and estimate the potential impact on the labor force using a representative sample such as the Current Population Survey. This typically results in estimates of how many jobs are 'offshorable' (Blinder, 2009), teleworkable (Dingel and Neiman, 2020), automatable (Gathmann, Grimm and Winkler, 2024), impacted by Large Language

Models (Eloundou et al., 2024), etc.

Recent work by O*NET incorporates the use of ChatGPT and NLP methods to enhance the taxonomy (Lewis and Morris, 2024; Klein et al., 2025; Lewis, Gregory and Morris, 2025). Computer science researchers adopting O*NET's taxonomy also increasingly use machine learning and NLP methods to extract detailed task information from occupational text such as job descriptions (see, e.g., (Putka et al., 2023; Rounds, 2023)). Handa et al. (2025), for example, map requests from users of the large language model Claude to O*NET's list of task statements to identify the complementarity and automatability of specific tasks.

### 2.1.2 Limitations of O*NET data

Many of the limitations of O*NET below arise from researcher's misuse of data that is not longitudinal or designed to understand intra-occupational task heterogeneity. Papers that seek to measure occupational exposure to technological shocks may draw faulty conclusions. Little has changed since, over a decade ago, occupational health researchers examining the suitability of O*NET data to determine occupational exposure to risk factors issued a cautionary note advising against its use (Cifuentes et al., 2010). Citing poor statistical power, infrequent data collection, and confusing concepts operationalized however a researcher chooses to do, the authors concluded that O*NET's task-based measurement of occupations, while promising in its design, lacked proven predictive value or convergent validity, and suggested therefore that the scientific value of its data was yet to be determined.

While much of the economics literature assumes that there is change in demand for tasks over time, survey data from O*NET does not measure intra-occupational variation or change in tasks over time (Autor, 2013). O*NET's survey based data collection provides superb indicative task information for each occupation, but it is not designed to be longitudinal. Each occupation is updated infrequently, with an average of 71 observations per occupation from a single point in time, with data collection for some occupations occurring most recently in 2006, according to the metadata reported in O*NET Version 29.1. This severely impairs the credibility of research that uses O*NET data as a blunt tool to understand contemporary change within jobs and occupations.

A second limitation with the existing O*NET database is the difficulty of interpreting values for each O*NET element. The calculation of occupational measures for Importance, Level, and Relevance may be "opaque" and difficult to interpret (Autor, 2013), and subject to researcher degrees of freedom (Cifuentes et al., 2010). While surveys directly inform the Work Values (1.B.2.), Work Styles (1.C.1.), and Work Activities (4.A.) of workers in specific occupations, the crosswalk between Work Activities and the calculation of values reported in other O*NET elements, including Abilities (1.A.) and Skills (2.A.), is theoretically driven. Assumptions of the O*NET model impose requirements that all tasks

and detailed work activities exist only within a single occupation. Theoretical assumptions driving the calculations that generate skills data and task data may not be empirically justified.

A third limitation of O*NET as a taxonomic structure is insufficient detail on some elements (such as skills) that are not elaborated at the same level of detail of others (such as task). Thousands of detailed and labeled text elements are often necessary to pursue accurate Natural Language Processing (NLP) analysis that follows a taxonomic knowledge structure. For the purpose of extracting structured data from job ads, O*NET's content model is foundational, but many existing O*NET content model elements lack sufficient elaboration or adequate text descriptions for text classification and extraction purposes.

## 2.2 Job Ad Data

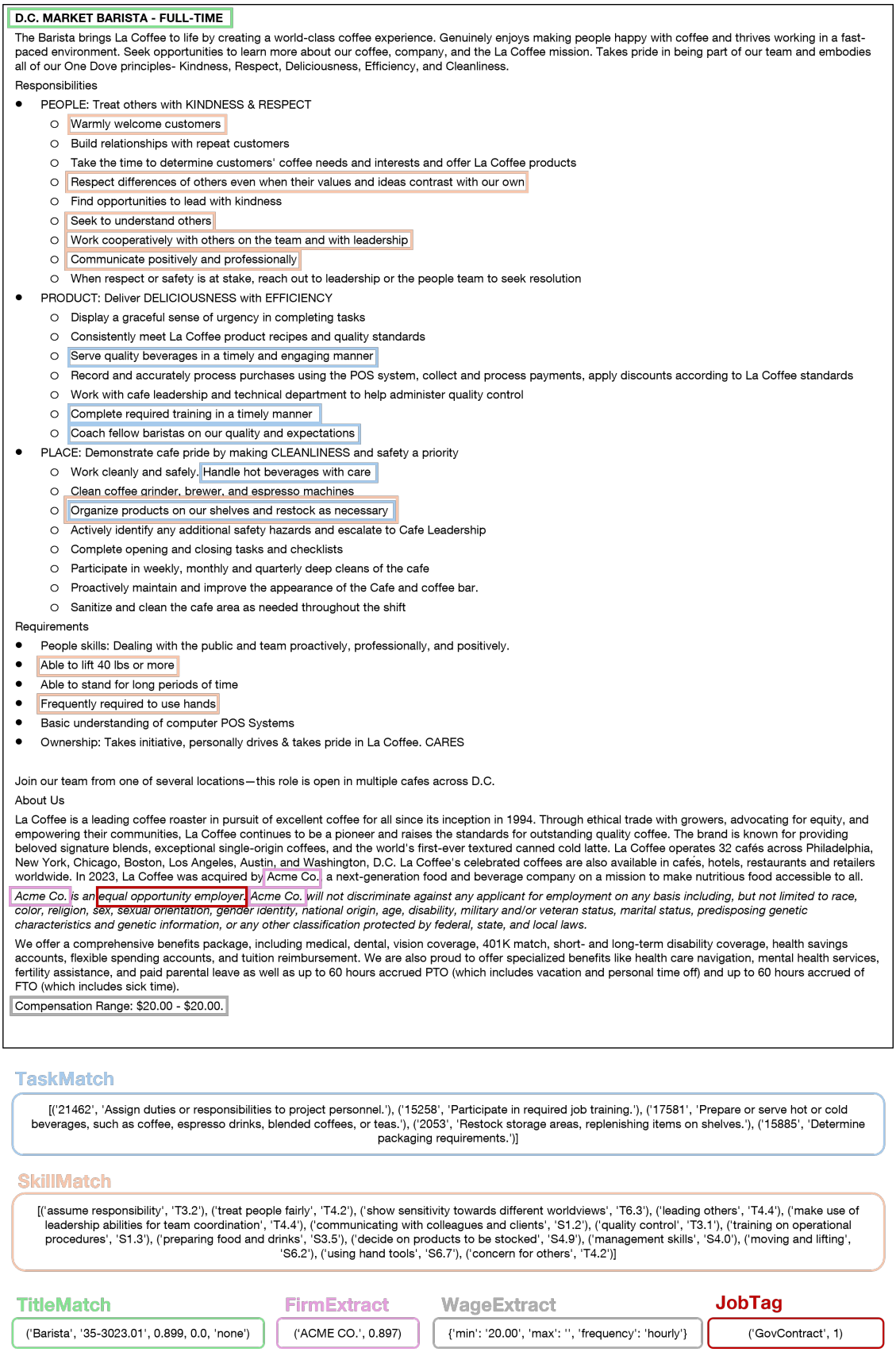
Real-time large-scale online job ad data and other newer sources of information have significantly enhanced researchers' capabilities to understand labor markets in recent decades (Horton and Tambe, 2015). For practitioners, projects started over 30 years ago have continuously delivered online job ad data to frontline workforce development professionals to help job seekers in search, referrals, and matching (Eberts and O'Leary, 2003). For over a decade, labor market intelligence data from job ads have been used by employers in workforce planning, in education and curriculum planning, career planning, and economic development (Carnevale, Jayasundera and Repnikov, 2014). Policy-makers, media, and the general public also frequently rely on aggregate job ad data to understand labor market trends.

Job advertisements often contain granular information on the tasks and skills needed to do a job, required education, licenses, qualifications and preferences, and often include details of working conditions, wages, benefits and more. To illustrate the wealth of information a job advertisement contains that can be mapped to codes from O*NET and other standard dictionaries, Figure 2 displays a job ad and the actual codes extracted with the ML tools we develop and describe in Section 3.1. While Figure 2 highlights capabilities to extract occupation information, skills requirements, task detail, firm name and industry, and wage information, a great deal of additional information that could be structured is not displayed. We separately describe how we use custom dictionaries for benefits, working conditions and culture, and recruitment practices to capture additional elements of context below.

### 2.2.1 Uses of Job Ad Data

Job ad data contributes to research on changing skills (Hershbein and Kahn, 2018; Clemens, Kahn and Meer, 2021), labor market structure (Azar, Marinescu and Steinbaum, 2022), the polarization of job

Figure 2: An illustrative job ad with features extracted by the Job Ad Analysis Toolkit (JAAT).




Note: Each individual ML tool (TaskMatch, SkillMatch, TitleMatch, FirmExtract, JobTag) is built with custom, manually audited and validated training data. Actual JAAT outputs are displayed and mapped to their approximate locations in the original job ad. We anonymize the original employer name to *La Coffee* and the parent company to *Acme Co.*

skills (Alabdulkareem et al., 2018), the importance of language in jobs (Marinescu and Wolthoff, 2020), and strategic management and recruitment strategy (Sauerwald and Norlander, 2024), and many more areas. Despite this, aggregate job ad data and other labor market information from proprietary sources used in academic papers is rarely made available. Exceptions include labor market concentration (Choi and Marinescu, 2024) and outside options (Schubert, Stansbury and Taska, 2024) data.

All job ad data have limitations, summarized well in a technical report (Lancaster, Mahoney-Nair and Ratcliff, 2019). Researchers are often careful to acknowledge and, where possible, adjust for these. As advertisements, they are employer's statements intended to attract workers, and may be less detailed than actual job descriptions, contain omissions, and inaccuracies. Online job ads are known to over-represent highly-educated workers and large firms, and to over- or under-represent certain occupations and industries. Job ad data management practices have improved over time, but a single online job posting may represent no or multiple actual vacancies (Hashizume, 2024).

### 2.2.2 Limitations of Proprietary Job Ad Data

Several companies, prominently Lightcast (formerly Burning Glass Technologies or BGT), Revelio Labs, LinkUp, and WageScape license job ad data to academic researchers for research purposes. Commercial providers typically sell access to structured data that has been built from job ad text, without disclosure of methods for creating structured information from text, or warranties or description regarding either methods or accuracy. Providers often state that they source their data from web scraping of employer webpages and job boards, often leading to duplication, and undertake proprietary processes to de-duplicate and clean the data. In general, models used to build data are proprietary and unavailable for independent use or testing.

Rising use of proprietary data in academic research risks hindering scientific advances (Lazer et al., 2020). As a source of information about work and occupations, O*NET provides invaluable insight and grounded data from worker interviews, but has not previously been combined with job ad data, in part because proprietary data providers have developed bespoke libraries and definitions of job and occupational features such as skills (National Academy of Sciences, 2024). Generally, these taxonomies are not made readily available for inspection or use and are difficult to cross-walk to standard sources like O*NET. Because many taxonomies depend upon unsupervised learning and are not combined with theory or foundational taxonomies, the choice of $k$ clusters leads to arbitrary, incompatible, and confusing definitions of skills. For reasons of replication and equity, scientific research standards include making code and data public and 'knowing your data source' (American Economics Association, 2023), and more generally, making research findable, accessible, interoperable, and reusable (FAIR) (Stall et al., 2019). Only one independent technical analysis of NLx and a major commercially provided

dataset is available: a University of Virginia team accessed both the Lightcast (BGT at the time) and NLx data to test the suitability of each data source for use in workforce development (Lancaster, Mahoney-Nair and Ratcliff, 2019). The Lightcast data is likely the most frequently used in academic research and often described as representing the 'near universe' of online job ads (Hansen et al., 2023).

Benchmarking Lightcast data against NLx in the UVA report finds that in a direct comparison of a sample of job ads in a region in a period, BGT has 24% more observations than NLx. However, 29% of BGT observations are duplicates while NLx has only 6% duplicates. According to the UVA analysis, the correlation coefficient between the number of observations in a region in the datasets is 0.996. Independent researchers' findings, summarized in the University of Virginia report, are that accuracy for education, occupation and experience fields in BGT is under 80%, there are missing values for 36% of employer names, salary is provided for 7% of observations, educational requirements are extracted for 53% of observations, and experience for 52 percent. BGT data has more structured data fields: for example, BGT's cleaned data includes an occupation family for 96.6% of job ads, while NLx had 82.7% at the time of the Virginia report.

# 3 Data

Since 2007, the National Labor Exchange (NLx) has been the leading platform for job ad distribution in the United States. NLx is a not-for-profit partnership between the Direct Employers Association, which runs the national job ad syndication network, and the National Association of State Workforce Agencies. NLx obtains data from over 300,000 employers that hire workers directly, and distributes job ads to a network of state workforce agencies and online job ad portals. Since 2021, with backing by the National Science Foundation and Bill and Melinda Gates Foundation, the NLx Research Hub has given researchers "a trusted and transparent source of job vacancy data" with a goal to "make real-time job ad information a public utility for the first time, broadening opportunities for research, analytics and product development."

Labor exchanges in the U.S. were established in 1933 under the Wagner-Peyser Act, and intermediate job-seekers and employers to facilitate efficient labor market matching while also creating opportunities to develop labor market insight from their operations (Balducchi, Eberts and O'Leary, 2004). Under the Vietnam Era Veterans' Readjustment Assistance Act (VEVRAA), federal contractors must meet job posting requirements, including that postings be filed with state unemployment offices. NLx, and America's Job Bank (AJB) before it, ensure contractor compliance with these and other recruitment-related requirements.[2]

Top recommendations in the November 2024 report of the Workforce Information Advisory Council – a group of 14 national leaders in workforce information – included strengthening the NLx, standardizing job postings data, creating pilot programs, and building tools and minimally viable data products for real-time use (Hirsch and Hui, 2024). Researchers can access the NLx corpus under an academic data use agreement. NLx's structured data fields are for the most part blank if the original creator of the job ad did not manually populate the field at the time of creation. We transform the job description, zip code, and date fields from NLx that are present over 97% of job ads to create our dataset. We use NLx wage information where available for minimum (4.6%) and maximum (4.15%) wage ranges, and build a new comprehensive wage variable including wage information we extract from job ad text. We use company name from NLx for the 61% of job ads where it is available, and transform company names into industry codes by a process described below. For wage and company name, we developed custom tools to extract these features, and then create a standardized wage variable and industry code to prepare the data for research use. For all other areas, we built entirely new data tools to extract

[2]Launched in 1995, AJB was an online job ads portal supported with funding from the U.S. Department of Labor and with input and involvement from large employers and state workforce development agencies. Free for employers and job-seekers, it was one of the most heavily trafficked websites on the early web. With more than 2.2 million monthly postings, 600,000 resumes, and 450,000 employers, it held what was the largest repository of online job ads at the time it was shuttered in 2007 (Frauenheim, 2007). The 1995-2006 archive of online job ads once managed by AJB was destroyed following defunding; attempts to recover the job ad text that was once part of AJB through Freedom of Information Act requests to the Minnesota, New York, and U.S. Department of Labor were unsuccessful.

O*NET features from job ad descriptions.

The remainder of this section introduces the major purpose and descriptive statistics for the building blocks of our aggregate dataset of job ad features. Section 3.1 introduces the Job Ad Analysis Toolkit (JAAT). Section 3.2 describes dictionaries of terms and knowledge maps we run through the job ads, including O*NET's tools and technologies dictionaries. Section 3.3 details the construction of additional variables necessary for creation of an aggregated dataset, including the month variable used in the construction of time series data and procedures for mapping from Zip Code to higher geographical levels. Appendix A describes the specific elements of O*NET structure we map to job ad features for extraction. Appendix C provides additional detail on methodology and validation procedures.

## 3.1 Methods: The Job Ad Analysis Toolkit (JAAT)

The Job Ad Analysis Toolkit (JAAT) is an open-source collection of machine tools developed for extraction of standardized information from job ads. Table 1 summarizes the machine learning models built and the multiple natural language processing methods we use to build structured data from job ad text. JAAT features include TitleMatch 3.1.1, SkillMatch 3.1.2,, TaskMatch 3.1.3, FirmExtract 3.1.4, WageExtract 3.1.5, and JobTag 3.1.6.

We adopted a trial-and-error method in iterative cycles of developing training data, building models, testing and validation, and manual augmentation of training data by "humans in the loop" (Rudin, 2018). This approach improves performance and reduces biases in language models (Choudhury, Starr and Agarwal, 2020) and provides labeled output with a high degree of correspondence with human understanding (Adadi and Berrada, 2018; Gunning et al., 2019). We track the performance of over 100 iterative stages of these models with different training data and computational approaches in a laboratory log. We searched for and tested other open-source contributions as well, but pursued novel development in nearly all cases.

We found that embedding an existing O*NET taxonomy or a newly labeled list of concepts as initial training data, searching for similar embedded sentences from a random sample of job ads, and augmenting the training data by bolstering it with highly similar sentences of real job ad text achieved promising results. Examining the output of each model, we made both manual augmentations after hand-reviewing high-frequency chunks, and semi-automated modifications of training data where we found high-precision results appearing above a specific model threshold. We choose thresholds based on strategic audits of a random sample of observations appearing at a threshold, and an examination of the distribution of the data. We generally seek a threshold where matches are 90% accurate according to manual coding of a random sample.

Where no prior knowledge base of labeled text existed, we follow in the tradition of interpretive text analysis (Gephart, 1997): we engage in inductive and deductive sampling of text, qualitatively analyze job ad language, and identify and label concepts we seek to code manually. Validity in the content analysis tradition typically starts with keyword searches, and includes strategic manual audits of high-frequency keywords and text phrases, and random manual audits of labeled output to ensure high content validity (Neuendorf, 2017).

There is no benchmark data for assessing the accuracy of NLP tools in this domain. We perform a number of post-hoc tests to compare model output to labeled data from multiple sources. We emphasize these demonstrations of convergent validity: where disaggregated public data is available, we compare model performance to those results on sentences, job titles, etc. With aggregate data, similar to the appendix of Hershbein and Kahn (2018), we also present summary descriptive statistics from the dataset we create, and compare these to benchmark statistics from BLS and Census sources. We also demonstrate how the output of independently constructed models and diverse methods of data extraction differ or converge, indicating whether the NLP tools are working as intended. We encourage users to independently test and inspect JAAT model results. Upon release of the aggregate data, users should inspect results carefully and compare them to other statistics. Appendix C provides additional detail on the methods used and known limitations with specific models. We provide the tools as is, as they are used in the construction of data, and in future directions, indicate how the iterative processes we deployed can continue to improve model performance.

Table 1: Job Ad Analysis Toolkit (JAAT) Models

| Module | Tool | Base Model | Type | # Parameters | Train Score | Validation Score |
|---|---|---|---|---|---|---|
| **TaskMatch** | Task Classification<br>https://huggingface.co/loyoladatamining/task-classifier-mini-improved2 | BERT-tiny | Fine-tuned (Binary) | 4.4M | 99.44 (F1) | 99.44 (F1) |
| | Task Matching<br>https://huggingface.co/thenlper/gte-small | GTE-small | Embedding | 30M | - | - |
| **SkillMatch** | Skill Classification<br>https://huggingface.co/loyoladatamining/skill-classifier-base | BERT-small | Fine-tuned (Binary) | 29M | 98.15 (F1) | 98.32 (F1) |
| | Skill Matching<br>https://huggingface.co/thenlper/gte-large | GTE-large | Embedding | 330M | - | - |
| **TitleMatch** | Title Matching<br>https://huggingface.co/thenlper/gte-small | GTE-small | Embedding | 30M | - | - |
| | Value Scoring<br>https://huggingface.co/loyoladatamining/title_value | DeBERTa-v3-base | Fine-tuned (regression) | 86M | 27.00 (MSE) | 34.08 (MSE) |
| | Feature Classification<br>https://huggingface.co/loyoladatamining/title_feature | DeBERTa-v3-base | Fine-tuned (Multi-label) | 86M | 81.40 (Acc.) | 81.53 (Acc.) |
| **FirmExtract** | Firm Name Extraction<br>https://huggingface.co/loyoladatamining/firmNER-v3 | DeBERTa-v3-base | Fine-tuned (Sequence) | 86M | 94.40 (F1) | 94.47 (F1) |
| **WageExtract** | Wage Classification<br>https://huggingface.co/loyoladatamining/is_pay | BERT-tiny | Fine-tuned (Binary) | 4.4M | 96.82 (F1) | 96.85 (F1) |
| | Wage Extraction<br>https://huggingface.co/loyoladatamining/wage-ner-v2 | DeBERTa-v3-base | Fine-tuned (Sequence) | 86M | 99.74 (F1) | 99.80 (F1) |
| | Wage Frequency Classification<br>https://huggingface.co/loyoladatamining/pay-freq-v2 | DeBERTa-v3-base | Fine-tuned (Multi-class) | 86M | 99.20 (F1) | 99.64 (F1) |
| **JobTag** | Job Feature Classification<br>https://github.com/Job-Ad-Research-at-QSB-LUC/JAAT | sklearn RandomForest | Trained (Binary) | - | - | - |

Note: An overview of the various language model-based tools used in the modules of the Job Ad Analysis Toolkit (JAAT). JAAT leverages a combination of pre-trained encoder-only embedding models, which are primarily used for semantic matching tasks, and fine-tuned language models, used for more specialized tasks. In the case of fine-tuning, we train a variety of models, including binary, multi-class, multi-label, and sequence classification models. The resulting models, their parameters sizes, and their training performance (on the selected validation metric) are included. Note that in the case of JobTag, we use simple RandomForest classification models.

### 3.1.1 TitleMatch

TitleMatch disambiguates job title features, returning standard SOC-O*NET codes, hierarchy, and features described below. While job titles are not perfect indicators of occupations, we follow economists (Atalay et al., 2018, 2020), epidemiologists (SOCCER) (Russ et al., 2014, 2023) and independent researchers (SOCkit) (Howison, 2022; Howison, Long and Hastings, 2023; Howison et al., 2024) in building computational models to return occupation codes from job titles.

O*NET's list of reported job titles and emerging job titles and associated occupation codes forms the basis of our model that matches job titles to occupation. The first step of TitleMatch involves a semantic matching procedure using a GTE-SMALL model, similar to that found in TaskMatch and SkillMatch. O*NET sample titles are used as a foundation, to which an instance in question is matched, following a simple nearest neighbor selection. For TitleMatch, we do not choose a minimum similarity threshold, thus always returning the best-matched title (and its corresponding occupation code from O*NET).

**Summary of Model Performance** We test TitleMatch and Sockit performance against a high-quality source of administrative data. The Department of Labor releases disaggregated Labor Condition Application Disclosure Data that employers are required to complete to lawfully place foreign-born guest workers at a worksite.[3] These include employer's self-reports of job titles mapped to occupation codes (Gibbons et al., 2019). For each foreign worker program, we combine data from 2008 – 2024 and reduce 7.5 million employer filings into unique non-null combinations of job titles and occupation codes with more than 5 observations that match a six-digit occupation code in the SOC system (n = 77,562, weighted = 2.86 million employer filings). The dataset contains occupation codes for 661 of the 867 SOC codes. No tool could match all these job title-SOC combinations from job titles alone, as codes vary within job title in the administrative data: the average number of different six-digit occupation codes per unique title in the dataset is 20.8. Job titles appear in multiple occupations for many valid reasons, including that job titles do not perfectly indicate occupations, and that human raters often disagree. Strategic behavior may also affect the selection of occupation in the LCA data, as guest worker minimum wages are set to the prevailing wage within an occupation and region (DeVaro and Norlander, 2021).

Table 2 reports the result of comparing TitleMatch and Sockit using the LCA data. TitleMatch consistently outperforms Sockit at returning occupation codes that match those assigned to job titles in the data. Weighted by the number of employer filings of a given job title and occupation code combination, TitleMatch matches 72% of the LCA data at the two-digit level , 62% at the four-digit

[3] See https://www.dol.gov/agencies/eta/foreign-labor/performance. Accessed September 8, 2025.

Table 2: TitleMatch and SockIt: Percent Job Title-SOC Match with LCA Data

| Test / Tool | 2-Digit SOC | 4-Digit SOC | 6-Digit SOC |
|---|---|---|---|
| Sockit | 0.53 | 0.39 | 0.22 |
| Sockit (Wtd.) | 0.62 | 0.51 | 0.29 |
| Sockit Matches Any Occ Within Title | 0.65 | 0.54 | 0.39 |
| Sockit Matches Any Occ Within Title (Wtd.) | 0.74 | 0.67 | 0.56 |
| TitleMatch | 0.62 | 0.49 | 0.32 |
| TitleMatch (Wtd.) | 0.72 | 0.62 | 0.47 |
| TitleMatch Matches Any Occ Within Title | 0.75 | 0.64 | 0.49 |
| TitleMatch Matches Any Occ Within Title (Wtd.) | 0.86 | 0.78 | 0.67 |
| TitleMatch Matches Data or Sockit | 0.81 | 0.70 | 0.55 |
| TitleMatch Matches Data or Sockit (Wtd.) | 0.88 | 0.82 | 0.70 |

level and 47% at the six-digit level. In terms of matching any of the occupations listed by an employer within a given job title, TitleMatch occupation codes match 86%, 78%, and 67% of cases weighted by frequency at the 2- 4- and 6-digit levels respectively, indicating that its results match analysis done by human expert coders for much of the data. Under majority voting rules (matching a job title occupation code combination identified in either LCA data or Sockit results), TitleMatch matches 88%, 82%, and 70% at the 2- 4- and 6-digit levels respectively, indicating that its results align closely with administrative data and other tools developed to code occupations from job titles. In future directions and limitations, we describe how this analysis could be used to improve TitleMatch.

**Comparison to OES** To assess the representativeness of the NLx corpus coded by TitleMatch, we contrast the distribution of job ads across major 2-digit occupational groups in the NLx data with TitleMatch to the 2-digit occupational distribution for all employed workers in the U.S. using data from the BLS Occupational Employment Statistics (OES) survey (U.S. Bureau of Labor Statistics, 2025*a*). OES data is collected in May and November of each year, and each annual OES report includes results from the prior 3 years of data collection. For each year, we first compare the percentage of jobs within each 2-digit occupation from OES to the comparable three-year data from NLx with occupation coded by TitleMatch.

For the most recent year, Figure 3 illustrates what studies of Lightcast / Burning Glass job ad data also show (Hershbein and Kahn, 2018): job ad data do not reflect the distribution of jobs in the overall economy, and certain occupational groups are significantly over- or under- represented. Similar to Burning Glass data, NLx data coded by occupation with TitleMatch has a significantly higher proportion of Management, Computer and Mathematical Occupations, Healthcare Practitioners, and Business and Financial Operations occupations than the labor market as a whole.

Figure B.7 plots changes in the correlation over the period observed. Earlier years are shaded darker, and arrows display the direction in the most recent year. It can be seen that management

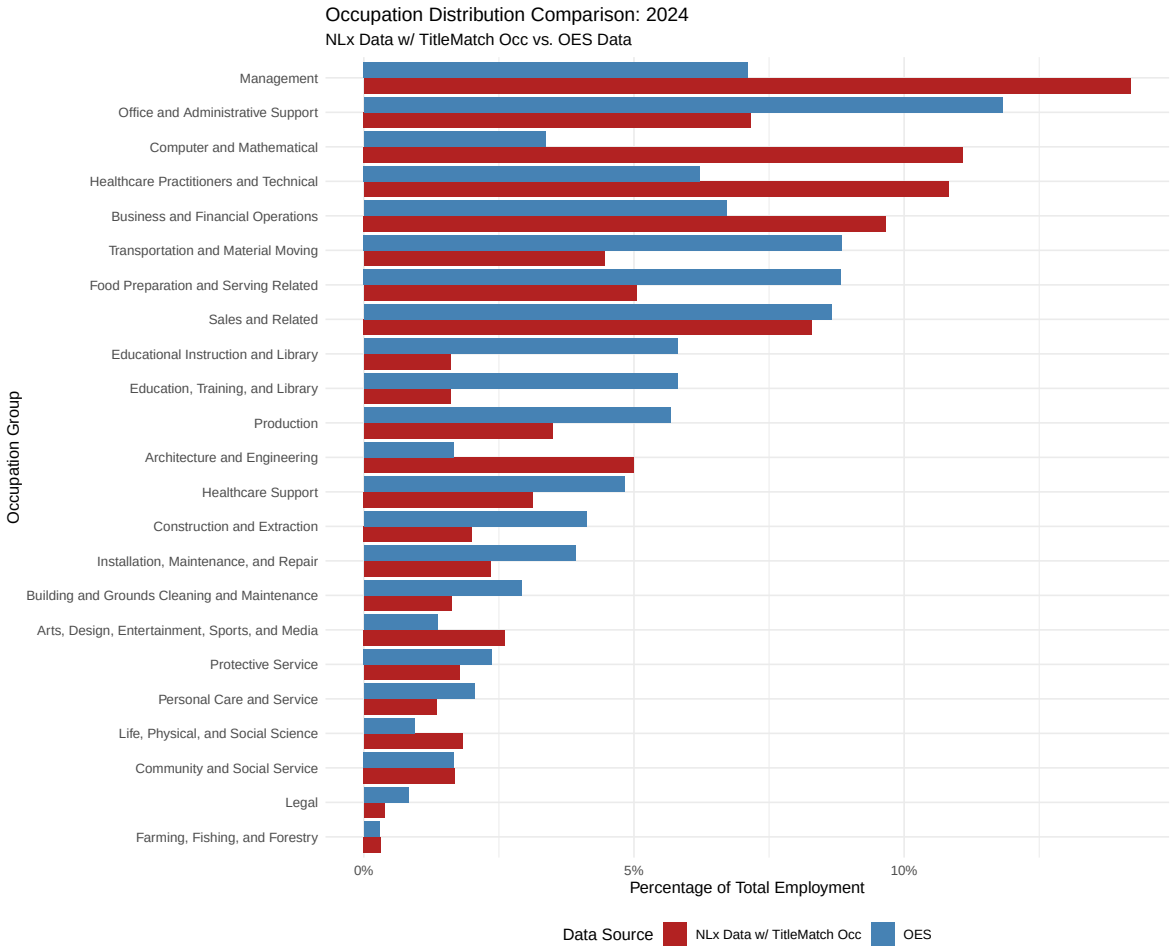



Figure 3: Occupation Distribution Comparison (3-Year Sample), 2024.

over-representation has become more significant over time, while under-representation of office and administrative support occupations has decreased over time. In several occupations, there is meaningful year-to-year fluctuation in the representativeness of the NLx data. As in Hershbein and Kahn (2018), research designs with job ad data must carefully select appropriate weights, controls, and strategies given significant changes in underlying composition of the source. Additional figures in Appendix Section B provide correlation plots and bar chart comparisons for earlier years.

**Additional Features from Titles** Following title matching to occupation, TitleMatch results also returns a hierarchy value and features of the job advertised in the title. Other features extracted from titles are presented in Appendix Table E.1. Hierarchy values and features are extracted using distinct fine-tuned DeBERTa-v3-base models, as indicated in 1. Hierarchy values returned are a number within the range [-10 through 60], as indicated in excerpt in Appendix Table C.3, where -10 represents trainees and interns, 0 represents a non-managerial role, 10 represents a first-level manager, and increasing levels of managerial responsibility increment by tens up to the Chief Executive Officer (60).

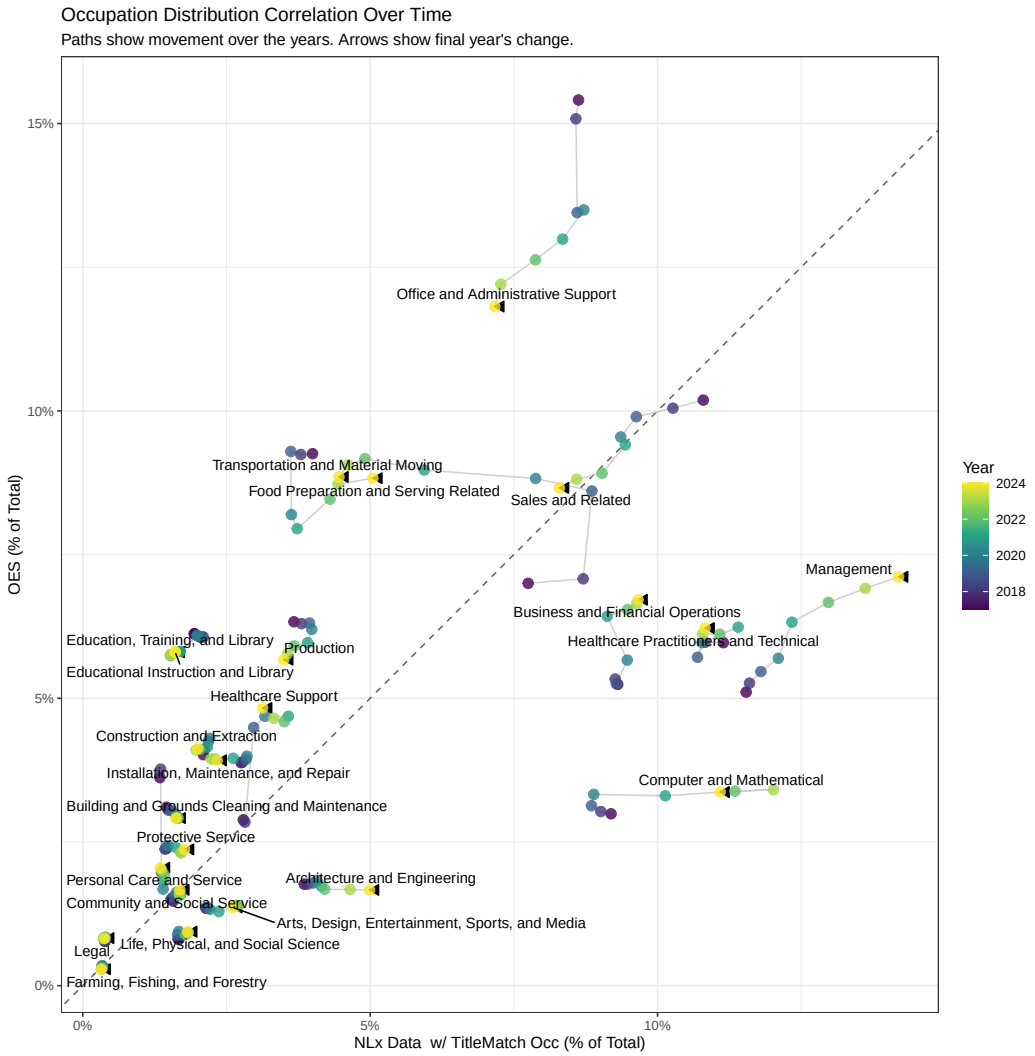



Figure 4: Trajectory of Occupation Correlation Over Time.

### 3.1.2 SkillMatch

SkillMatch follows a two-stage model to first classify "skill sentences" and then perform a semantic similarity search against a list of skills. We examined O*NET's skills taxonomy along with skill taxonomies from the European Skills, Competences, and Occupations (ESCO) database, the OECD, and the World Economic Forum (WEF). We found the ESCO v. 1.2.0 database to be the most detailed, and manually developed crosswalks between O*NET, ESCO, WEF, and OECD. We incorporated example text and labels from each of these taxonomies, and thereby increased the number of examples assigned to labels from the ESCO skills taxonomy.

As in TaskMatch, we constructed our training dataset by first beginning with the texts labeled by experts who developed the above mentioned taxonomies. These base texts were used to run a similar augmentation procedure on a random sample of 100,000 job ads, where semantic matching was performed to find the most and least similar sentences. Thus, we build a dataset with a roughly even

split of $\sim$250k "positive" skill sentence examples and $\sim$250k "negative" not-skill example sentences.

To fine-tune the binary language model-based classification model, we opted for a slightly larger version of BERT (BERT-SMALL), due to initial testing that indicated the "tiny" version was not sufficient to capture the nuances of skill sentence classification. The resulting fine-tuned model achieved a 98.32 F1 score on the validation set. Accordingly, we used a larger, more capable embedding model (GTE-LARGE) for the semantic matching portion of SkillMatch. Early analysis suggested that model accuracy dropped markedly below 0.87, was very high above 0.90, and that high-precision results could also be obtained between 0.87 and 0.89. Two independent raters coded 100 randomly selected observations within this threshold. Inter-rater reliability indicated moderate agreement (Cohen's Kappa = 0.58). This small strategic audit suggested that a threshold of 0.87 for cosine similarity provided results that were 90% accurate. This threshold was employed as a default for SkillMatch.

**Post-Hoc Validation.** Due to the unique two-stage pipeline of SkillMatch (and also the ensuing TaskMatch in Section 3.1.3), we sought the validate the performance of both stages, namely in the binary classification of skill versus non-skill statements, and subsequently the semantic matching of skill statements to skill codes. We follow a two-part validation, leveraging the LLM-as-a-Judge paradigm (Zheng et al., 2023) for an estimation of performance at scale, which is internally validated on a smaller sample by two coders.

For the validation data, we choose a different corpus of job postings, albeit with a likely similar distribution in terms of semantics. This consisted of 213k job postings between the months of March and April 2022, from the Career One Stop platform (U.S. Department of Labor, Employment and Training Administration, 2022). All of these postings were run through our SkillMatch pipeline, where we saved the individual statement-level (sentence) decisions at both stages, i.e., the binary classification, and in the case of a skill statement, the matched skill code.

The 213k job postings consisted of 5.34 million sentences, of which 2.78 million were marked by SkillMatch's classifier as being a skill statement. We formed the first validation set by randomly sampling 10k sentences marked as skill statements, and 10k marked as not. We then crafted a few-shot LLM prompt, with the task of deciding whether a given sentence was indeed a skill statement or not. This prompt is provided in Table C.1 of the Appendix.

We opted to use two LLMs for judging, one closed-source (GPT-4O-MINI) and one open-source (LLAMA-3.3-70B-INSTRUCT). We also conduct a smaller-scale validation with two human coders who are well-versed in O*NET and job posting language. These coders both perform validation on 100 samples from the validation set. The results of both the LLM and human validation are presented in Table 3.

| Validator | TPR | FPR | TNR | FNR | F1 | Agree | Strict Acc. | Lenient Acc. |
|---|---|---|---|---|---|---|---|---|
| GPT-4o-mini | 0.581 | 0.419 | 0.883 | 0.117 | 0.685 | 0.888 | 0.699 | 0.811 |
| Llama-3.3-70B | 0.733 | 0.267 | 0.821 | 0.179 | 0.767 | | | |
| Human 1 | x | x | x | x | x | x | x | x |
| Human 2 | x | x | x | x | x | | | |

Table 3: Validation results for LLM-as-a-Judge and human coders on SkillMatch binary classification. We provide True Positive, False Positive, True Negative, and False Negative rates, as well as the resulting F1 scores. In addition, we indicate how often the pairs of coders (LLMs or humans) agree, and the resulting accuracy scores in a strict setting (SkillMatch corresponds to *both* coders) or a lenient setting (corresponds to at least one).

From the 2.78 million sentences that were flagged as being skill statements, we also validate the semantic matching process, where each sentence is matched to the most similar skill code (via the code's title), and only the match results above a chosen threshold of similarity (in our case, 0.87) are kept. To illustrate how this process performs outside of its run on the full corpus, we choose a random sample of 1000 match results at all similarity scores in the range of [0.8, 1.0], rounded to two digits. In the case where 1000 results do not exist, we take the complete (maximum) number of results for that score. This resulted in a final validation set of 16597 statements, each with a corresponding matched skill.

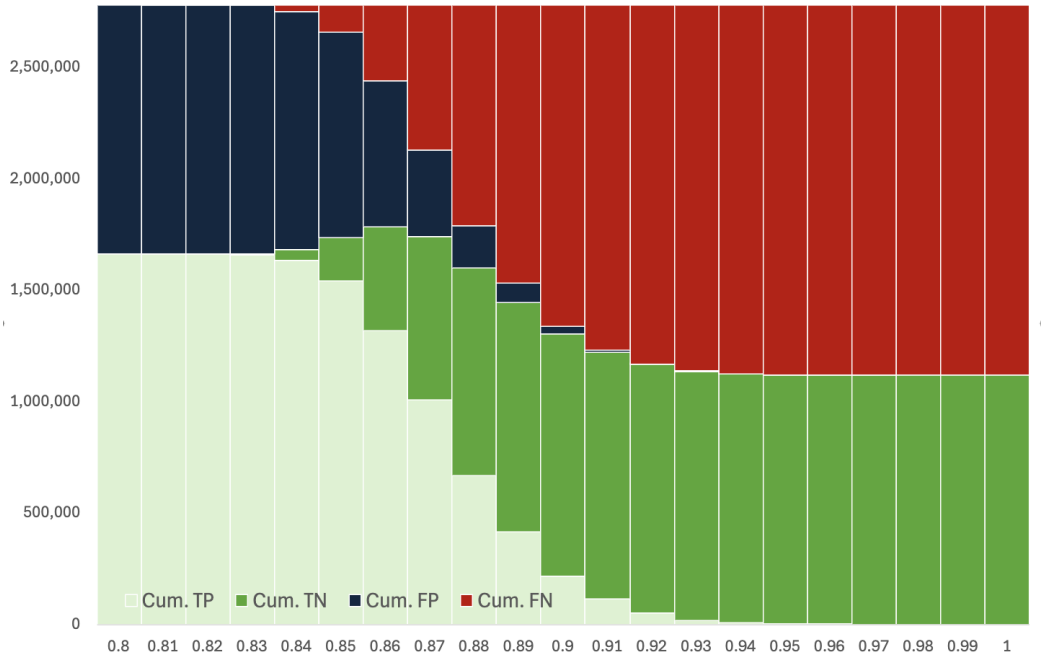
These statements were evaluated via LLM-as-a-Judge using a second prompt, found in Table C.2, which tasked the LLMs to provide a binary decision on whether the matched skill was an appropriate match or not given the skill statement. As with the first part of our validation, we peform a similar coding process with humans on a smaller set of 10 results per similarity score. The results of this validation round are presented in Table 4.

| Validator | 0.80-0.85 | 0.86 | 0.87 | 0.88 | 0.89 | 0.90 | 0.91 | 0.92 | 0.93 | 0.94 | 0.95-1.0 | Agree |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| GPT-4o-mini | 0.20 | 0.35 | 0.44 | 0.57 | 0.67 | 0.73 | 0.85 | 0.89 | 0.84 | 0.95 | 0.98 | 0.788 |
| Llama-3.3-70B | 0.52 | 0.73 | 0.82 | 0.87 | 0.90 | 0.91 | 0.94 | 0.97 | 0.93 | 0.99 | 0.99 | |
| Llama-GPT Kappa | 0.35 | 0.32 | 0.28 | 0.35 | 0.35 | 0.45 | 0.55 | 0.49 | 0.59 | 0.42 | 0.76 | |
| Human 1 | 0.39 | 0.77 | 0.87 | 0.92 | 0.85 | 0.85 | 0.85 | 0.98 | 1.0 | 1.0 | 1.0 | x |
| Human 2 | x | x | x | x | x | x | x | x | x | x | x | |

Table 4: Validation results for LLM-as-a-Judge and human coders on SkillMatch semantic matching. Results are given per similarity score (columns), and the values represent the percentage of correctly matched skills, as judged by the coders.

Figure 5 simulates of the effect of choosing a threshold somewhere between 0.8 and 1 on the proportion of True Positive, False Positive, True Negative, and False Negative rates, based on the observed distribution of 2.8 million sentences by match score as coded by SkillMatch, and based on the average estimates for accuracy by the LLMs at each threshold, as drawn from the above analysis. The visual illustrates that the initial choice of 0.87 as a threshold based on a small strategic sample is supported by the above analysis, as the optimal tradeoff between accepting False Positives and False Negatives, and the maximum sum of True Positive and True Negative is located in the range between

0.85 and 0.87.

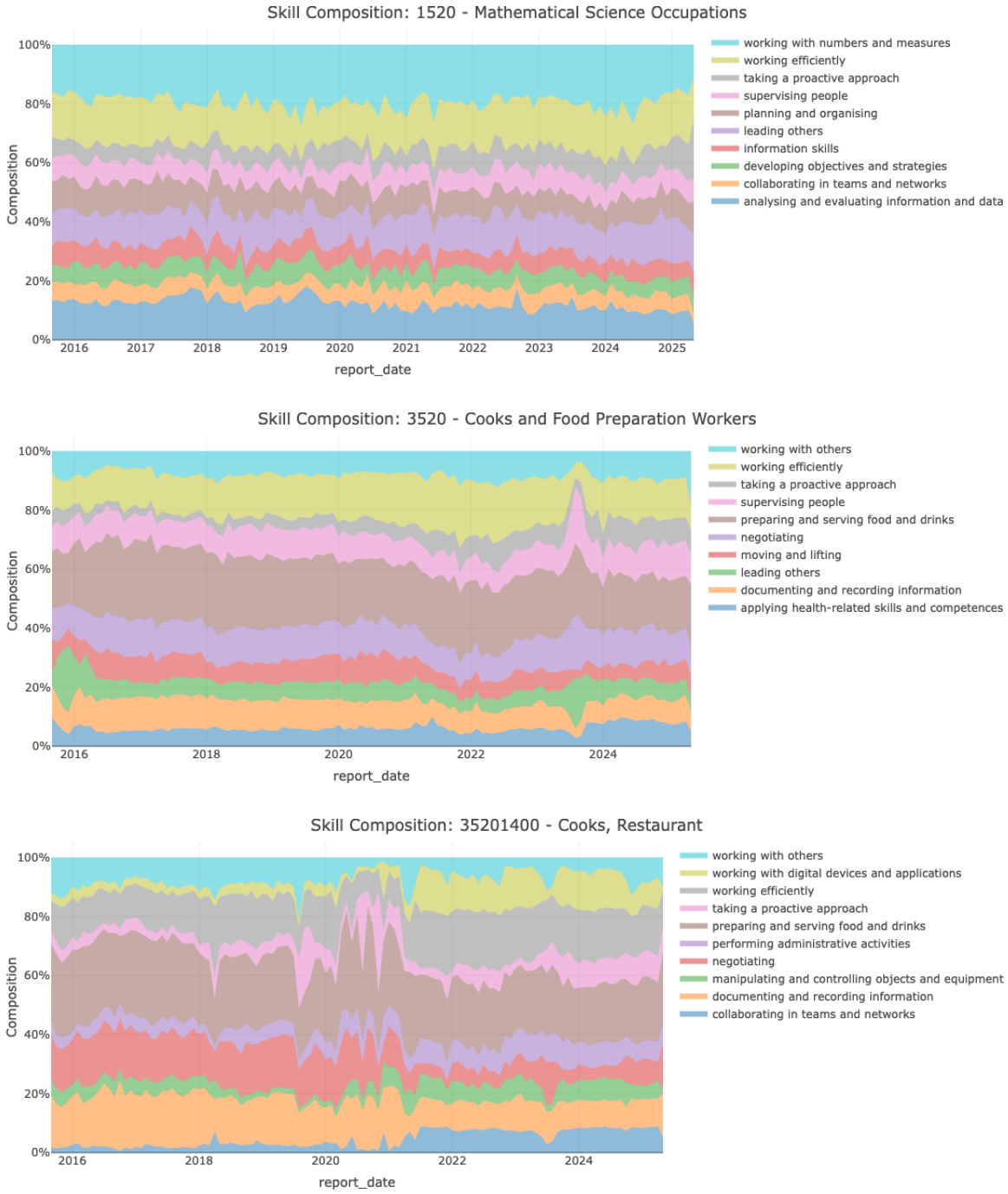



*Note:* Storing all data above a threshold that is too low yields too many false positives (dark blue). Storing only the data above a high threshold drops too many false negatives (dark red). For the construction of data, we chose a threshold above which results are high-precision based on a small strategic audit; this analysis suggests this method arrives close to the optimal threshold.

Figure 5: Estimated Distribution of TP, FP, TN, FN at Various Thresholds for 2.8M Job Ad Sentences

As with TitleMatch and TaskMatch below, the convergent validity of SkillMatch output may also be assessed by examining the aggregated data with a navigation tool. Figure 6 illustrates the top 10 SkillMatch results for two occupations at the minor group level – Mathematicians and Cooks and Food Preparation Workers – and Fast Food Cooks at the detailed occupation. Top skills for Mathematicians demonstrate month-to-month fluctuations but remain relatively stable over time. The top skill is working with numbers and measures, with analyzing and evaluating information and data, and information skills also in the top ten.

For Cooks and Food Preparation Workers, "preparing and serving food and drinks" is the top skill over the duration. There is a post-pandemic increased emphasis on working efficiently, but no apparent shock in the immediate pandemic aftermath. At the detailed occupation level, for restaurant cooks, the pandemic appears to be a seismic event, with durable aftershocks: while there is great volatility between 2020 - 2022, SkillMatch suggests that "collaborating in teams and networks" and "working with digital devices" emerge as significant top 10 post-pandemic skills desired by employers.




*Note:* The stability of skills in the Mathematician occupation can be contrasted with post-pandemic changes in skills desired in the Cooks occupation over a 10 year period. Changes are magnified at the detailed occupation level. Examining Restaurant cooks, after the COVID-19 pandemic, there is an observed increase in the demand for collaboration and digital skills.

Figure 6: Top 10 Tasks by Month for Mathematicians, Cooks, and Fast Food Cooks

### 3.1.3 TaskMatch

TaskMatch provides detailed structured information from job ads about the nature of work performed on the job. Each of over 20,000 O*NET task statements has a unique identifier, which is linked via a hierarchical taxonomy to detailed, intermediate, and general work activities, which can be cross-walked via O*NET to taxonomies of skills and abilities. TaskMatch bridges highly precise hand-created task statements based on interviews with workers (i.e., those found in O*NET) and the ability to generalize these statements to language in job ads that describe what a worker will do on the job.

TaskMatch is a two-stage model that first identifies task sentences in job ads, thereby reducing false positives and the computational overhead of running semantic matching over every sentence. The fine-tuned BERT model used in the first stage is lightweight (17 MB) and achieved an F1 score of 99.44 on the held-out validation set during training. The semantic matching between O*NET task statements and "candidate" task statements allows for expert-curated knowledge from O*NET to be scaled efficiently over large corpora, in a way that is not made possible by rigid approaches such as dictionary-based methods or computationally expensive methods such as using LLMs.

To enable independent assessment of the validity of TaskMatch and other tools, we develop a data exploration tool that provides users the ability to inspect the top 10 tasks from each month by major and detailed occupation codes, industry, and state. This allows for easy comparison to O*NET data. We restrict the data in the application to show only cells with more than 200 job postings. An illustration in Figure 7 for the Firefighter occupation demonstrates how the information extracted from job ads aligns with – and differs from existing O*NET data. Table 5 provides the top 15 tasks extracted for the Firefighter occupation (33-2011.00) and by comparison, the 6 most important tasks in the O*NET data. There are subtle differences between O*NET data for this and other occupations that suggest job ad informed O*NET task data might be preferable to O*NET's data for certain purposes.

**A. Top Tasks from NLx Job Ad Data**

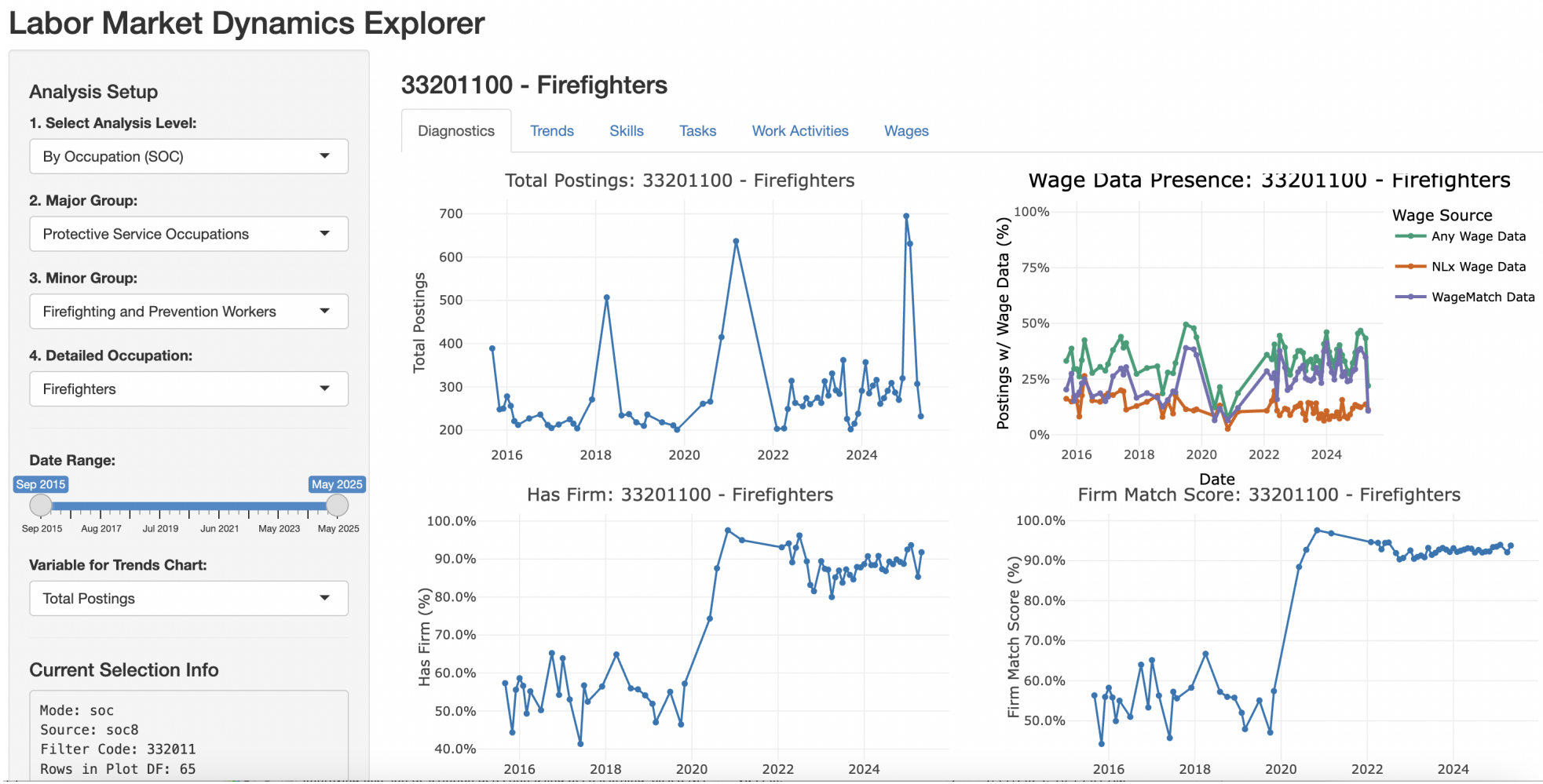
| Count | Months in Top 10 | Task |
|---|---|---|
| 15611 | 116 | Participate in firefighting efforts. |
| 8226 | 117 | Drive and operate fire fighting vehicles and equipment. |
| 6619 | 84 | Conduct wildland firefighting training. |
| 4017 | 95 | Clean and maintain fire stations and fire fighting equipment and apparatus. |
| 2441 | 56 | Work with or remove hazardous material. |
| 2164 | 48 | Rescue and evacuate injured persons. |
| 2118 | 51 | "Conduct fire |
| 1961 | 38 | "Communicate fire details to superiors |
| 1867 | 42 | Develop training materials and conduct training sessions on fire protection. |
| 1529 | 18 | Interview and hire applicants. |
| 1473 | 29 | Assign duties to other staff and give instructions regarding work methods and routines. |
| 1460 | 38 | Operate safety equipment and use safe work habits. |
| 1305 | 34 | "Direct |
| 995 | 26 | Maintain knowledge of fire laws and fire prevention techniques and tactics. |
| 905 | 23 | Supervise activities of other forestry workers. |

**B. Top Tasks from O*NET Data**

| Importance | Category | Task |
|---|---|---|
| 93 | Core | "Rescue survivors from burning buildings |
| 91 | Core | Dress with equipment such as fire-resistant clothing and breathing apparatus. |
| 90 | Core | "Assess fires and situations and report conditions to superiors to receive instructions |
| 90 | Core | "Move toward the source of a fire |
| 90 | Core | "Respond to fire alarms and other calls for assistance |
| 89 | Core | "Create openings in buildings for ventilation or entrance |
| 88 | Core | Drive and operate fire fighting vehicles and equipment. |
| 88 | Core | Inspect fire sites after flames have been extinguished to ensure that there is no further danger. |
| 87 | Core | "Position and climb ladders to gain access to upper levels of buildings |
| 87 | Core | "Select and attach hose nozzles |
| 86 | Core | Operate pumps connected to high-pressure hoses. |
| 84 | Core | "Maintain contact with fire dispatchers at all times to notify them of the need for additional firefighters and supplies |
| 84 | Core | Collaborate with other firefighters as a member of a firefighting crew. |
| 83 | Core | Patrol burned areas after fires to locate and eliminate hot spots that may restart fires. |
| 83 | Core | "Collaborate with police to respond to accidents |

*Note:* For the September 2015 - June 2025 period, 294,651 tasks are extracted from 26,987 job postings in the firefighter occupation. This table reports the top 15 based on data from an aggregated summary table on the top 10 tasks for each month and occupation.

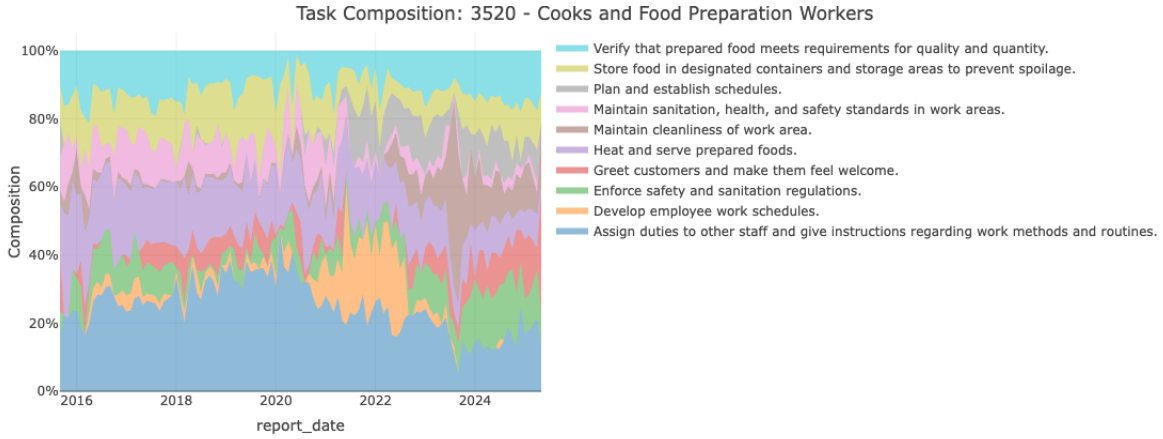
Table 5: The Top Fifteen Tasks for Firefighters in NLx Job Ad Data and O*NET Data.




*Note:* An illustrative example of diagnostics information for the firefighter occupation using the Job Ad exploration tool. This illustrates the number of postings each month, the availability of wage and firm (industry) data for the occupation, and the average match scores for occupation and industry matching for this occupation. Data for from months with fewer than 200 observations are suppressed.

Figure 7: A Job Ad Data Exploration Tool with Shiny

Data exploration of subgroups can provide additional demonstrations of the convergent validity of TaskMatch and other tools. As discussed above, a limitation of O*NET is the static representation of occupational tasks based on data collection from a single point in time. Many labor market observers assume that there is change in tasks within occupations over time. The data exploration tool enables users to investigate trends and change over time, and assists in assessing convergent validity in combination with knowledge of specific occupational and industry on trends involved. For example, Figure 8 illustrates change in the top 10 tasks over time for the minor occupation group "3520 - Cooks and Food Preparation Workers," an occupation affected greatly by pandemic health concerns and post-pandemic labor shortages often referred to as the "Great Resignation." In the 2020-2022 time period, task statements "Maintain sanitation, health, and safety standards" and "Developing employee work schedules" grew as a share of the top 10 tasks sought by employers, corresponding to trends described in trade publications (Littman, 2021).




*Note:* The increased share of tasks involving managing employee work schedules (orange) during the 2020-2022 period aligns with post-pandemic labor shortages and the "great resignation."

Figure 8: Top 10 Tasks by Month for Cooks

### 3.1.4 FirmExtract

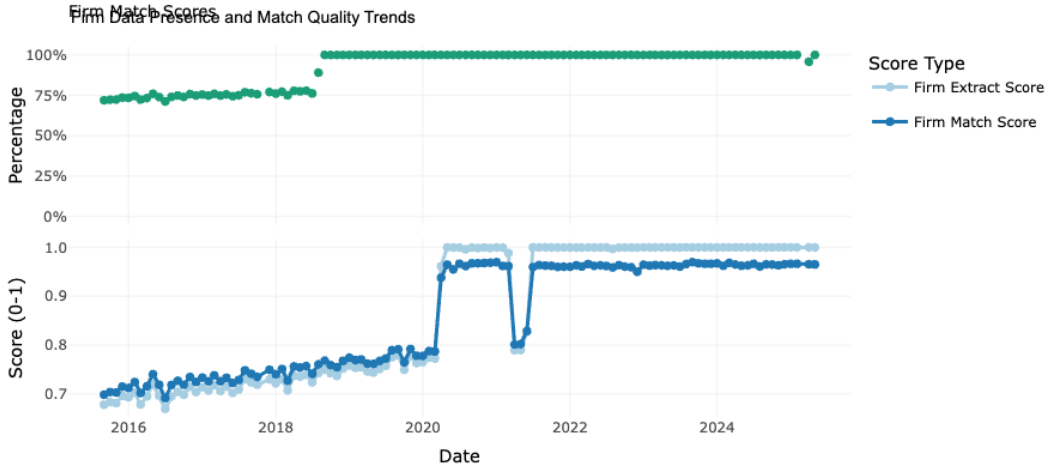
FirmExtract retrieves the firm name from the text description of job ads (FirmNER), with additional capabilities to clean and standardize firm names, and perform a similarity match to other sources of firm name information. In both the NLx descriptive data we examine, and in research using the Lightcast data, the metadata for firm is missing for approximately 40% of observations (Hershbein and Kahn, 2018; Lancaster, Mahoney-Nair and Ratcliff, 2019). We train a custom NER (Named Entity Recognition) model ("firmNER") to extract names from job ad text. FirmNER is created by fine-tuning a DeBERTa-v3-base model on quality labeled data – a large sample of job ad data with firm name present in the metadata.

In the next steps of FirmExtract, the extracted sequence representing the firm name is standardized

and fuzzy matched to an existing collection of known firm names in the United States. We standardize all extracted firm names using common firm record-linkage cleaners (Wasi and Flaaen, 2015). This cleaning protocol standardizes firm names that can be subject to multiple spellings: "Seven-Eleven", "7-11 Inc.", etc. We then fuzzy match firm names from job ads to firm names in a yearly file of U.S. establishments licensed from Data Axle for 2015-2023. Data Axle's information includes a unique establishment ID, and indicates relationships between establishments, subsidiaries, and parent companies. Data Axle fields also include industry (SIC and NAICS) for all firms, and sales volume, and number of employees for many observations.

The top panel of Figure 9 illustrates the percentage of job ads each month that could be matched to a unique firm ID, and thus industry. The bottom panel illustrates the average confidence score of the extraction and the match for the duration of the period. After extraction, firm names are available for approximately 75% of job ads in the NLx corpus prior to major improvements in data collection by NLx in 2018, after which we are able to obtain a firm name and a link to industry for nearly 100% of job ads.

Figure 9: Industry Availability in the Data




The percent of job advertisements matched to a specific firm (top) and the model score / confidence in the match (bottom).

To assess the representation of job postings by industry, we compare the distribution of job ads across major 2-digit industry groups in the NLx data with FirmExtract to the distribution for all employed workers in the U.S. with data from the Quarterly Census of Employment and Wages (QCEW) (U.S. Bureau of Labor Statistics, 2025*b*), as in the Hazell et al. (2022) analysis using Lightcast data. We take the average share of employment by 2-digit industry (QCEW) and the average share of job postings (NLx with FirmMatch Industry) for each year and plot the comparison in Figure 10. While there are noticeable differences between job ad data from NLx and similar figures built with

Lightcast data, there are similarities in that NLx postings with sector from FirmExtract significantly under-represent the manufacturing, construction, and accommodation and food services sectors, and over-represent information and professional, scientific, and technical services.

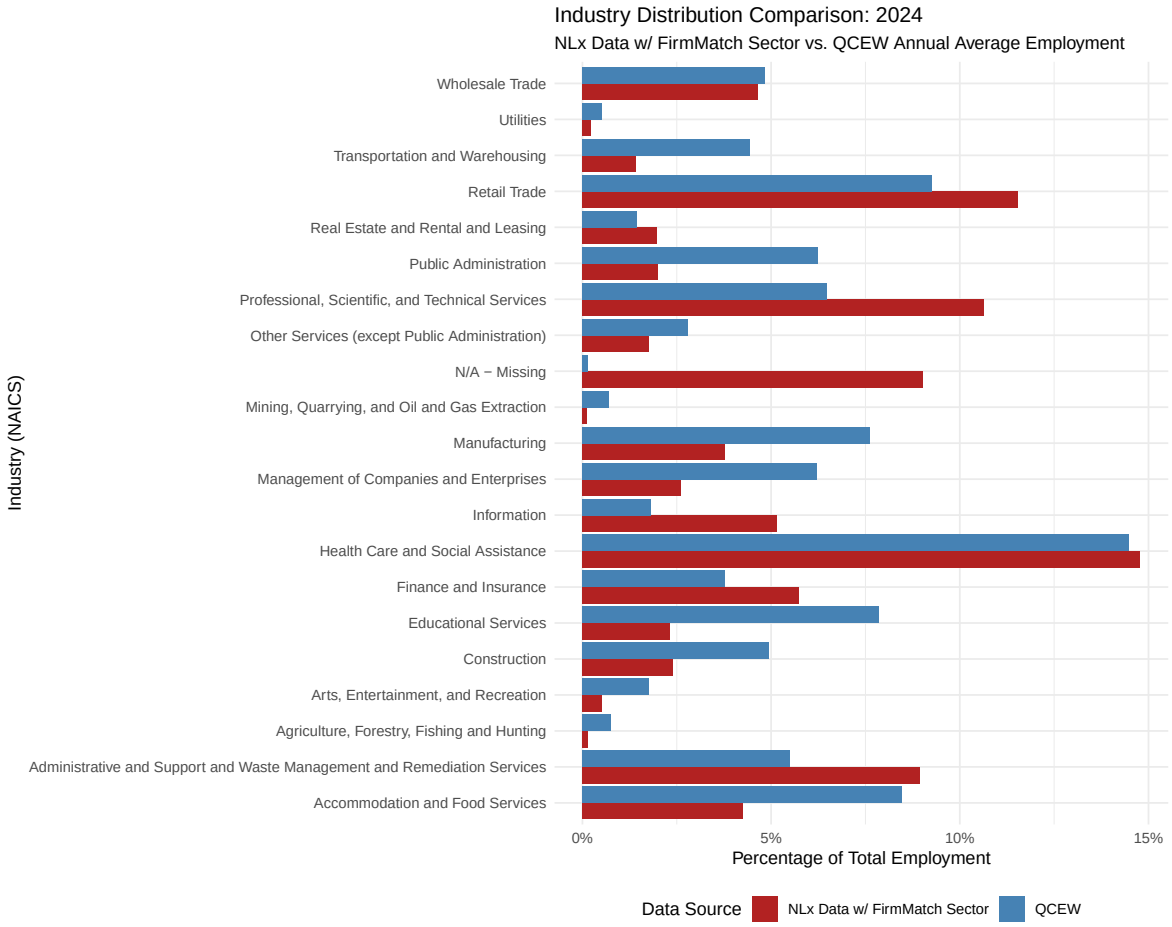



Figure 10: Industry Distribution Comparison, 2024.

### 3.1.5 WageExtract

WageExtract retrieves pay frequency, minimum, and maximum wages from the unstructured text of job ads. We developed WageExtract by randomly sampling job ad sentences that contain a list of regular expressions plausibly related to wages. We developed regular expressions to extract wages from this text and manually audited and corrected each scenario present in the training data. We then constructed a training dataset that distinguishes between sentences containing wage information, and those that do not. Using this, we fine-tune a lightweight BERT-TINY binary classification model, which quickly and efficiently identifies sentences with potential wage information. This model achieves a 96.8% F1 score on the validation set.

We then fine-tuned a DEBERTA-V3-BASE model for sequence classification in order to extract the

spans of text containing the wage statements from the identified wage sentences. In particular, we use custom tags to delineate whether an identified span refers to a lower range wage value (MIN) or upper range wage value (MAX). The resulting model achieves an F1 of 99.8 on the validation set, which measures the accuracy of predicting the correct *spans* containing wage information. Given the model outputs, we design a simple parsing algorithm to separate the extracted spans into distinct MIN and MAX return values. In addition to the nominal wage values, we also train a multi-class DeBERTA-v3-base classification model to extract the pay frequency expressed by the wage values. These labels for the training dataset were obtained likewise via crafted regular expressions, and manually checked for validity. The training dataset represented a subset of 22k examples, and the resulting model predicts pay frequency with a F1 of 99.6.

We combine the results of WageExtract with NLx's structured wage information and post-process results to remove outliers and standardize the wage as an annual salary, using either the point wage provided or the midpoint of the wage range provided. For the duration studied, NLx structured data includes a minimum or maximum wage for 4.62% and 4.15% of job postings. With WageExtract, we obtain wage information for far more observations. Figure 11 illustrates that the availability of wage information in our data hovers between 10% and 13% before 2022, and dramatically increases beginning in 2022. In our dataset, the percent of job ads with wage data in the text reaches 39.6% in May 2025.

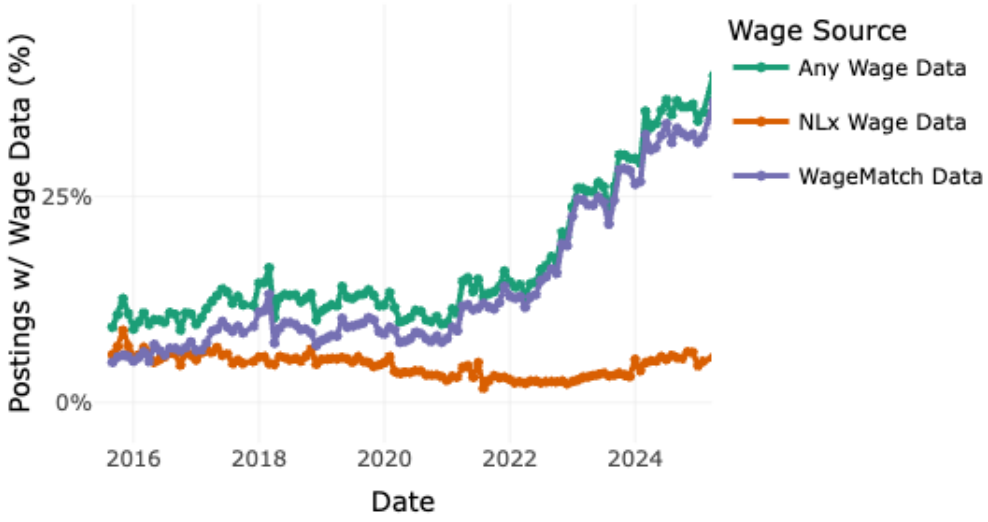



Figure 11: WageExtract: The percent of postings with wage information.

For comparison purposes, an analysis of structured data provided by Lightcast, Batra, Michaud and Mongey (2023) report that 14% of job ads had any wage information between 2012 and 2017, and 8% had point data. Using data from Lightcast, Hazell et al. (2022) state that 5% of job ads include point wage data from 2010-2019.

Figure 12 demonstrates that availability of wage information varies significantly by state, and increases significnatly following the passage of pay transparency laws. Colorado enacted the first state-wide pay transparency law in 2019, and took effect in January 2021. Other states have followed: Washington's law took effect January 2023; New York in September 2023; Illinois' law took effect January 2025. With each law, the percent of job ads including a wage that we extract increases significantly following passage. Arnold, Quach and Taska (2022) reports that availability of pay information increases by 30 percentage points following the implementation of Colorado's law and reaches 70% by the year following the law taking effect. We find similar results in our dataset, for Colorado and other states.

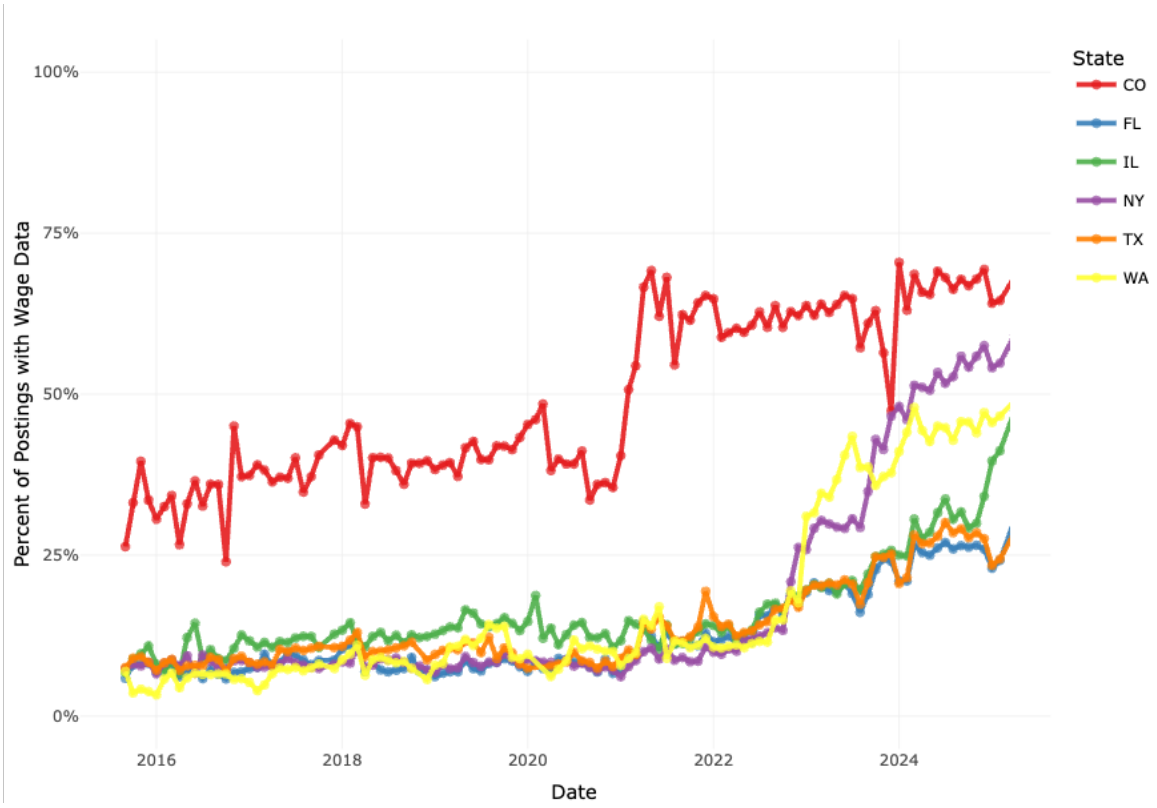



Figure 12: WageExtract: Wage Information Availability in States with Pay Transparency laws, Florida, and Texas.

Figures 13 and 14 display boxplots comparing OES data with wage results from WageExtract and occupation coding from TitleMatch. Batra, Michaud and Mongey (2023) caution strongly against using wage information from job postings as a proxy for administrative wage data. While we repeat their finding that job ads consistently have wages that are "lower in high wage occupations and higher in low wage occupations relative to the OES", further research is needed.

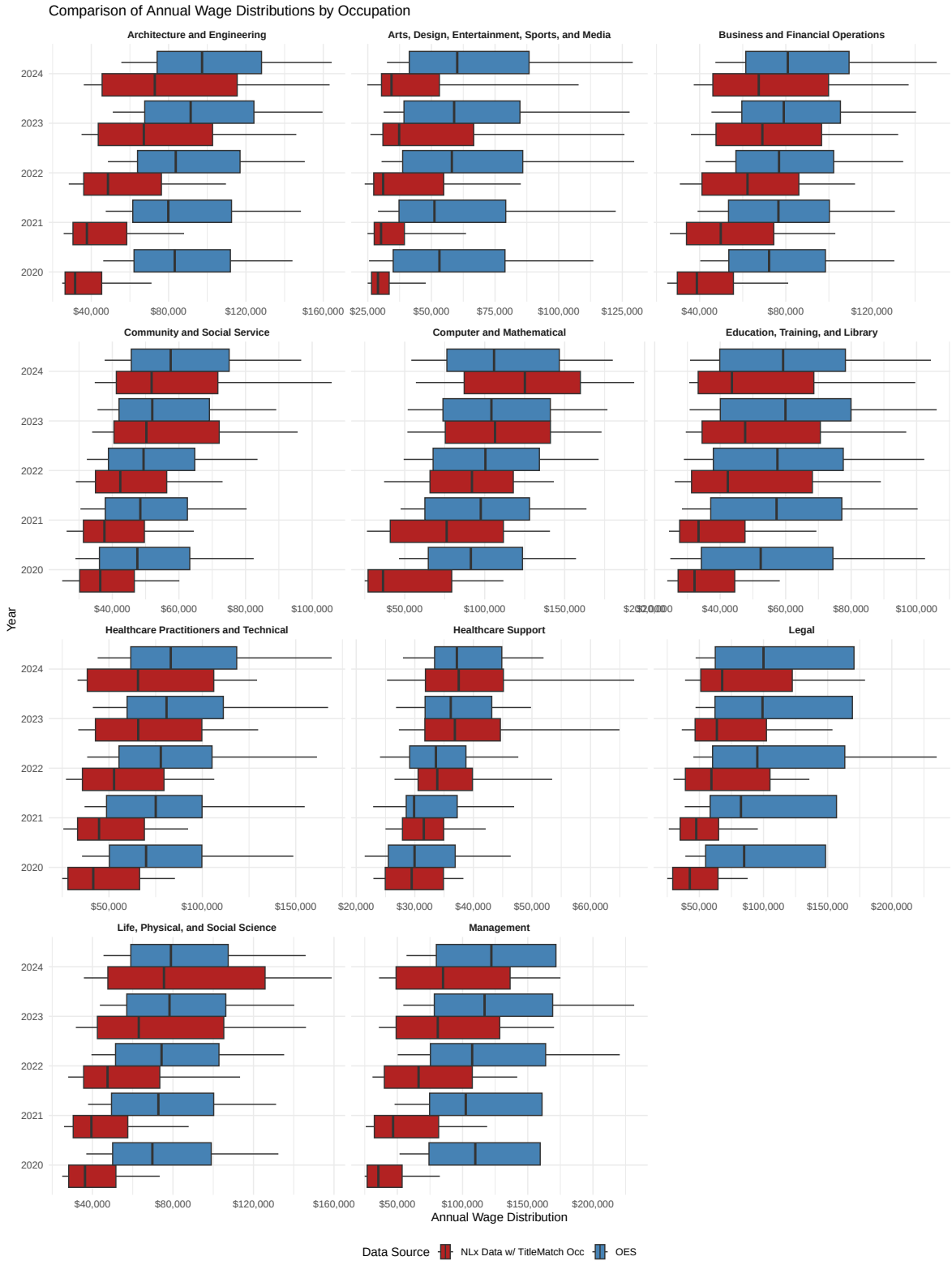



Figure 13: Wage Distributions, 2020-2024, for SOC codes below 32.

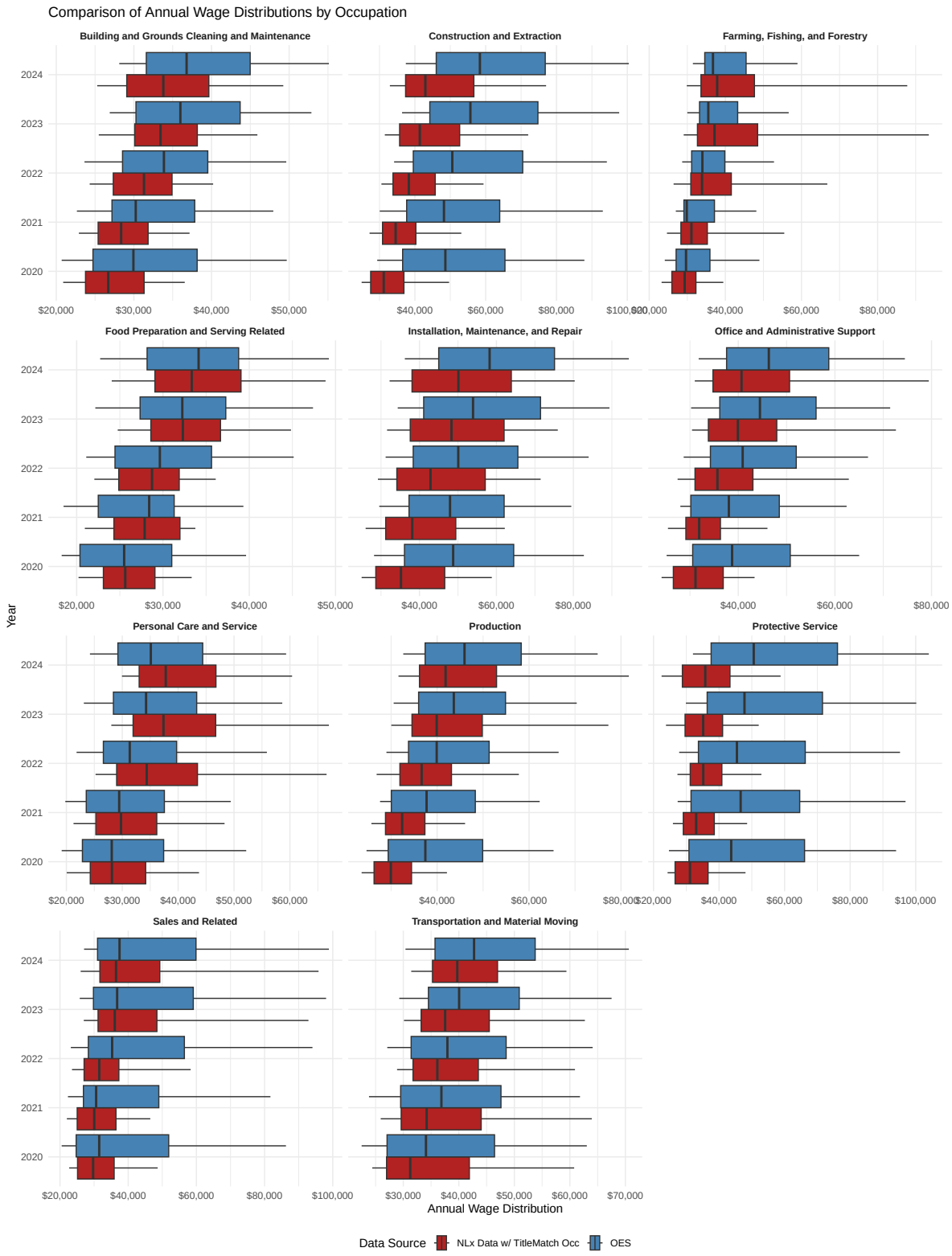



Figure 14: Wage Distributions, 2020-2024, for SOC codes above 32.

### 3.1.6 JobTag (CRAML)

The JobTag module of the Job Ad Analysis Toolkit (JAAT) classifies job ad text into user-defined categories using niche classifiers built with the Context Rule Assisted Machine Learning (CRAML) tool (Meisenbacher and Norlander, 2023). In particular, the nine classifiers are Random Forest classifiers, train on expert validated "context windows" relevant to each niche class. For example, the class 'ind_contractor' loads a classifier that indicates whether a job ad contains language indicating the job is on a contract basis, 1099, or otherwise indicated for 'independent contractors.' Ten pre-defined classifiers are currently included in the JobTag module. JobTag directly illustrates the merits of CRAML's domain-specific classifiers that are fine-tuned on expert-curated context rules.

The JobTag module is extensible in that it can support any number of newly added classifiers, accomplished via the definition of a new class and its keywords, and then via the validation of extracted context windows based on these keywords. In this way, should other researchers or practitioners develop and publish niche classifiers, this module allows for coverage of novel, emergent, and specialized interest in data extraction from job ads.

Figure 15 illustrates change in the percent independent contractor over time in different major sectors. Appendix B contains additional illustrations by different groupings of occupation and industry for a limited subset of the JobTag variables, features extracted from titles by TitleMatch, and dictionaries described below. The online data navigator provides additional opportunities to explore change in the prevalence of these binary indicator variables over time, occupation, industry, and geography.

## 3.2 Dictionaries

We exact match terms using pre-existing and novel dictionaries that map to O*NET's taxonomy of work. As described in Appendix A, we map O*NET's structure to additional dictionaries we adopt or build. Custom dictionaries we develop are presented in Appendix E Tables E.1 - E.4, and include dictionaries for job title features (used to build training data for TitleMatch), benefits, education, shifts, and drug, background and criminal background checks. Standard dictionaries from the management literature are described in this section and presented in more detail in Appendix E.6.

To execute dictionary-based strategies, we use patent-pending analytic engines that scale large 'knowledge maps' with unique concept identifiers and association rules over unstructured text with exact matching (Price, Boyda and Bobay, 2024). 'Knowledge maps' find and match one or more keywords to a standard label or code at high speed. Capable of addressing negation and complex association rules such as the presence of multiple unique concept identifiers within a specified span, lists of terms, such as the O*NET dictionary of 21,841 tools and technologies, are run against the corpus and return UNSPSC codes associated with the presence or absence of the dictionary term(s)

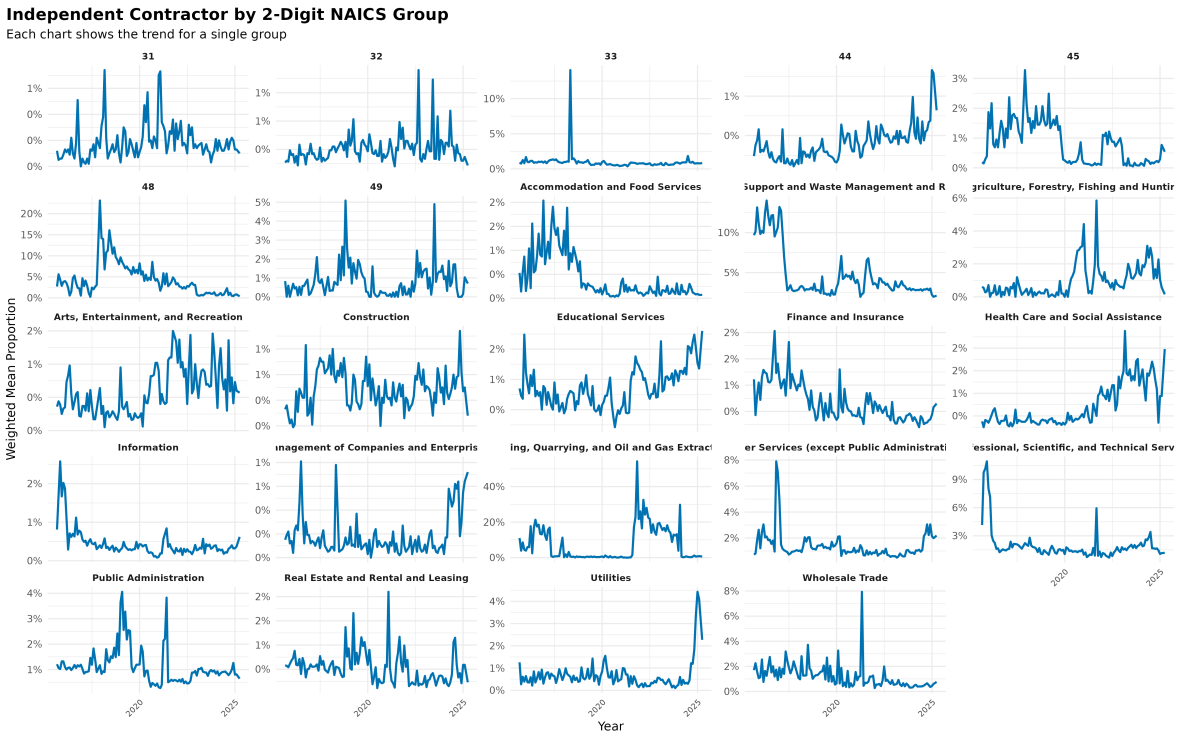



Figure 15: Independent Contractor by major NAICS sector.

within each job ad.

We visualize results for several of the O*NET dictionaries in an abbreviated fashion here to provide suggestive description of how this may be of use to other researchers. General occupational interests used in vocational interests and career planning and based on Holland's (1997) RIASEC (Realistic, Investigative, Artistic, Social, Enterprising, and Conventional) framework, are captured using a dictionary of RIASEC keywords developed by O*NET (Rounds, Putka and Lewis, 2022). Figure 16 illustrates that as a share of the total RIASEC keywords extracted, there has been an approximately 1% decline in the share of enterprising and social keywords, and a 1% increase in investigative keywords between 2015 and 2025.

As described in Appendix A, O*NET's taxonomy incorporates all important elements of work, but does not sufficiently elaborate them. To illuminate the organizational context of work within the O*NET framework, we developed several novel dictionaries and deployed standard dictionaries from the business and management literature. For example, several novel dictionaries indicate various aspects of scheduling predictability and flexibility of the job, and other novel dictionaries include a dictionary of diversity, equity, and inclusion terms based on a subset of the list of terms recently reported to the subject of scrutiny by the federal government (Yourish et al., 2025).

Figure 17 illustrates several indicators of job flexibility and predictable schedules for three large occupations: Home Health Aides, Nurses, and Retail Salespersons. Specific shift includes phrases

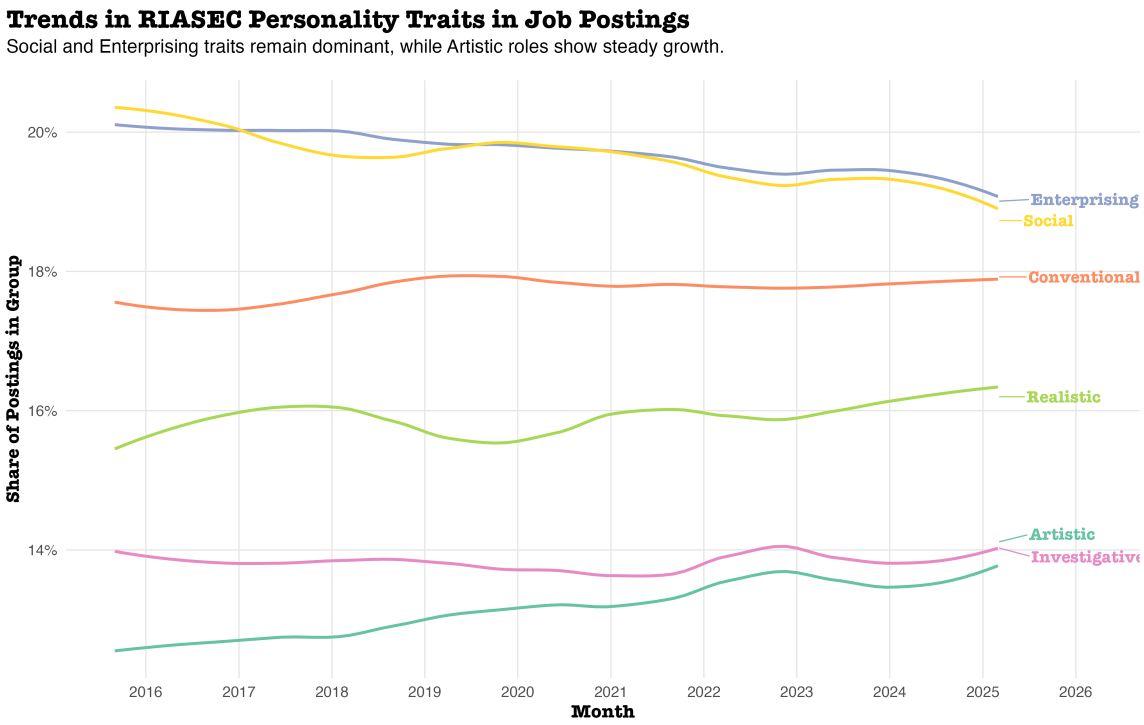



Figure 16: RIASEC Keywords as a Percent of All RIASEC keywords.

associated with a specific, predictable shift; flexible schedule indicates all kinds of unpredictable and flexible schedules, including those that indicate a willingness to accommodate workers' preferences; flexible for employer indicates a desire for workers who can work hours that the employer prefers. The results suggest a rise in flexible schedules and predictable shifts in the last decade across these three large occupations. While registered nurses and home health aides generally have a low percent of postings with expectations that the worker be flexible for employer needs compared to retail sales, there was a significant increase in expectations for flexibility around employer needs for home health aides during the 2020-2022 post-pandemic time period.

Standard dictionaries presented in the Appendix E.6 including recruiting signals (Banks et al., 2018); time horizon (Brigham et al., 2014; Brochet, Loumioti and Serafeim, 2015; DesJardine and Bansal, 2019); innovation (Heyden, Sidhu and Volberda, 2015); digital orientation (Kindermann et al., 2021); creativity (Haselhuhn, Wong and Ormiston, 2022); attention (Levy, 2005); ambidexterity (McKenny et al., 2018*a*); entrepreneurial orientation (McKenny et al., 2018*b*; Short et al., 2009); Value Orientation (Moss et al., 2018); Corporate Social Responsibility (CSR) (Pencle and Mălăescu, 2016); Sustainability Orientation (Vaupel et al., 2023); Market Orientation (Zachary et al., 2011).

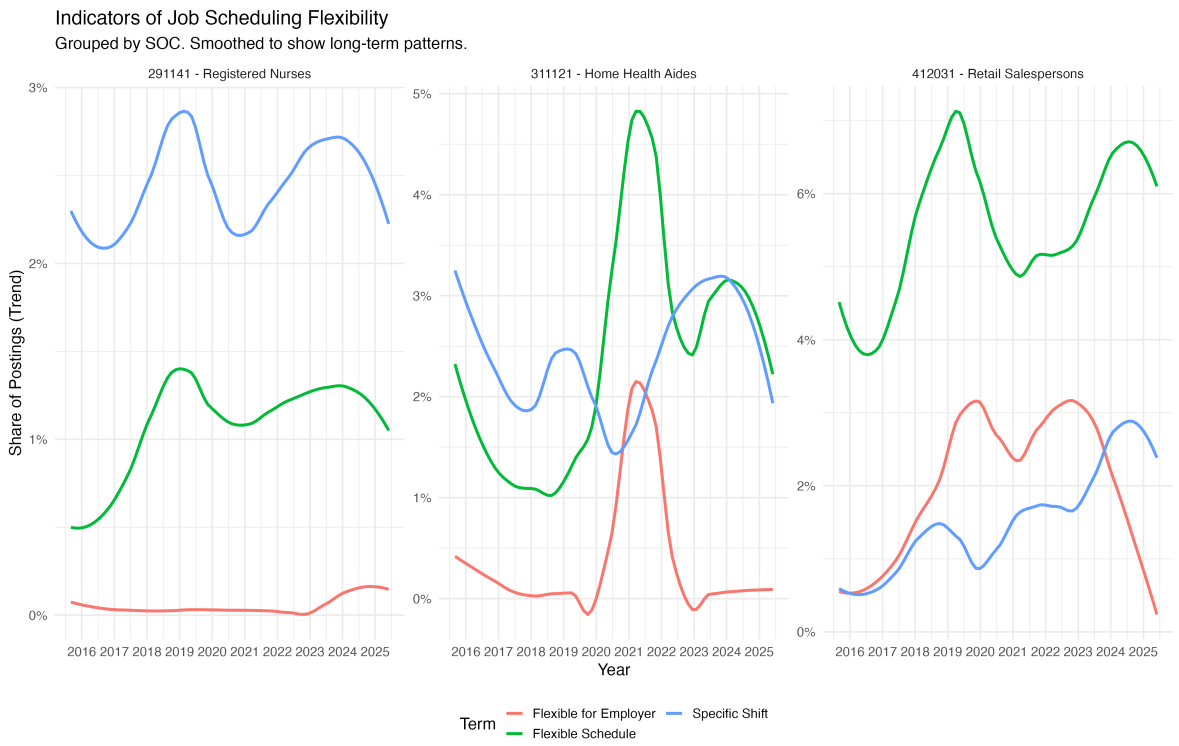



Figure 17: Indicators of Job Flexibility.

## 3.3 Aggregations: Month, Occupation, Industry, Geography

We aggregate data at month, occupation, industry, and geographic levels in order to build data that is usable for research and practitioner purposes. Occupation and industry aggregations at the 2-digit, 4-digit, and 6-digit level are performed with the output of TitleMatch and FirmMatch, respectively, as described above. We create sums, means, and percentile variables to reflect the underlying data within a "month" that we create as described below.

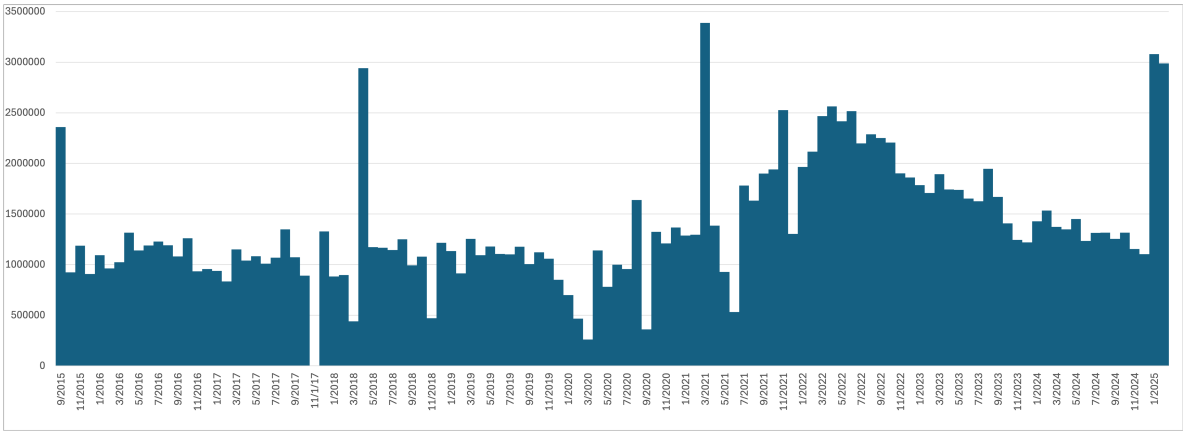
### 3.3.1 Building a Time Series: Month

For months after September 2015, NLx provides monthly 'job' files with one row per unique job ad, grouped by the month of a job ad's last *date_compiled*. The NLx field *date_compiled* provides "the last date [a job ad] appeared in the job feed." As displayed in Figure 18, monthly job files begin September 2015. One month – November 2017 – is missing. Several months contain an unusually large numbers of job ads.

NLx has improved systems for collecting and storing job ad data over time. A major data warehouse upgrade in 2021 added comprehensive job history tables that track more precise windows of dates when job ads were posted. For periods prior to 2015, the NLx API provides access to additional job postings.

As the monthly file contains only the job postings in the month in which they are last posted,

Figure 18: Count of Monthly Job Ads by Date Compiled




its contents include many postings that were visible in earlier monthly periods. Analyzing NLx data, Hashizume (2024) finds half of job postings from Fortune 500 firms are available for 37 days or less, and that there is a long tail of "evergreen jobs" posted for long durations: 25% of job postings are up for more than 90 days, and 10% for more than 180 days. Employers often use evergreen postings to fill multiple vacancies for stable roles with steady demand over long durations – these assist employers seeking to meet just-in-time hiring needs. At the extreme end, 20 job postings have been posted for the entire 16 years the NLx has data, dating back to 2007, leading some to call these "phantom" or "ghost" jobs. As Hashizume (2024) describes, both states and employers address vacancy duration differently: while most states do not impose a time limit, several states cull their records after 30/60/90 day periods, and many employers set initial 30, 60, or 90 days as initial windows for postings, and then may extend these. Hashizume (2024) finds that employer differences seem to drive the variation in job posting duration.

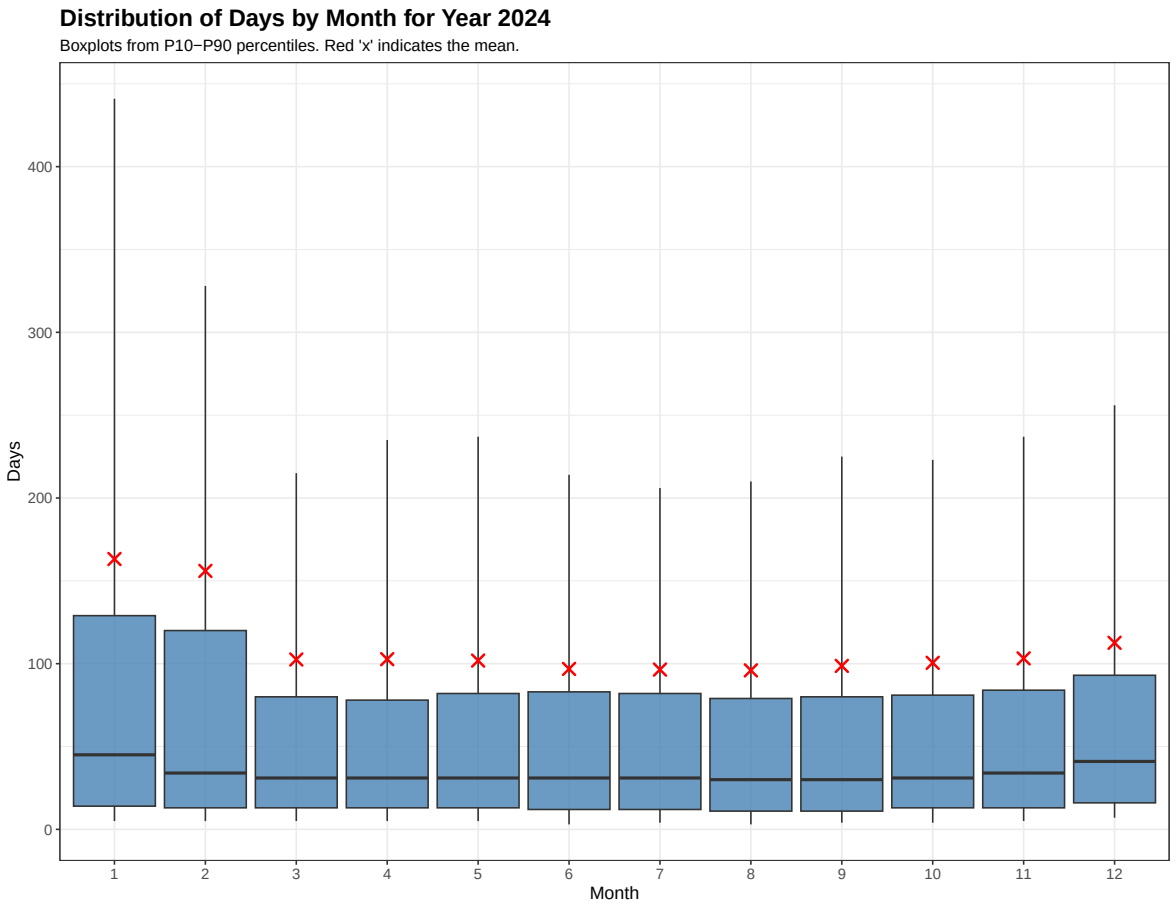
As each monthly file can be subject to large fluctuations and includes only the postings that were last posted in that month, we investigate techniques to reflect that many of these postings were on display in earlier months, smooth the data appropriately, and reduce the potential for noise in a given monthly 'jobs' file to drive results.

### 3.3.2 Data smoothing

Researchers may wish to smooth data. We analyze the job history table provided by NLx and compare NLx aggregate counts to estimates of job openings provided in the Job Openings and Labor Turnover Survey (JOLTS) (Bureau of Labor Statistics, U.S. Department of Labor, 2025). Research on job vacancies provide a variety of findings and techniques to address online job posting durations. An analysis of JOLTS data from 2001-2009 estimates a mean vacancy length of 14-25 days, and that each

job opening yields between 1 - 1.8 hires (Davis, Faberman and Haltiwanger, 2013). With LinkUp data, Chen and Li (2023) drop job ads that are posted for more than 180 days and find an average posting duration of 36.5 days and median of 23 days. Mueller et al. (2024) find a mean posting duration in Austrian job postings of 30.5 days. Using UK job postings data, Bassier, Manning and Petrongolo (2025) find a mean posting duration in UK job postings of 17-18 days.

Figure 19: Distribution of NLx Job Posting Duration in 2024




We first analyze the monthly 'job history table' available since June 2021 to understand patterns of job duration in NLx data. We find the maximum duration for each job posting appearing in each monthly job history file. This answers the question: how long have job postings active in a particular month been posted? Figure 19 provides boxplots of durations of job postings in the monthly job history files in calendar year 2023. The median ranges between 29 and 47 days, the mean between 125 and 158 days, and the 75th percentile of durations is between 80 and 120 days. It appears that more than half of job postings compiled in a typical month will also appear in the prior month, and more than 25% will appear 3 or more months prior to the month in which they are compiled.

Based on the above analysis, we sought to assess the tightness of fit between different approaches

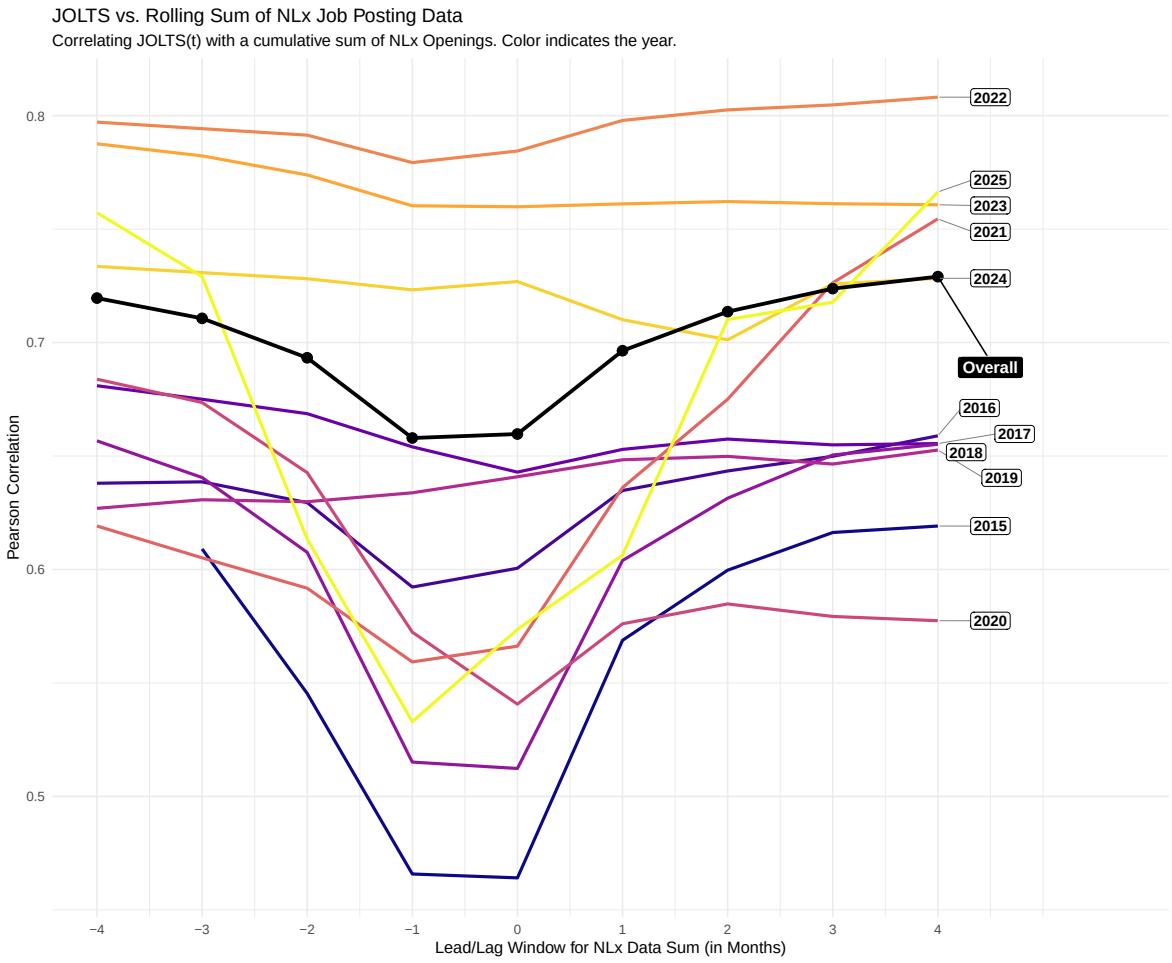
to aggregating the NLx postings data and JOLTS data. We aggregate the national monthly count of NLx job postings from the monthly jobs file. For testing purposes, we create the monthly count of postings from a single monthly file (at time 0) and using the monthly file of one or more of the preceding or subsequent monthly files within a 4-month range. This 'rolling sum' approach allows us to assess whether adding leading or lagging months to a focal month would improve the fit between our dataset and JOLTS data.

Figure 20 illustrates the correlation between alternative rolling aggregations of the monthly count and JOLTS monthly job openings data. The overall and yearly correlations illustrate several descriptive findings: 1) adding one lagging month typically decreases the correlation, 2) the correlation has increased in recent years compared to earlier, and 3) there is a stronger correlation when adding 3 or 4 months of leading data following the publication of the monthly file. The overall correlation is 0.7 when adding three months of data, including for all years since 2020 (a year where both job postings and official statistics collections may have been especially affected by pandemic related challenges). Because gains after three months appear to be small, and based on the above analysis, researchers may choose to add leading data to each focal month's dataset for analysis purposes.

#### 3.3.3 Geocoding: State and Zip Code to Commuting Zone

For geographic analysis, we standardize the state field from NLx metadata, and use the zip code from the NLx metadata to construct 1990 commuting zone level data. We link zip codes from job ads to county FIPS codes provided by Din and Wilson (2020). If a zip code includes more than one county FIPS, we assign the zip code to the county with the greater proportion of businesses. We link from FIPS codes to 1990 commuting zones / local labor market areas (Tolbert and Sizer, 1996).

Figure 20: The Correlation Between JOLTS and NLx Postings



# 4 Results

Our results are capable of providing both telescopic and wide-angle evidence of how labor market demand changes, with an unprecedented number of features of work within and across occupations over time, and by industry, and geography. We highlight several results for several key audiences: 1) implications for researchers with illustrations of the potential to measure labor market change and shocks in granular detail and unprecedented large-scale description of national trends in skills, tasks, and management practices; 2) implications for workforce development and education planning professionals with an illustrative sample report on an occupation. Each result presented here requires more in-depth research. Users of this information should proceed with an awareness of the known limitations of this new data, which are described more fully in the concluding section.

## 4.1 Applications for Research

Aggregate data released and tools in the JAAT toolkit are available for research purposes. Trained on job advertisement text and custom built, the JAAT toolkit can be especially helpful for researchers examining and comparing the extracted and structured information in job ads and other sources of information. While the data released contain a great deal of information that may be useful for specific studies, JAAT tools may be of assistance when working with raw NLx or proprietary job ad datasets for specific study designs. Here, we demonstrate how the aggregate data can be of use in the study of how technological and other shocks affect the organization of work and jobs as an aid to exploration and discovery as well as an input to future work focused on causal inference.

### 4.1.1 A Wide-Angle Lens on National Trends

In a turbulent ten year period of economic change including a global pandemic, hot labor markets, supply-chain disruptions, and technological shocks, trace evidence of trends that gradually transform the labor market, and shocks that quickly change employer demand in labor market should be evident in the 2015-2025 time frame. Many research agendas that depend upon access to the scale and real-time data to understand these changes and shocks can be furthered by these data.

Tight pre-pandemic and post-pandemic labor markets (referred to as the "great resignation") occurred incrementally and are likely to effect more than the overall demand for labor for the duration of a period with extremely low unemployment. The release of ChatGPT on November 30, 2022 heralded rapid advances in generative AI, and the COVID-19 pandemic, beginning in March 2020, are examples of shocks – sudden unexpected events with far-reaching implications. Such shocks can have short and long term impacts. Over a longer period of time, generative AI is likely to transform occupations

as workers adopt technology to replace their effort on certain tasks Handa et al. (2025). COVID-19 closed many workplaces to all but "essential workers" in the short-term, with effects on the composition of available jobs, and likely had lasting impacts on the organization of work at different levels of education, in specific sectors, and occupations.

Event studies and plausibly exogenous technological and policy shocks are rare in research with job postings. Horton and Tambe (2025) demonstrate that it is possible to trace the extinction of technologies through job postings, and Sauerwald and Norlander (2024) trace the impact of board members' prior political experience on recruitment policies after a political upset in 2017. However, no prior public research-use dataset has been able to follow change at a highly-granular level, including change within occupation at the task-level, in combination with thousands of features such as technologies, tools, using the standard taxonomies for skill and other features.

To illustrate significant national trends and shocks in labor market demand demand by employers, we demonstrate: 1) meaningful shocks that alter demand for specific task bundles in the last ten years; 2) changes in the measured linguistic complexity of job postings over the 10 year period; and 3) national trends involving the gradual rise of project-based work, which is detected at the task level, especially after the pandemic, and the rise of interpersonal work activities as a share of postings; and 3)

Figure 21 presents a time series of tasks related to cleaning, inventory and purchasing, recruiting, and scheduling from the 2015-2025. This figure presents both both count and percent (as a percent of job postings in a month) to illustrate that the composition of job postings in the labor market and number of jobs involving a task being offered are different but both relevant. Moving clockwise from the top left quadrant of Figure 21, we first illustrate change in cleaning tasks (Task IDs 23536, 20779, 9515, 20790, 23102, 23557). A national increase in the count of cleaning-related tasks within job postings is evident after 2020 and the pandemic. However, the proportion of jobs including cleaning tasks does not increase until 2023. This may be due to changes in the composition of labor demand during the period of pandemic restrictions: while there was increased demand for cleaning among jobs that were open, the overall labor market demand may have been biased against in-person jobs that require cleaning. Pandemic-induced shifts in consumer demand and international supply chains disrupted inventories following the onset of COVID-19.

As a percent and a count, tasks related to inventory and purchasing increased dramatically in months immediately before and after January 2021 (Task IDs 15704, 1333, 72). The bottom right panel illustrates a dramatic increase in scheduling related tasks (Task IDs 23695, 18658, 1106) beginning in 2021, a period referred to as "the great resignation" and one in which employers made efforts to retain workers. Flexible scheduling and stable work hours are a top worker amenity, especially for women and

other groups, with significant implications for labor market participation (Bell, 2020). The bottom left quadrant displays tasks involving recruitment and hiring (Task IDs 22954, 23170, 17611, 9711, 18858, 21313). Recruitment peaks pre-pandemic and post-pandemic are interrupted by a plunge in recruitment related tasks during the peak pandemic period.

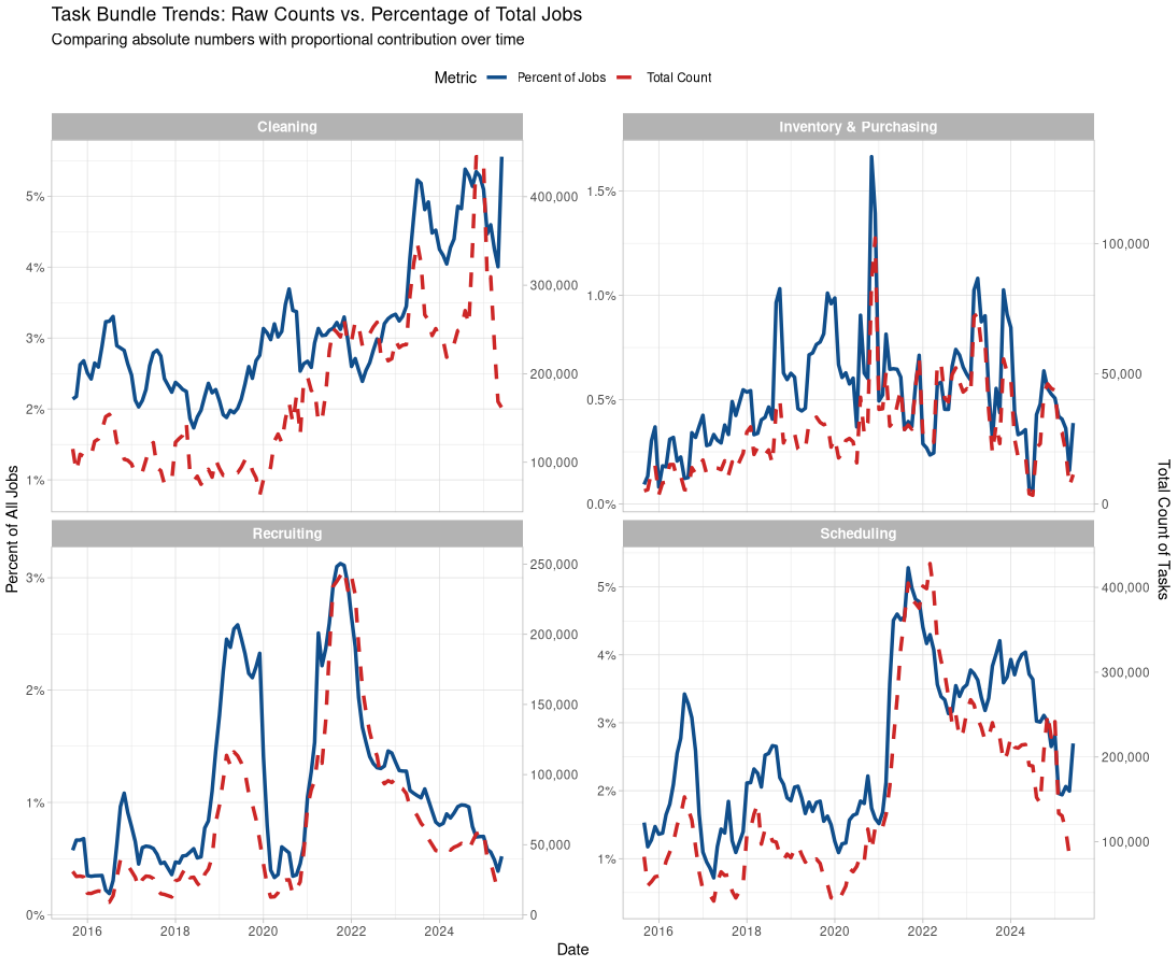



Figure 21: National trends in task bundles related to cleaning, inventory and purchasing, recruiting and scheduling, since 2015.

We construct the Fleisch-Kincaid readability score for all job ads and provide the average by occupation and other groups. Figure 22 demonstrates a stark decline in the readability of job ads in the months following the COVID-19 pandemic, likely due to changes in the composition of jobs posted in that time period and a shift in labor market demand toward jobs requiring higher levels of education. In the months after December 2022, there is a sharp decrease in the linguistic complexity of job postings nationally, which may suggest the adoption of large language model tools that can aid recruiters in writing the postings, and job seekers in accessing jobs through improved readability.

Examining the 10 most common task statements with a decade of data, we note a significant growth in the share held by Task ID 21462 "Assign duties or responsibilities to project personnel."

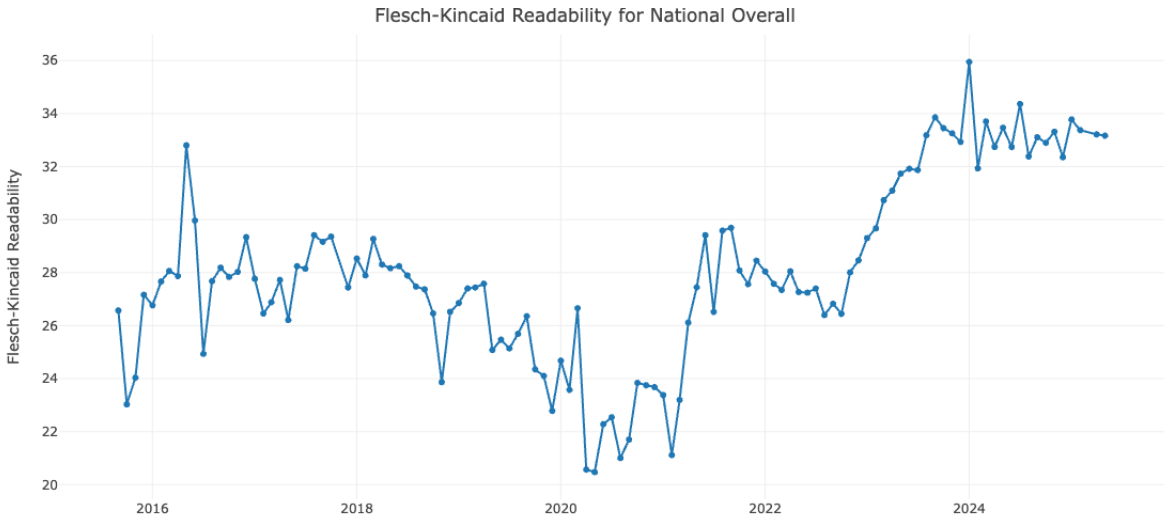
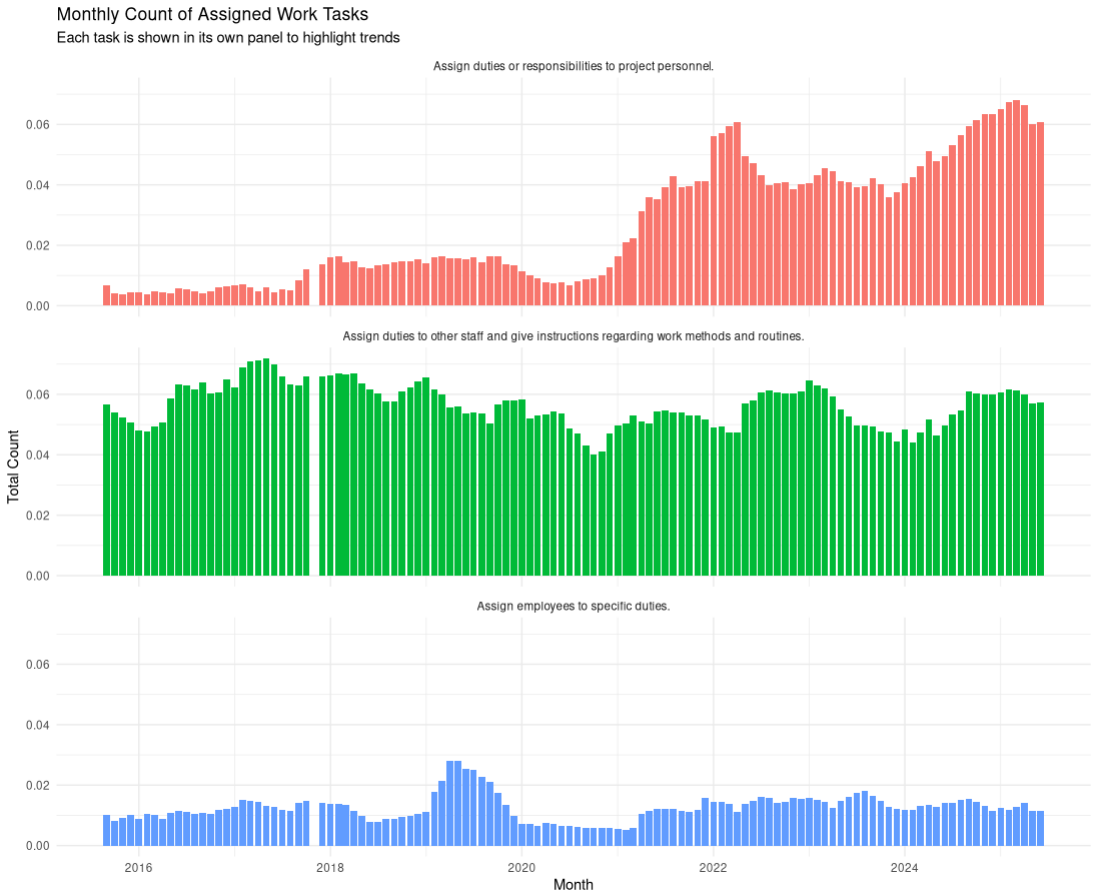



Figure 22: National Changes in the Readability of Job Ads

This growth has occurred in conjunction with an overall stability in Task IDs related to assigning work to employees or staff, e.g., Task IDs 659, "Assign employees to specific duties.", and Task ID 9583 "Assign duties to other staff and give instructions regarding work methods and routines." Figure 23 Panel A (top) illustrates that, for all occupations, the project related Task ID is 0.6% of Task IDs in 2015 versus 6.0% in 2025, while assignments of work to staff or employees is relatively flat over the time period, moving from 5.1% to 5.6%. The subtle difference that the model picks up between unique Task IDs related to projects versus staff or employees merits deeper investigation, and would be washed away at even the next level of aggregation, the detailed work activity, where these Task IDs fall under DWA 4.A.4.b.4.I13.D06 with the label "Assign duties or work schedules to employees."

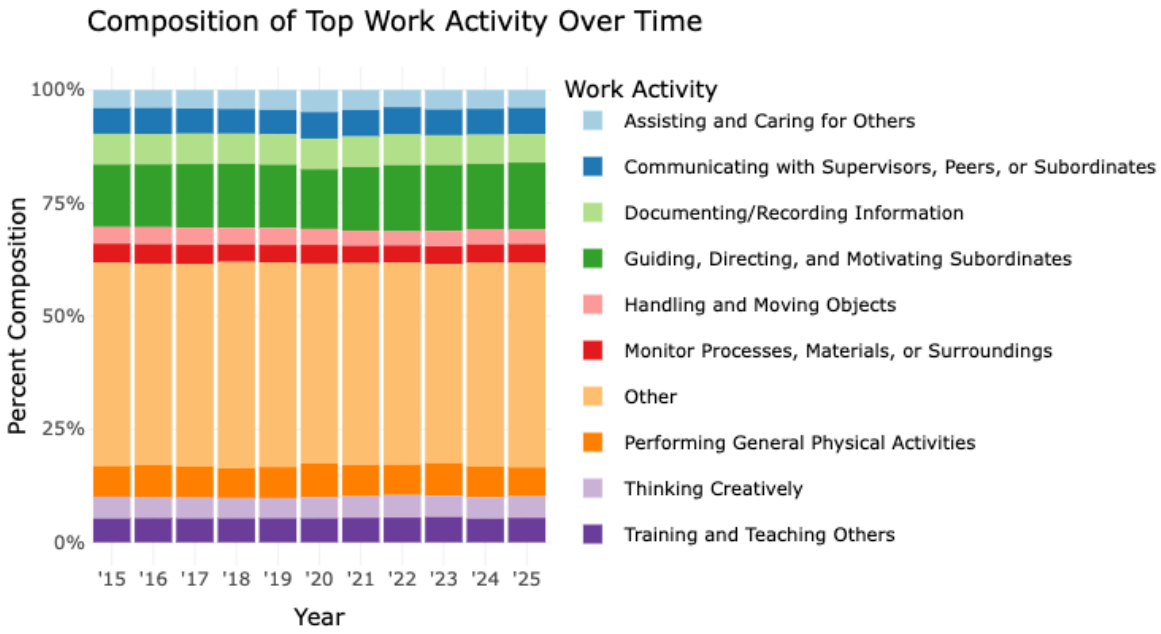
The last major trend we note is the rise of interpersonal work activities, including managing people, communication, and providing consultation. This change is visible at the work activities level. As seen in Figure 24, for all occupations, the percent of jobs involving supervision of people increases from 13.8% to 14.8% from 2015-2025. Again, the change can be seen more dramatically within detailed occupations. In the software developer occupation, managing people grows from 13.7% to 17.5% of work activities over the duration.




*Note:* The growth of project-based work is visible nationally as a percent of all job postings.

Figure 23: TaskMatch: The growth of project-based work

Figure 24: TaskMatch at the Activity Level: Increased Emphasis on People Management



### 4.1.2 A Telescope for Finer-Grained Analysis

Our dataset is capable of tracking change within detailed occupations, as suggested above in the rise of health-related tasks and scheduling for cooks and food preparation workers post-pandemic in Section 3.1.3 and Figure 8. The opportunity to track fine-grained changes provides additional means to assess labor market change. Taking one example from the prior section, the shift toward project-based work is particularly dramatic in specific occupations. Figure 25 illustrates that, for Software Developers, assigning duties to project personnel did not appear as a top 10 task in 2015, and grew to 43.8% of all tasks for the occupation by 2025. Further investigations by occupation and industry, especially when paired with subject matter knowledge and expertise, could suggest whether this is a shift in the organization of work toward flatter, specific purpose project-based teams, and/or fissuring of employment and outsourcing, and a growth in management of vendor staff rather than employees of the direct firm.

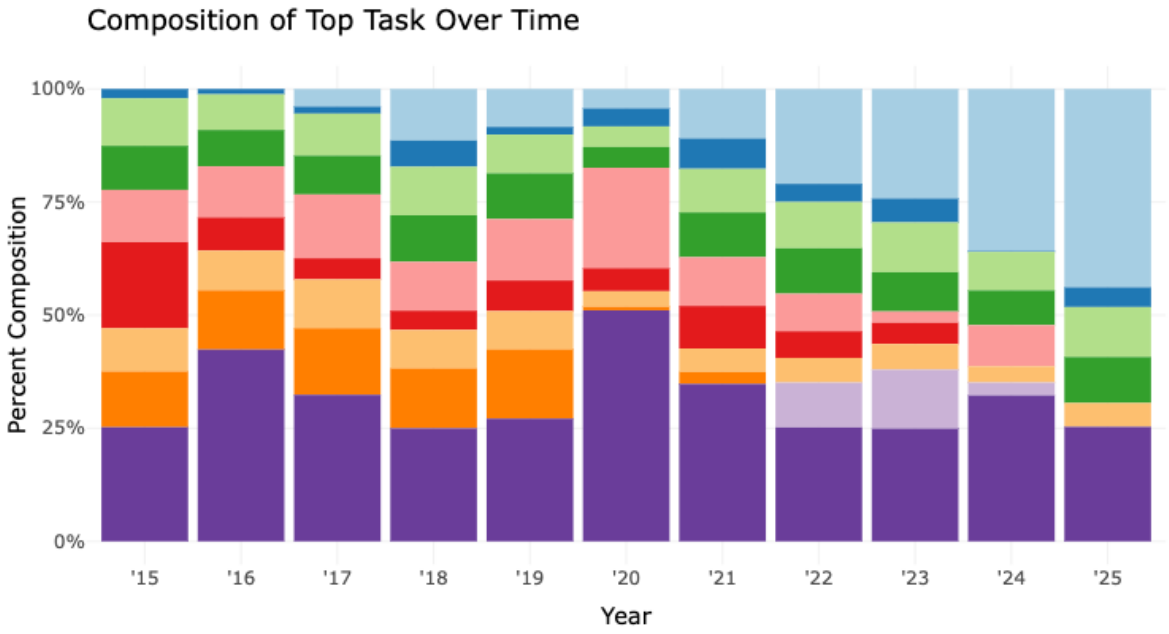



Figure 25: Growth in Project-Based Work and Interpersonal Activities in Software Developer occupation (15-1252.00). At top are the Top 10 Tasks for Software Developer. The growth of project-based tasks is depicted in light blue.

As another example from the above section, adoption of generative AI technology in recruitment posting may be less or more prevalent in specific industries and/or geographies. Figure 26 features the change in readability of job ads nationally in contrast to different sectors including information, public administration, education, and health care. Information sector job postings become more readable faster than the national trend following the launch of generative AI tools, while adoption in the education and health care sectors are slower but converging toward the national level, while public administration postings have a smaller change.

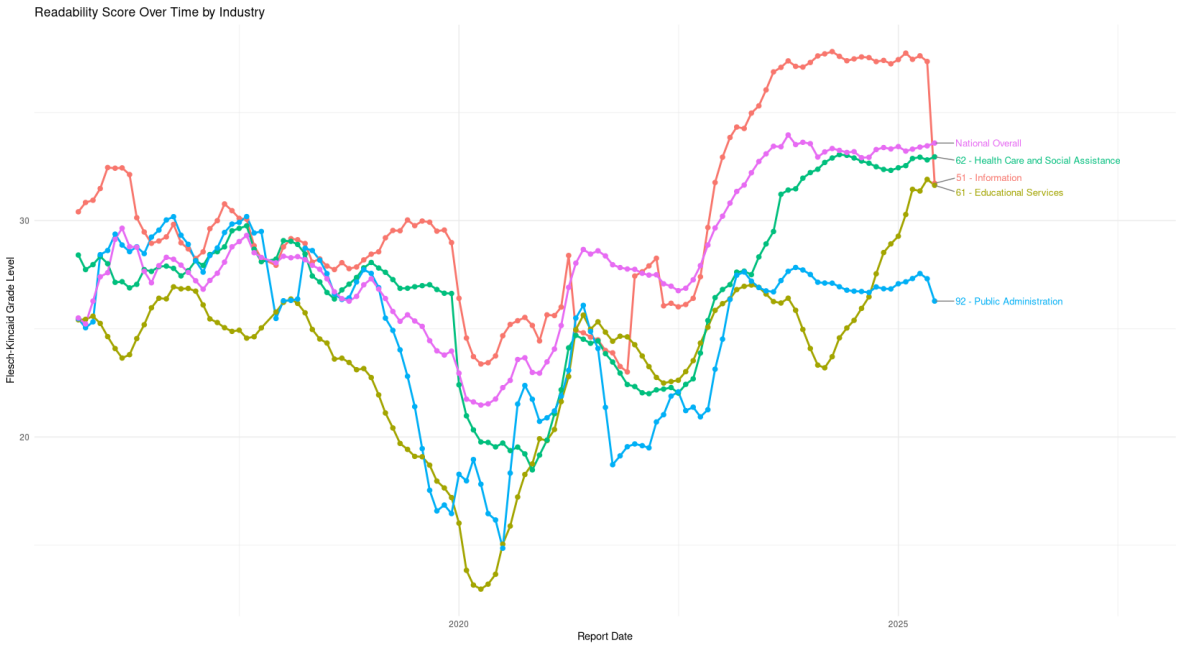



Figure 26: Information Sector Changes in the Readability of Job Ads

Along with the shock to the entire labor market with the widespread adoption of large language models (LLMs), our dataset allows us to see historical changes in key occupations that may presage broader transformations. For example, machine learning algorithms and data mining are understood to be essential to advances in computer science, AI, and LLMs. Using TaskMatch output, Figure 27 indicates that within the Computer and Information Research Scientists occupation (15-1221.00), there has been considerable growth in job ads indicating tasks that involve ML algorithms and data mining.

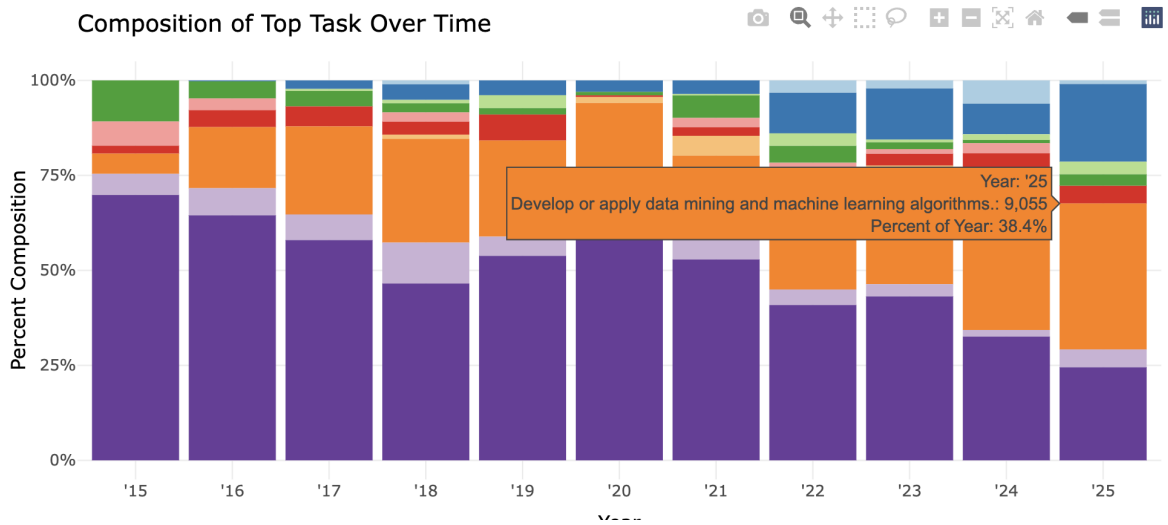



Figure 27: Transformation of the Computer and Information Research Scientist Occupation

This gradual, decade-long rise in tasks related to machine learning algorithms and data mining in

the computer and information research scientists occupation is an example of a discovery otherwise not readily available for researchers to access from current data sources. We also leverage the available list of tools and technologies from O*NET, and custom dictionaries built to track mentions of "Artificial Intelligence" and its variants (AI, e.g.), as well as a more detailed list of technical terms related to AI (e.g., "machine learning", "neural networks"), to assess changes in technology adoption. Figure 28 illustrates that as a percent of all job postings in the computer and mathematical occupations, products related to databases appear most frequently for much of the duration, with technical AI terms rising rapidly in the mathematical sciences and exceeding the percent of data analysis software more recently.

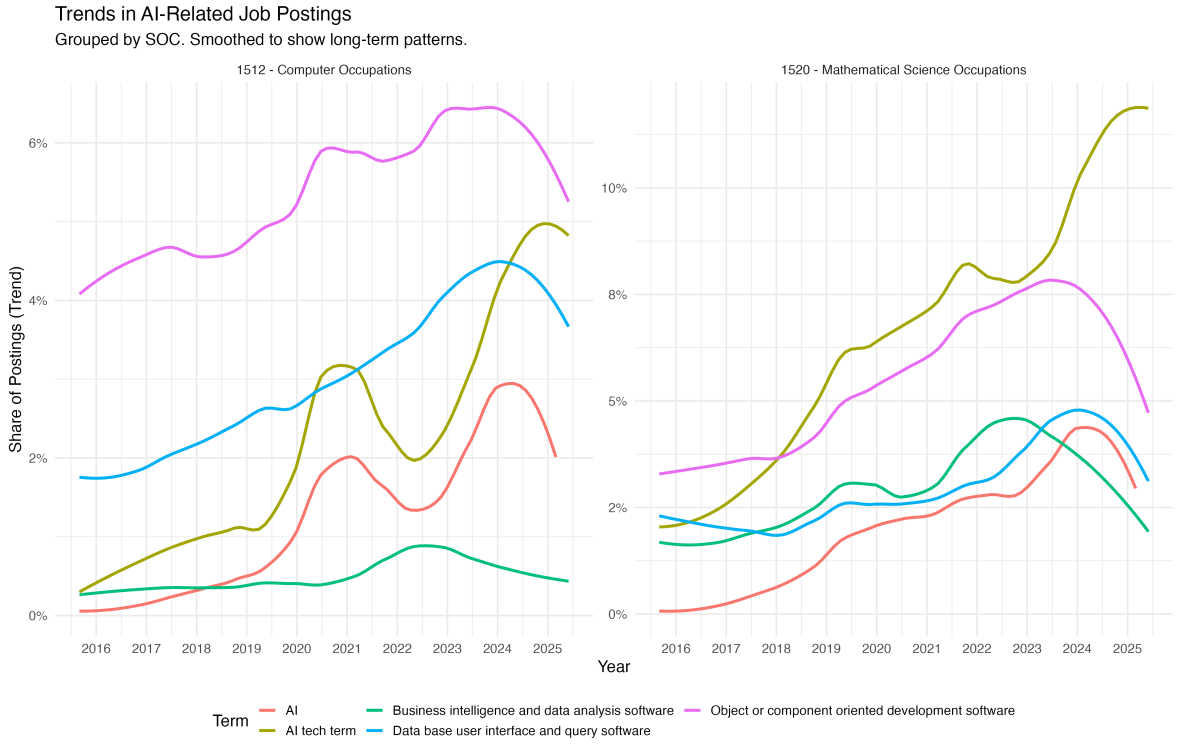



Figure 28: AI and Related Technologies in Computer and Mathematical Occupations.

Visualizing the aggregate data can serve as a vehicle for exploration suggesting areas that require additional in-depth research. Aggregate data will be made available upon publication.

## 4.2 Applications for Practitioners: A Sample Occupational Report

We provide an abbreviated and illustrative report in Appendix D for one occupation using the data exploration tool we develop to highlight how an O*NET informed by job ad data might improve the understanding of changes in labor markets for practitioners. Figure 4 illustrated a significant decline in the percent of all U.S. workers in the Office and Administrative Support occupations. To peer into possible changes within an occupation that is in decline as an overall share of the workforce, we

examine Secretaries and Administrative Assistants, an intermediate level of aggregation (inclusive of 43-6011, 43-6012, 43-6013, and 43-6014).

An examination of a decade's worth of job ads demonstrates that in the Secretaries and Administrative Assistants occupations: 1) performing administrative activities as a percent of work activities and as a skill is in decline, 2) training others is a rising work activity, 3) working with and leading others and supervising people are rising skills in this occupation, 4) taking a proactive approach is also rising, and 5) at the task level, greeting, protecting property, and placing displays were the top 3 tasks in 2015, while in 2025, the top 3 were all related to assigning duties to others.

# 5 Future Directions

This paper contributes new data and tools that point to a direction for advancing research and practice by combining O*NET and job ad data. By overcoming limitations to access created by proprietary methods and data providers, this research and the NLx Research Hub data can support further development of a vibrant and more open ecosystem for labor market information (Lane and Potok, 2024; Hirsch and Hui, 2024). While this pilot project demonstrates potentialities of NLx job ad data and O*NET, it "takes a village" to build a robust data ecosystem that share and manage data well, as information scientists have demonstrated (Borgman and Bourne, 2021). In other scientific fields, centralized repositories, dictionaries, taxonomies, and ontologies have received more funding over many more years than comparable resources such as O*NET. Further refinement and elaboration of standard taxonomies when combined with job postings data is possible, including with the data generated here and the JAAT toolkit. Models of multiple levels of government and public-private-philanthropy-research collaborations demonstrate a path forward to create data infrastructure and have both demonstrated value and capabilities to provide security around sensitive information (Cunningham et al., 2021).

The aggregate data we build can support many more detailed investigations, some suggested in the manuscript body. Especially useful will be studies that incorporate firm-level HRM policies. Existing measurement of firm-level HR and management practices is often reliant on a single respondent survey of an establishment. Research on the impact of HRM practices still faces this issue 25 years after Gerhart et al. (2000) identified the substantial measurement errors due to incumbent methods. Measurement and assessments of job quality, DEI practices, firm corporate social responsibility, and environmental, social, governance policies – often featured in job postings and one of the features of organizational context we capture – often suffer from opaque analysis, with measures obtained via obscure methods, with low convergent validity and far less coverage than is possible from job ads (Whelan and Atz, 2023; Eccles, Kastrapeli and Potter, 2017; Van Holt and Whelan, 2021). While imperfect, job ad and other textual sources of information in the public domain (or as managed by a trustee with a mission to support research) have great potential to advance knowledge and practice.

Results of post-hoc tests with LLMs and SkillMatch, and testing TitleMatch with LCA data, provides data and direction for future iterations of these models. While benchmark data does not exist, and while there are limitations with LCA data and with Career One Stop data used here to test our models out of sample, further work could yield benchmark datasets for model building and testing in the job ad classification domain. High-frequency missed matches between JAAT output and LLM judges or LCA data could be manually audited, and added in the future to the TitleMatch training data. For TitleMatch, low-frequency job-title occupation code combinations could be removed to aim for a model that returns the modal occupation code for a given title.

The Job Ad Analysis Toolkit is available now for other researchers to extract structured information from the unstructured text of job ads. Other research groups could similarly pursue data and tool contributions that improve upon measurement, including those developed here. While purpose-built for job ad text, JAAT tools may be of use in extracting information from other labor market corpora according to standardized schemes. Our aggregate data created will be released upon publication, and may be useful to developers, practitioners, and researchers as an input to further research. We plan to continue to update the data monthly until at least December 2025 under current funding and plans.

## 5.1 Limitations

Given the volume of data construction, beyond the tests and small n validation work we report in this paper, we cannot make representations about the accuracy of individual features extracted or provide a confidence level for individual subgroups in the aggregated dataset. With any ML model, there may be issues arising from inaccuracies in the training data. As models improve, and in our own work, better and different results may be obtainable from the same training data. Users of the data should assess its suitability for specific purposes.

Researcher judgment in interpreting aggregate output is required – the meaning of and the keywords associated with concepts such as CSR can change rapidly (Van Holt and Whelan, 2021). We perform no cleaning of suspect identifications in post-processing and do not hide model output, and caution users to use judgment and audit model outputs on their own data. We notice, for example, TaskMatch returns tasks related to recruitment interviews for many occupations where recruitment is likely not a task that the worker will perform on the job, but one that a candidate must undertake during the job screening process. The output of individual tools in the JAAT toolkit is subject to nuanced interpretation.

For example, the sample of reported titles can incorporate a single title into multiple occupations. Other elements from LCA data such as skills, required education, and career level, and information extracted from job ads, such as NAICS, tasks, and tools, could contribute to a more robust occupation coding model that considers more information than job title alone. This step is likely necessary for the many job titles that do not fit neatly within a single occupation. We leave in all observations, but better prediction of occupation should be possible in the future by incorporating results of the LCA analysis and additional information we extract from job ads.

While aspiring to FAIR standards (Stall et al., 2019), and providing a dataset with the ML models and dictionaries that built it, we cannot release all information. We cannot share training data due to agreements that limit the ability to share any information obtained at the row-level. Our agreement with NLx allows disclosure of information at aggregate levels only, only for the purpose of an approved

project. This means that underlying training data for models we build, which include short excerpts of real job ad text and their associated labels, cannot be shared. Pursuing additional research topics with NLx's data can be done through their project review process.

There is no standard benchmark job ad description dataset online to permit benchmark testing of model performance. This limits the ability of researchers to develop models independently using the same training data, or to compare our results against an agreed upon benchmark of labeled data. Because we use established dictionaries and taxonomies as the principal approach to extraction and do not interrogate all of the text features extracted, whether the appropriate label is used for a given feature is not addressed. As a result, while we adopt the standard dictionaries and taxonomy of O*NET, existing taxonomies are known to be non-exhaustive in their coverage of features as well (Bowker and Star, 1999), and we pursue novel creation of only a small number of JobTags or dictionaries. Results are necessarily ad hoc and another researcher might achieve different results: control over the answer to questions such as "is a hotdog a sandwich?" might appear unimportant, but many important legal and business cases revolve around similar questions (Abend, 2023, pgs. 3-27), and researchers and practitioners often have diverse use cases.

## 5.2 Conclusion

We build custom ML models and tools to create an aggregated dataset to understand change in the workplace using job ad descriptions and data from the National Labor Exchange Research Hub and O*NET's structured taxonomy as a basis feature extraction. Potential applications for researchers and practitioners are described. Intended to overcome limitations in O*NET data collection and proprietary job ad datasets, the data and models built here are suggested to have high potential for future use in research and practice. Further research is needed.

# References


**Abend, Gabriel.** 2023. *Words and Distinctions for the Common Good: Practical Reason in the Logic of Social Science.* Princeton University Press. Google-Books-ID: JieqEAAAQBAJ.

**Acemoglu, Daron, and David Autor.** 2011. "Skills, tasks and technologies: Implications for employment and earnings." In *Handbook of labor economics*. Vol. 4, 1043–1171. Elsevier.

**Adadi, Amina, and Mohammed Berrada.** 2018. "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)." *IEEE Access*, 6: 52138–52160.

**Alabdulkareem, Ahmad, Morgan R. Frank, Lijun Sun, Bedoor AlShebli, César Hidalgo, and Iyad Rahwan.** 2018. "Unpacking the polarization of workplace skills." *Science Advances*, 4(7): eaao6030. Publisher: American Association for the Advancement of Science.

**American Economics Association.** 2023. ""AEA Data Legality Policy and Explanations" and "AEA Data and Code Availability Policy"."

**Arnold, David, Simon Quach, and Bledi Taska.** 2022. *The impact of pay transparency in job postings on the labor market.* SSRN.

**Atalay, Enghin, Phai Phongthiengtham, Sebastian Sotelo, and Daniel Tannenbaum.** 2018. "New technologies and the labor market." *Journal of Monetary Economics*, 97: 48–67.

**Atalay, Enghin, Phai Phongthiengtham, Sebastian Sotelo, and Daniel Tannenbaum.** 2020. "The Evolution of Work in the United States." *American Economic Journal: Applied Economics*, 12(2): 1–34.

**Autor, David, H.** 2013. "The "task approach" to labor markets: an overview." *Journal for Labour Market Research*, 46(3): 185–199.

**Azar, José, Ioana Marinescu, and Marshall Steinbaum.** 2022. "Labor Market Concentration." *Journal of Human Resources*, 57(S): S167–S199. Publisher: University of Wisconsin Press.

**Balducchi, David, Randall Eberts, and Christopher O'Leary.** 2004. "Labor Exchange Policy in the United States." *Upjohn Press*.

**Banks, George C., Haley M. Woznyj, Ryan S. Wesslen, Katherine A. Frear, Gregory Berka, Eric D. Heggestad, and Heather L. Gordon.** 2018. "Strategic Recruitment Across Borders: An Investigation of Multinational Enterprises." *Journal of Management*, 45(2): 476–509.

**Bassier, Ihsaan, Alan Manning, and Barbara Petrongolo.** 2025. "Vacancy Duration and Wages." *The Review of Economics and Statistics*, 1–28.

**Batra, Honey, Amanda Michaud, and Simon Mongey.** 2023. "Online job posts contain very little wage information." National Bureau of Economic Research.

**Bell, Alex.** 2020. "Job Amenities and Earnings Inequality."

**Blinder, Alan S.** 2009. "How many US jobs might be offshorable?" *World Economics*, 10(2): 41–78.

**Borgman, Christine L, and Philip E Bourne.** 2021. "Why it takes a village to manage and share data." *arXiv preprint arXiv:2109.01694*.

**Bowker, Geoffrey, and Susan Leigh Star.** 1999. "Sorting things out." *Classification and its consequences*, 4.

**Brigham, Keith H., G. T. Lumpkin, G. Tyge Payne, and Miles A. Zachary.** 2014. "Researching Long-Term Orientation: A Validation Study and Recommendations for Future Research." *Family Business Review*, 27(1): 72–88. Publisher: SAGE Publications Inc.

**Brochet, Francois, Maria Loumioti, and George Serafeim.** 2015. "Speaking of the short-term: disclosure horizon and managerial myopia." *Review of Accounting Studies*, 20(3): 1122–1163.

**Bureau of Labor Statistics, U.S. Department of Labor.** 2025. "Job Openings and Labor Turnover Survey (JOLTS)." https://www.bls.gov/jlt/, Accessed: 2025-07-24.

**Carnevale, Anthony P., Tamara Jayasundera, and Dmitri Repnikov.** 2014. "Understanding online job ads data." A technical report. MS o. PP Center on Education and the Workforce. https ....

**Chen, Ciao-Wei, and Laura Yue Li.** 2023. "Is hiring fast a good sign? The informativeness of job vacancy duration for future firm profitability." *Review of Accounting Studies*, 28(3): 1316–1353.

**Choi, Hyeri, and Ioana Marinescu.** 2024. "Data for labor market concentration using Lightcast (formerly Burning Glass Technologies)." *Data in Brief*, 55: 110647. ISBN: 2352-3409 Publisher: Elsevier.

**Choudhury, Prithwiraj, Evan Starr, and Rajshree Agarwal.** 2020. "Machine learning and human capital complementarities: Experimental evidence on bias mitigation." *Strategic Management Journal*, 41(8): 1381–1411. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/smj.3152.

**Cifuentes, Manuel, Jon Boyer, David A. Lombardi, and Laura Punnett.** 2010. "Use of O*NET as a job exposure matrix: A literature review." *American Journal of Industrial Medicine*, 53(9): 898–914. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ajim.20846.

**Clemens, Jeffrey, Lisa B Kahn, and Jonathan Meer.** 2021. "Dropouts need not apply? The minimum wage and skill upgrading." *Journal of Labor Economics*, 39(S1): S107–S149.

**Cunningham, Jessica, Anna Hui, Julia Lane, and George Putnam.** 2021. "A Value-Driven Approach to Building Data Infrastructures: The Example of the MidWest Collaborative." *Harvard Data Science Review*, 4.

**Davis, Steven J., R. Jason Faberman, and John C. Haltiwanger.** 2013. "The establishment-level behavior of vacancies and hiring." *The Quarterly Journal of Economics*, 128(2): 581–622. ISBN: 0033-5533 Publisher: MIT Press.

**Deming, David J.** 2017. "The Growing Importance of Social Skills in the Labor Market*." *The Quarterly Journal of Economics*, 132(4): 1593–1640.

**DesJardine, Mark, and Pratima Bansal.** 2019. "One Step Forward, Two Steps Back: How Negative External Evaluations Can Shorten Organizational Time Horizons." *Organization Science*, 30(4): 761–780.

**DeVaro, Jed, and Peter Norlander.** 2021. "Wage Theft, Economic Conditions, and Market Power: The Case of H-1B Workers." *Economic Conditions, and Market Power: The Case of H-1B Workers (May 27, 2021).*

**Din, Alexander, and Ron Wilson.** 2020. "Crosswalking ZIP codes to census geographies." *Cityscape*, 22(1): 293–314.

**Dingel, Jonathan I., and Brent Neiman.** 2020. "How many jobs can be done at home?" *Journal of Public Economics*, 189: 104235.

**Eberts, Randall, and Christopher O'Leary.** 2003. "A Frontline Decision Support System for Georgia Career Centers." *Book Chapters.*

**Eccles, Robert G, Mirtha D Kastrapeli, and Stephanie J Potter.** 2017. "How to integrate ESG into investment decision-making: Results of a global survey of institutional investors." *Journal of Applied Corporate Finance*, 29(4): 125–133.

**Eloundou, Tyna, Sam Manning, Pamela Mishkin, and Daniel Rock.** 2024. "GPTs are GPTs: Labor market impact potential of LLMs." *Science*, 384(6702): 1306–1308.

**Frauenheim, Ed.** 2007. "What Killed America's Job Bank?" *Workforce Management*, 86(13): 33–. Place: Chicago Publisher: MediaTec Publishing Inc.

**Gathmann, Christina, Felix Grimm, and Erwin Winkler.** 2024. "AI, Task Changes in Jobs, and Worker Reallocation."

**Gephart, Robert.** 1997. "Hazardous measures: An interpretive textual analysis of quantitative sensemaking during crises." *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior*, 18(S1): 583–622.

**Gerhart, Barry, Patrick M Wright, GARY C MC MAHAN, and Scott A Snell.** 2000. "Measurement error in research on human resources and firm performance: how much error is there and how does it influence effect size estimates?" *Personnel psychology*, 53(4): 803–834.

**Gibbons, Eric M., Allie Greenman, Peter Norlander, and Todd Sørensen.** 2019. "Monopsony Power and Guest Worker Programs." *The Antitrust Bulletin*, 64(4): 540–565. Publisher: SAGE Publications Inc.

**Gunning, David, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang.** 2019. "XAI—Explainable artificial intelligence." *Science Robotics*, 4(37): eaay7120.

**Handa, Kunal, Alex Tamkin, Miles McCain, Saffron Huang, Esin Durmus, Sarah Heck, Jared Mueller, Jerry Hong, Stuart Ritchie, Tim Belonax, et al.** 2025. "Which Economic Tasks are Performed with AI? Evidence from Millions of Claude Conversations." *arXiv preprint arXiv:2503.04761*.

**Hansen, Stephen, Peter John Lambert, Nicholas Bloom, Steven J Davis, Raffaella Sadun, and Bledi Taska.** 2023. "Remote work across jobs, companies, and space." National Bureau of Economic Research.

**Haselhuhn, Michael P., Elaine M. Wong, and Margaret E. Ormiston.** 2022. "Investors respond negatively to executives' discussion of creativity." *Organizational Behavior and Human Decision Processes*, 171: 104155.

**Hashizume, Marissa.** 2024. "Timing Fortune 500 Job Opportunities."

**Hazell, Jonathon, Christina Patterson, Heather Sarsons, and Bledi Taska.** 2022. "National wage setting." *University of Chicago, Becker Friedman Institute for Economics Working Paper*, , (2022-150).

**He, Pengcheng, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen.** 2021. "DeBERTa: Decoding-enhanced BERT with Disentangled Attention."

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

**Heyden, Mariano L. M., Jatinder S. Sidhu, and Henk W. Volberda.** 2015. "The Conjoint Influence of Top and Middle Management Characteristics on Management Innovation." *Journal of Management*, 44(4): 1505–1529.

**Hirsch, Lesley, and Anna Hui.** 2024. "Reimagining the Workforce and Labor Market Information System for the 21st Century. Workforce Information Advisory Council Recommendations."

**Holland, John L.** 1997. *Making vocational choices: A theory of vocational personalities and work environments, 3rd ed. Making vocational choices: A theory of vocational personalities and work environments, 3rd ed*, Odessa, FL, US:Psychological Assessment Resources. Pages: xiv, 303.

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

**Horton, John J., and Prasanna Tambe.** 2025. "The Death of a Technical Skill." *Information Systems Research*. Publisher: INFORMS.

**Howison, Mark.** 2022. "Replication files for: "Occupational models from 42 million unstructured job postings"."

**Howison, Mark, Joe Long, and Justine Hastings.** 2023. "Recommending Career Transitions to Job Seekers Using Earnings Estimates, Skills Similarity, and Occupational Demand."

**Howison, Mark, William O. Ensor, Suraj Maharjan, Rahil Parikh, Srinivasan H. Sengamedu, Paul Daniels, Amber Gaither, Carrie Yeats, Chandan K. Reddy, and Justine S. Hastings.** 2024. "Extracting Structured Labor Market Information from Job Postings with Generative AI." *Digit. Gov.: Res. Pract.* Just Accepted.

**Kindermann, Bastian, Sebastian Beutel, Gonzalo Garcia de Lomana, Steffen Strese, David Bendig, and Malte Brettel.** 2021. "Digital orientation: Conceptualization and operationalization of a new strategic orientation." *European Management Journal*, 39(5): 645–657.

**Klein, Katherine, Mitchell Pontikes, Jeffrey A Dahlke, Dan J Putka, Brittany F Crawford, Matthew C Reeder, and Phil Lewis.** 2025. "Identification of Emerging Tasks in the O* NET System: A Revised Approach."

**Lancaster, Vicki, Devika Mahoney-Nair, and Nathaniel J Ratcliff.** 2019. "Technology report review of burning glass job-ad data." *University of Virginia, Biocomplexity Institute and Initiative Social and Decision Analytics Division*.

**Lane, Julia, and Nancy Potok.** 2024. "Democratizing Data: Our Vision." *Harvard Data Science Review*, , (Special Issue 4). https://hdsr.mitpress.mit.edu/pub/m1o4oblm.

**Lazer, David M. J., Alex Pentland, Duncan J. Watts, Sinan Aral, Susan Athey, Noshir Contractor, Deen Freelon, Sandra Gonzalez-Bailon, Gary King, Helen Margetts, Alondra Nelson, Matthew J. Salganik, Markus Strohmaier, Alessandro Vespignani, and Claudia Wagner.** 2020. "Computational social science: Obstacles and opportunities." *Science*, 369(6507): 1060–1062. Publisher: American Association for the Advancement of Science.

**Levy, Orly.** 2005. "The influence of top management team attention patterns on global strategic posture of firms." *Journal of Organizational Behavior*, 26(7): 797–819. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/job.340.

**Lewis, Phil, and Jeremiah Morris.** 2024. "Supplementing the O* NET Sources of Additional Information: A Preliminary Exploration of the Use of ChatGPT."

**Lewis, Phil, Christina Gregory, and Jeremiah Morris.** 2025. "Adding Drone-Specific Tasks to the O* NET Database: Initial Identification of Emerging Tasks using ChatGPT."

**Li, J., K. Larsen, and A. Abbasi.** 2020. *TheoryOn: A design framework and system for unlocking behavioral knowledge through ontology learning.* par.nsf.gov. Publication Title: MIS Quarterly.

**Littman, Julie.** 2021. "Why aren't restaurant workers coming back? Here's what the data shows. | Restaurant Dive."

**Li, Zehan, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang.** 2023. "Towards General Text Embeddings with Multi-stage Contrastive Learning."

**Lou, Bowen, and Lynn Wu.** 2021. "AI on Drugs: Can Artificial Intelligence Accelerate Drug Development? Evidence from a Large-Scale Examination of Bio-Pharma Firms." *Management Information Systems Quarterly*, 45(3): 1451–1482.

**Marinescu, Ioana, and Ronald Wolthoff.** 2020. "Opening the Black Box of the Matching Function: The Power of Words." *Journal of Labor Economics*, 38(2): 535–568. Publisher: The University of Chicago Press.

**McKenny, Aaron F., Herman Aguinis, Jeremy C. Short, and Aaron H. Anglin.** 2018*a*. "What Doesn't Get Measured Does Exist: Improving the Accuracy of Computer-Aided Text Analysis." *Journal of Management*, 44(7): 2909–2933.

**McKenny, Aaron F., Jeremy C. Short, David J. Ketchen Jr., G. Tyge Payne, and Todd W. Moss.** 2018*b*. "Strategic entrepreneurial orientation: Configurations, performance, and the effects of industry and time." *Strategic Entrepreneurship Journal*, 12(4): 504–521. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sej.1291.

**Meisenbacher, Stephen, and Peter Norlander.** 2023. "Transforming Unstructured Text into Data with Context Rule Assisted Machine Learning (CRAML)." arXiv:2301.08549 [cs].

**Moss, Todd W., Maija Renko, Emily Block, and Moriah Meyskens.** 2018. "Funding the story of hybrid ventures: Crowdfunder lending preferences and linguistic hybridity." *Journal of Business Venturing*, 33(5): 643–659.

**Mueller, Andreas I., Damian Osterwalder, Josef Zweimüller, and Andreas Kettemann.** 2024. "Vacancy durations and entry wages: Evidence from linked vacancy–employer–employee data." *Review of Economic Studies*, 91(3): 1807–1841. ISBN: 0034-6527 Publisher: Oxford University Press UK.

**National Academy of Sciences.** 2024. *Artificial Intelligence and the Future of Work.* Washington, D.C.:National Academies Press.

**Neuendorf, Kimberly A.** 2017. *The content analysis guidebook.* sage.

**Nguyen, Khanh Cao, Mike Zhang, Syrielle Montariol, and Antoine Bosselut.** 2024. "Rethinking Skill Extraction in the Job Market Domain using Large Language Models." arXiv:2402.03832 [cs].

**of New York, City.** 2025. "NYC Careers Portal."

**Olsen, Anders Ørding, Wolfgang Sofka, and Christoph Grimpe.** 2016. "Coordinated Exploration for Grand Challenges: The Role of Advocacy Groups in Search Consortia." *Academy of Management Journal*, 59(6): 2232–2255. Publisher: Academy of Management.

**O *NET Development, National Center for.** 2025. "O*NET OnLine."

**Pencle, Nadra, and Irina Mălăescu.** 2016. "What's in the Words? Development and Validation of a Multidimensional Dictionary for CSR and Application Using Prospectuses." *Journal of Emerging Technologies in Accounting*, 13(2): 109–127.

**Price, Ronald N., Jason Boyda, and Kathleen Bobay.** 2024. "Systems and Methods for Processing Data Using Interference and Analytics Engines."

**Putka, Dan J., Frederick L. Oswald, Richard N. Landers, Adam S. Beatty, Rodney A. McCloy, and Martin C. Yu.** 2023. "Evaluating a Natural Language Processing Approach to Estimating KSA and Interest Job Analysis Ratings." *Journal of Business and Psychology*, 38(2): 385–410.

**Reimers, Nils, and Iryna Gurevych.** 2019. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks." 3982–3992. Hong Kong, China:Association for Computational Linguistics.

**Rounds, James.** 2023. "Using Machine Learning to Develop Occupational Interest Profiles and High-Point Codes for the O* NET System."

**Rounds, James, Dan J. Putka, and P. Lewis.** 2022. "Updating vocational interests information for the O* NET content model." *National Center for O*NET Development. https://www. onetcenter. org/dl_files/Voc_Interests. pdf.*

**Rudin, Cynthia.** 2018. "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead."

**Russ, Daniel E., Kwan-Yuet Ho, Calvin A. Johnson, and Melissa C. Friesen.** 2014. "Computer-Based Coding of Occupation Codes for Epidemiological Analyses." 347–350. New York, NY, USA:IEEE.

**Russ, Daniel E, Pabitra Josse, Thomas Remen, Jonathan N Hofmann, Mark P Purdue, Jack Siemiatycki, Debra T Silverman, Yawei Zhang, Jerome Lavoué, and Melissa C Friesen.** 2023. "Evaluation of the updated SOCcer v2 algorithm for coding free-text job descriptions in three epidemiologic studies." *Annals of Work Exposures and Health*, 67(6): 772–783.

**Sauerwald, Steve, and Peter Norlander.** 2024. "Political Directors and the Recruitment of Foreign Workers." *Journal of Management*, 01492063241300311. Publisher: SAGE Publications Inc.

**Schubert, Gregor, Anna Stansbury, and Bledi Taska.** 2024. "Employer concentration and outside options." *Available at SSRN 3599454*.

**Short, Jeremy C., J. Christian Broberg, Claudia C. Cogliser, and Keith H. Brigham.** 2009. "Construct Validation Using Computer-Aided Text Analysis (CATA)." *Organizational Research Methods*, 13(2): 320–347.

**Stall, Shelley, Lynn Yarmey, Joel Cutcher-Gershenfeld, Brooks Hanson, Kerstin Lehnert, Brian Nosek, Mark Parsons, Erin Robinson, and Lesley Wyborn.** 2019. "Make scientific data FAIR." *Nature*, 570(7759): 27–29. Bandiera_abtest: a Cg_type: Comment Number: 7759 Publisher: Nature Publishing Group Subject_term: Databases, Research management, Policy, Publishing.

**Tolbert, Charles M, and Molly Sizer.** 1996. "US commuting zones and labor market areas: A 1990 update."

**U.S. Bureau of Labor Statistics.** 2025*a*. "Occupational Employment Statistics Data." *https://www.bls.gov/oes/*, Data extracted on July 24, 2025.

**U.S. Bureau of Labor Statistics.** 2025*b*. "Quarterly Census of Employment and Wages (QCEW)." *https://www.bls.gov/cew/downloadable-data-files.htm*, Accessed on July 24, 2025.

**U.S. Department of Labor, Employment and Training Administration.** 2022. "Career One Stop." *https://www.careeronestop.org/*.

**Van Holt, Tracy, and Tensie Whelan.** 2021. "Research frontiers in the era of embedding sustainability: Bringing social and environmental systems to the forefront." *Journal of Sustainability Research*, 3(2).

**Vaupel, Mario, David Bendig, Denise Fischer-Kreer, and Malte Brettel.** 2023. "The Role of Share Repurchases for Firms' Social and Environmental Sustainability." *Journal of Business Ethics*, 183(2): 401–428.

**Wasi, Nada, and Aaron Flaaen.** 2015. "Record linkage using Stata: Preprocessing, linking, and reviewing utilities." *The Stata Journal*, 15(3): 672–697. ISBN: 1536-867X Publisher: SAGE Publications Sage CA: Los Angeles, CA.

**Whelan, Tensie, and Ulrich Atz.** 2023. "What We Measure Today Does Not Matter When It Comes to Jobs." *Available at SSRN 5213954*.

**Yourish, Karen, Annie Daniel, Saurabh Datar, Isaac White, and Lazaro Gamio.** 2025. "These Words Are Disappearing in the New Trump Administration." *The New York Times*.

**Zachary, Miles A., Aaron McKenny, Jeremy Collin Short, and G. Tyge Payne.** 2011. "Family Business and Market Orientation." *Family Business Review*, 24(3): 233–251.

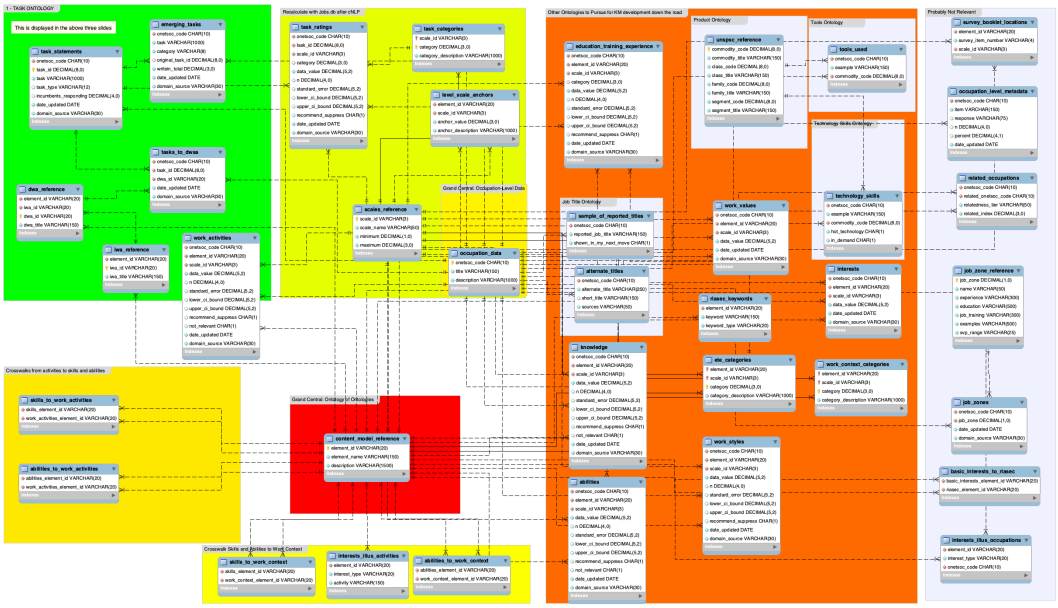
**Zhang, Mike, Kristian Jensen, Sif Sonniks, and Barbara Plank.** 2022. "SkillSpan: Hard and Soft Skill Extraction from English Job Postings." 4962–4984. Seattle, United States:Association for Computational Linguistics.

**Zheng, Lianmin, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica.** 2023. "Judging LLM-as-a-judge with MT-bench and Chatbot Arena." *NIPS '23*. Red Hook, NY, USA:Curran Associates Inc.

# A Appendix: O*NET Features Mapped

We initiated our work by inspecting the O*NET database scheme. We directly incorporated many of the O*NET tables and structures into the design of our work. Figure A.1 illustrates the entity relationship diagram for the database, and early notes we made to prioritize certain tables. Toward the center left, in dark red, is the Content Reference Model, a table is the "conceptual foundation of O*NET" and that joins all major features of the O*NET database into a single unified taxonomy for understanding the most important aspects of work.

Figure A.1: O*NET 29.1 Entity Relationship Diagram



In green on the top left side of the figure are task statements and crosswalks to work activities. In the yellow box to the right are occupational values for task intensity calculated in O*NET. The crosswalks between work activities (4.A.) and skills (2.A.) and abilities (1.A.) are on the lower left side in dark yellow. The box in dark red to its right is the Content Reference Model that ties together the various tables of the database. In orange on the right are tables that we used or considered using. On the far right are tables in O*NET that were less relevant to this work.

For the dataset we sought to create, O*NET provides the scaffolding, and in many cases the detail necessary for creating standard structured data. We pursued elements of the O*NET Content Reference Model with the opportunities provided by rich detail from frequently updated job ads data.

Table A.1 displays O*NET's content at a depth of two levels. Where the level of detail within the O*NET database and the survey method of data collection have limited O*NET's comprehensiveness, we augment its tables with real job ad text from NLx to boost the available training data for ML, and supplement O*NET by cross-walking its skill elements to the more elaborated ESCO taxonomy of skills.

Table A.1: O*NET Content Model Overview

| 1 Worker Characteristics | 2 Worker Requirements | 3 Experience Requirements |
|---|---|---|
| 1.A Abilities | 2.A Basic Skills | 3.A Experience and Training |
| 1.B Interests | 2.B Cross-Functional Skills | 3.B Basic Skills - Entry Requirements |
| 1.C Work Styles | 2.C Knowledge | 3.C Cross-Functional Skills - Entry Requirements |
| | 2.D Education | 3.D Licensing |

| 4 Occupational Requirements | 5 Occupation-Specific Information | 6 Workforce Characteristics |
|---|---|---|
| 4.A Generalized Work Activities | 5.A Tasks | 6.A Labor Market Information |
| 4.B Organizational Context | 5.C Title | 6.B Occupational Outlook |
| 4.C Work Context | 5.D Description | |
| 4.D Detailed Work Activities | 5.E Alternate Titles | |
| 4.E Intermediate Work Activities | 5.F Technology Skills | |
| | 5.G Tools | |

Table A.2: O*NET's structure covers nearly all elements related to work and provides scaffolding for extracting information from job ads.

#### A.0.1 O*NET 1: Worker Characteristics

O*NET includes measures of worker characteristics. Job ads often state a desire for worker characteristics. We did not seek to extract Abilities (1.A.), measures of which can be constructed with a cross-walk provided by O*NET that links Abilities to Work Activities (O*NET 4.A), (described in Section 3.1.3). Interests (1.B.) include general occupational interests (1.B.1) based on Holland's (1997) RIASEC framework, which we capture using a dictionary of RIASEC keywords provided by O*NET (Rounds, Putka and Lewis, 2022). We do not construct a separate model to complete elements of Work Values (1.B.2.) or Basic Occupational Interests (1.B.3) or Work Styles (1.C.). Some of these might be ascertained theoretically or by examining other sections of O*NET. For example, Dependability (1.C.5.a) as a worker characteristic is likely to be implicated in a job's description of a required shift or schedule (4.C.3.d.4), which is described below in Section A.0.4.

#### A.0.2 O*NET 2: Worker Requirements

Basic and Cross-Functional Skills (2.A. and 2.B.) can be constructed with a cross-walk provided by O*NET that links Skills to Work Activities (O*NET 4.A, described in Section 3.1.3). However, to increase the level of detail and directly capture empirical data on skills, we also develop a skill matching model (SkillMatch) described in Section 3.1.2 that is based upon the more elaborately detailed European Skills, Competences, and Occupations (ESCO) database v 1.20.

We do not construct a separate model for Knowledge requirements (2.C.). We build a dictionary (see Table E.3) to capture the Required Level of Education (2.D.1). We do not capture field of education requirements (2.D.3.). We capture jobs that reference Spanish language skills, but not other

languages.

### A.0.3 O*NET 3: Experience Requirements

We do not capture the number of years of related work experience required (3.A.1). We capture whether training is described in a job ad (3.A.3.). We do not systematically capture which basic or cross-functional skills are specifically mentioned as requirements for entry into the occupation (3.B. or 3.C.); these could perhaps be calculated from data we provide on skills in entry-level jobs within an occupation. We do calculate a Fleisch-Kincaid readability score for each job posting, which informs reading comprehension entry requirements (3.B.1.a). We capture whether an occupational license is indicated in the job posting (3.D.), but not which one (3.D.2.), or whether it is preferred or required. We capture several entry requirements that are imposed by the organization (3.D.5) or government (3.D.5.a), such as drug and criminal background checks (see Table E.4) and requirements related to work authorization and visa sponsorship. We capture presence of a labor union and/or a professional association (3.D.5.C).

### A.0.4 O*NET 4: Occupational Requirements

O*NET hierarchically structures generalized work activities (4.A.) into intermediate activities (4.E.) that span occupations, and detailed (4.D.) work activities that reside within an occupation, which are further detailed in over 20,000 occupation-specific task statements (5.A). We identify task statements within each job advertisement separately from identifying occupational codes from titles. Technical detail is provided in Section 3.1.3. Therefore, we do not impose a requirement built into O*NET that a task lies within only a single occupation.

Organizational Context (4.B.) includes "characteristics of the organization that influence how people do their work." Organizational context includes structural characteristics (4.B.1) such as human resources systems and practices (4.B.1.b.) and recruitment and selection processes (4.B.1.b.1). We develop novel dictionaries for Diversity, Equity, Inclusion, and Belonging (DEIB), elements of the rewards system (4.B.1.b.3), including benefits from a dictionary we build. We develop a custom NER model ("WageExtract") to extract wage information from job ad text (see section 3.1.5).

Social Processes (4.B.2) include culture, values and principles of the organization. A number of the dictionaries that relate generally to organizational context (4.B.) are standard dictionaries from the business and management literature, presented in the Appendix E.6.

Work Context (4.C.) includes features we capture, including whether work must be done physically close to others, and required work schedules for the position (see Appendix Table E.5).

### A.0.5 O*NET 5: Occupation-Specific Requirements

Task (5.A.) extraction is described further in the methods section 3.1.3 on TaskMatch. Titles (5.C.) and Alternate Titles (5.E.) are described above as related to occupation, with additional technical detail on "TitleMatch" in Section 3.1.1. Many features from titles are extracted and described in Table E.1. A dictionary of technologies (5.F.1.) and tools (5.G.1) is included in O*NET.

The O*NET database includes United Nations Standard Products and Services Codes (UNSPSC) for a list of bespoke tools ranging from 'abdominal binders' to 'Zylonite files', as well as technologies such as 'Microsoft Office.' We remove terms that generate many false positives. For example, we removed 96 items that generated excessive false positives (e.g., "scale", "range") from the O*NET tools and technologies dictionaries. We add Artificial Intelligence keywords from Lou and Wu (2021) and a novel list of AI related terms.

### A.0.6 O*NET 6: Workforce Characteristics

Labor Market Information (6.A.) and Occupational Statistics (6.A.1.) incorporate "information related to economic conditions and labor force characteristics of occupations." We utilize job advertisements to calculate occupational and industry labor market concentration within commuting zones using the Herfindahl Hirschman Index (HHI), as done in earlier research (Azar, Marinescu and Steinbaum, 2022) with CareerBuilder.com data. While we do not attempt an Occupational Outlook (6.B.) or Projections (6.B.1.), data extracted here could be well-suited for that purpose.

# B Descriptive, Robustness, and Comparison Figures

This appendix contains additional figures comparing the industry, occupation, and wage distributions dataset with official government benchmarks (QCEW and OES) in multiple time periods.

## B.1 Annual Industry Distribution Comparison (vs. QCEW)

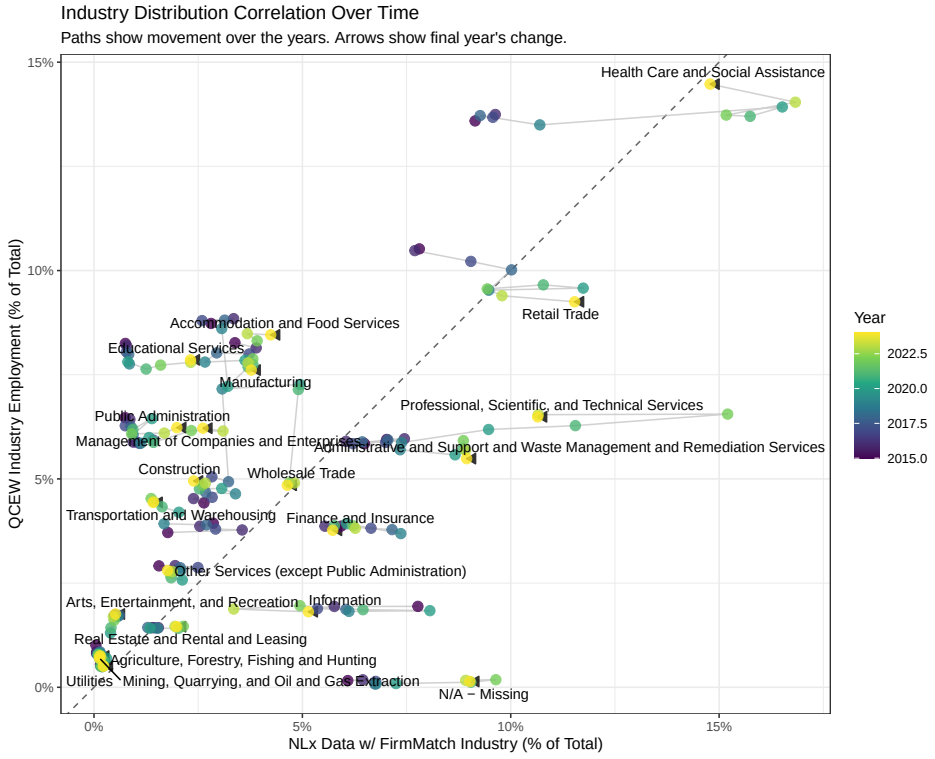



Figure B.1: Trajectory of Industry Correlation Over Time.

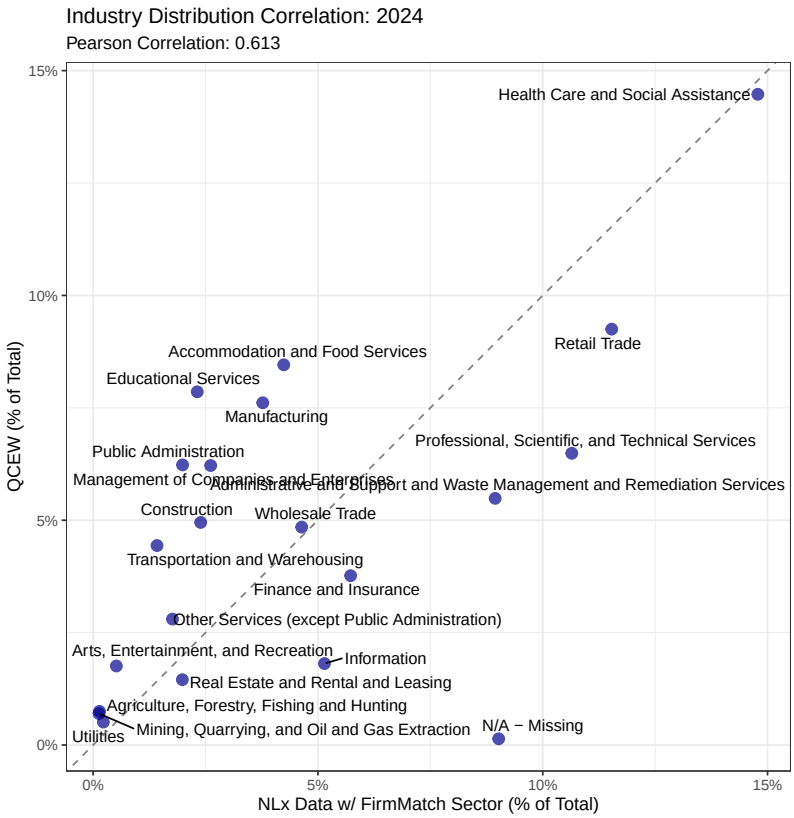



Figure B.2: Industry Correlation, 2024.

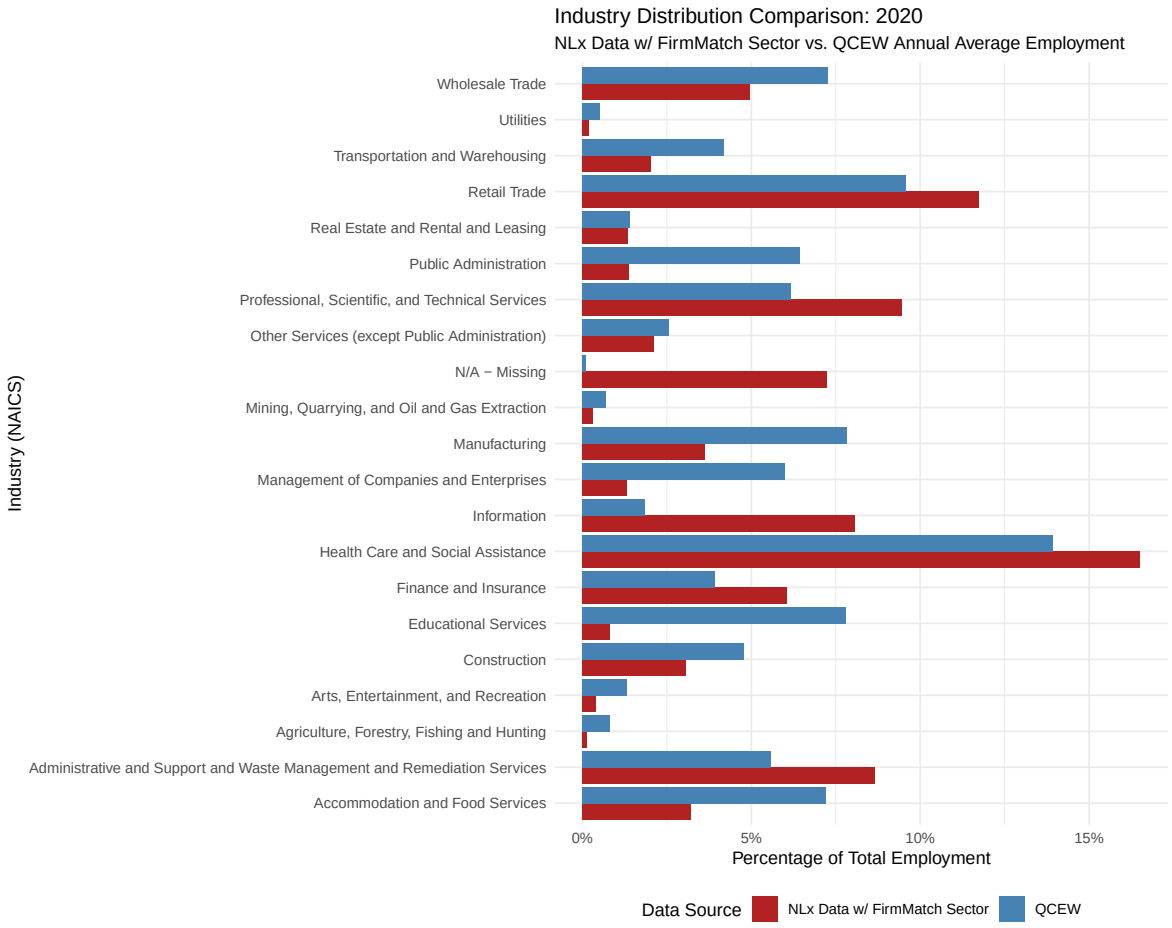



Figure B.3: Industry Distribution Comparison, 2020.

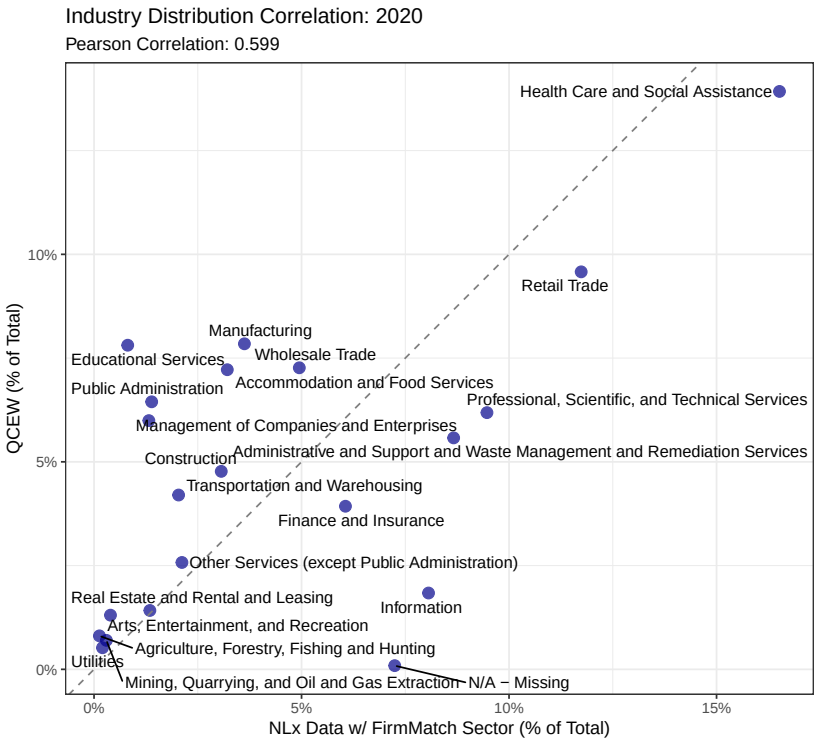



Figure B.4: Industry Correlation, 2020.

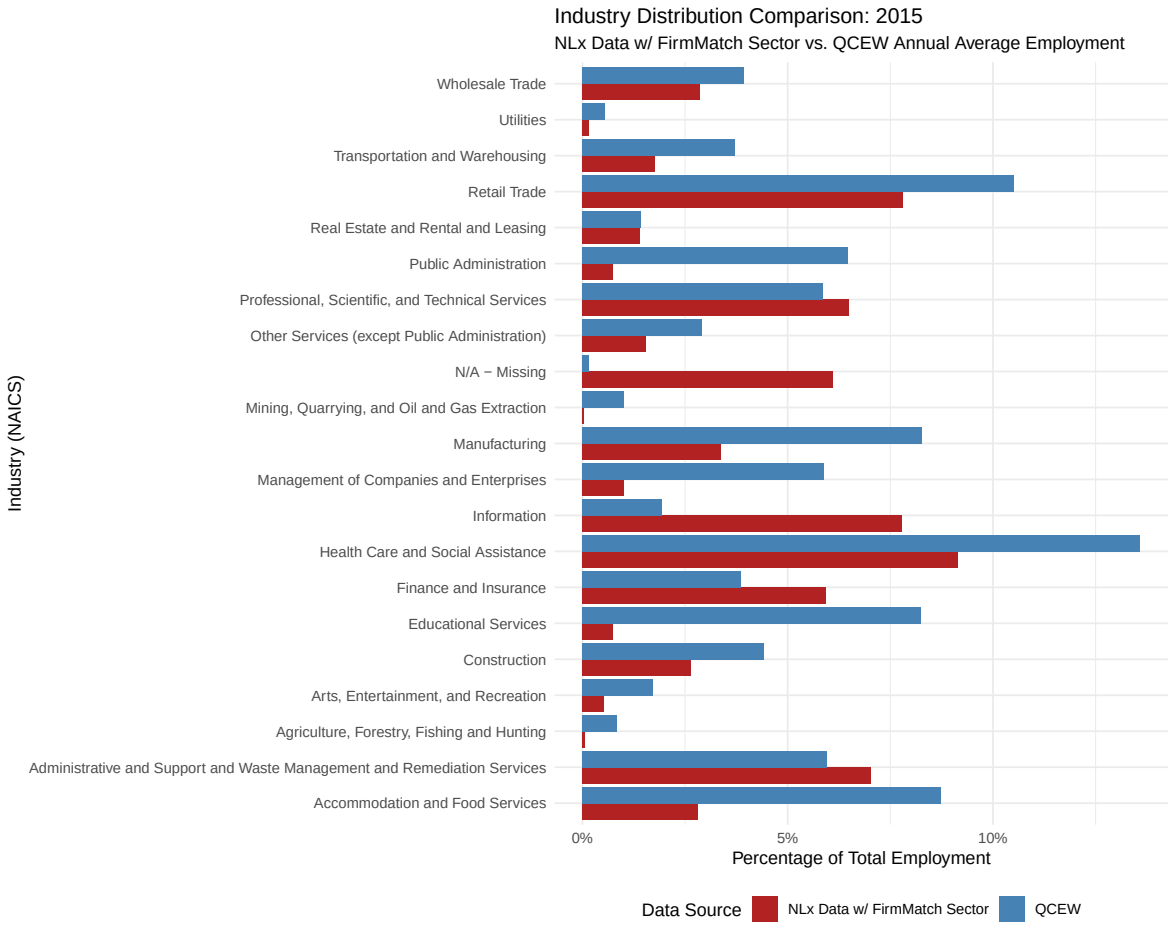



Figure B.5: Industry Distribution Comparison, 2015.

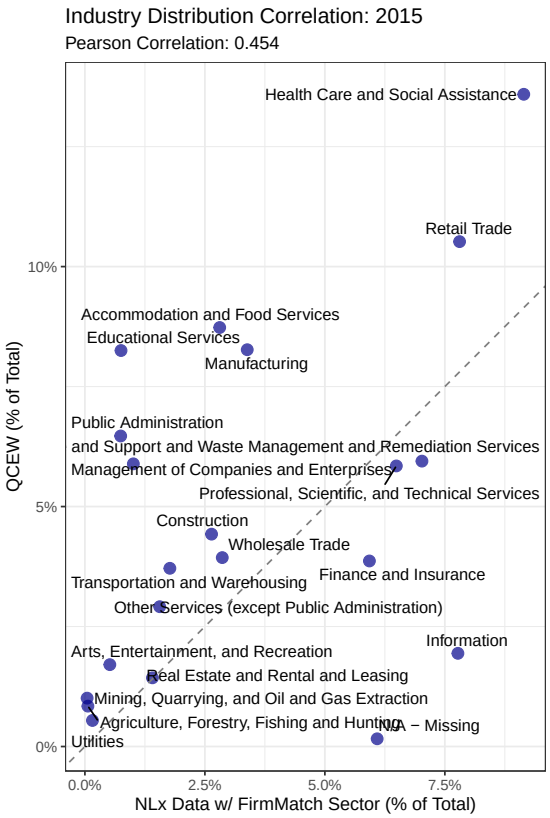



Figure B.6: Industry Correlation, 2015.

## B.2 Annual Occupation Distribution Comparison (vs. OES)

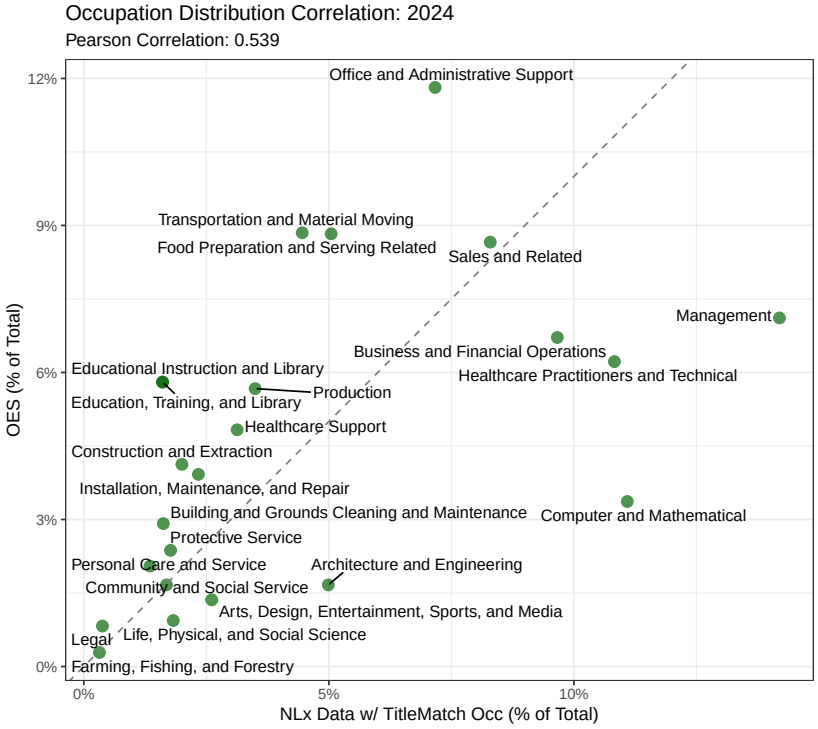



Figure B.7: Occupation Correlation (3-Year Sample), 2024.

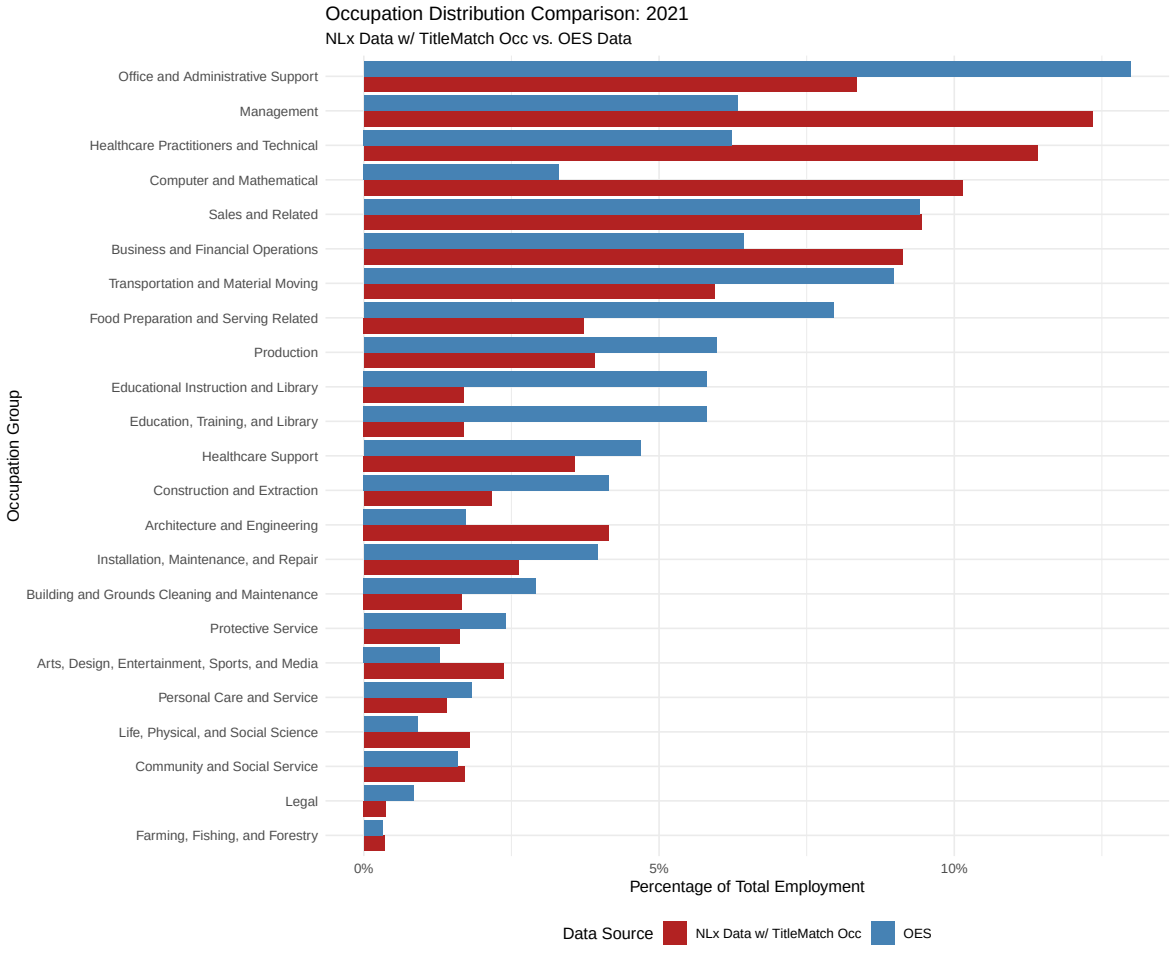



Figure B.8: Occupation Distribution Comparison (3-Year Sample), 2021.

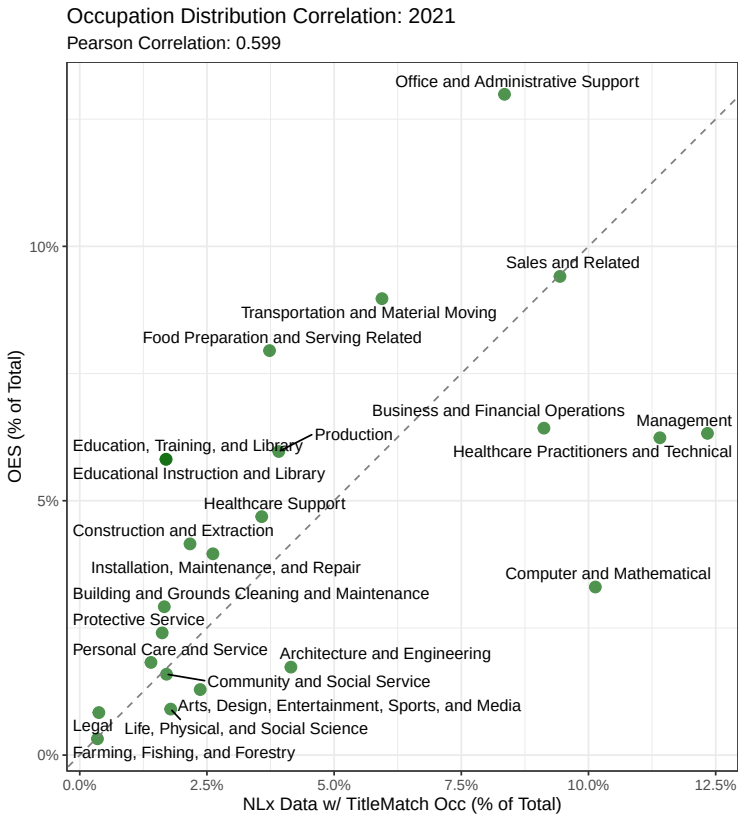



Figure B.9: Occupation Correlation (3-Year Sample), 2021.

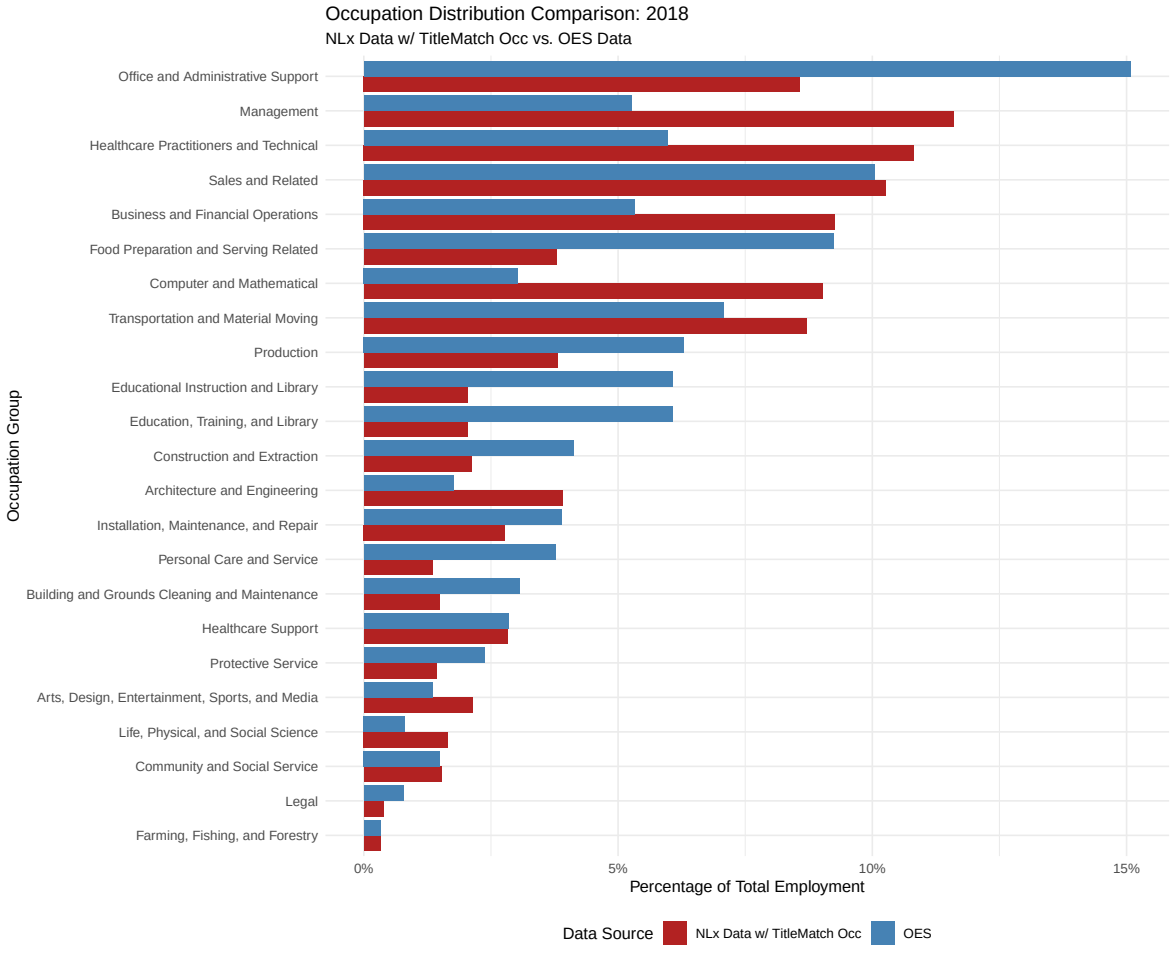



Figure B.10: Occupation Distribution Comparison (3-Year Sample), 2018.

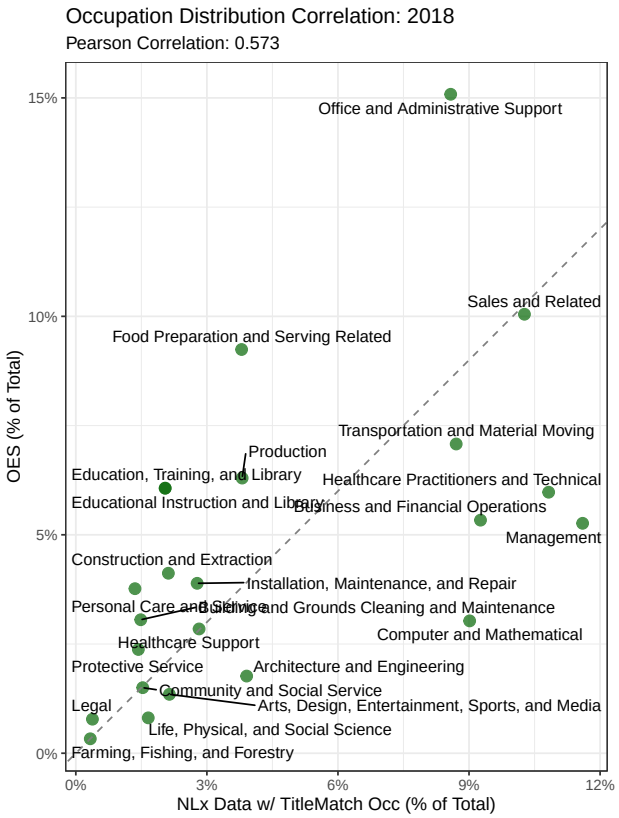



Figure B.11: Occupation Correlation (3-Year Sample), 2018.

## B.3 2015-2019 Wage Distribution Comparison (vs. OES)

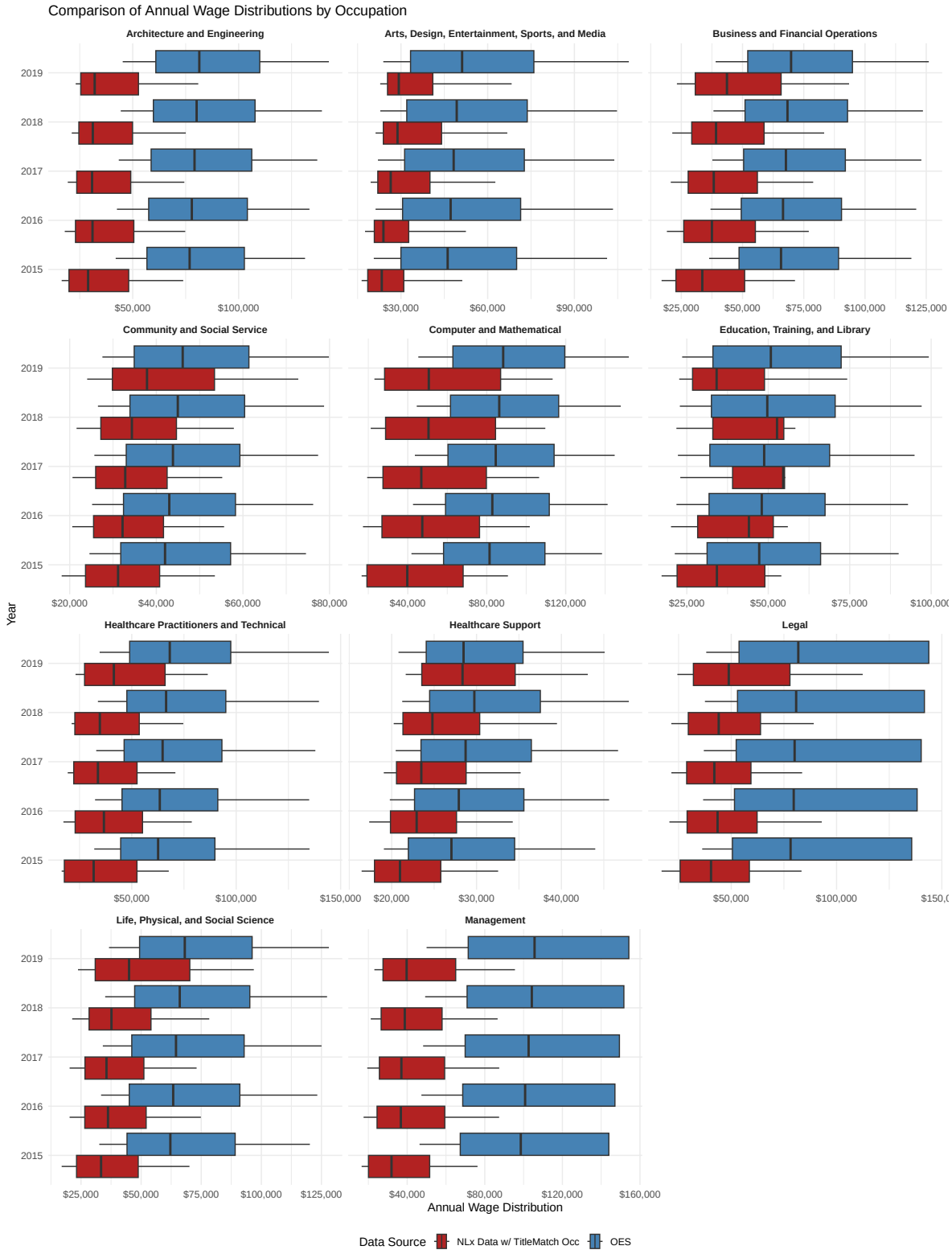



Figure B.12: Wage Distributions, 2015-2019, for SOC codes below 32.

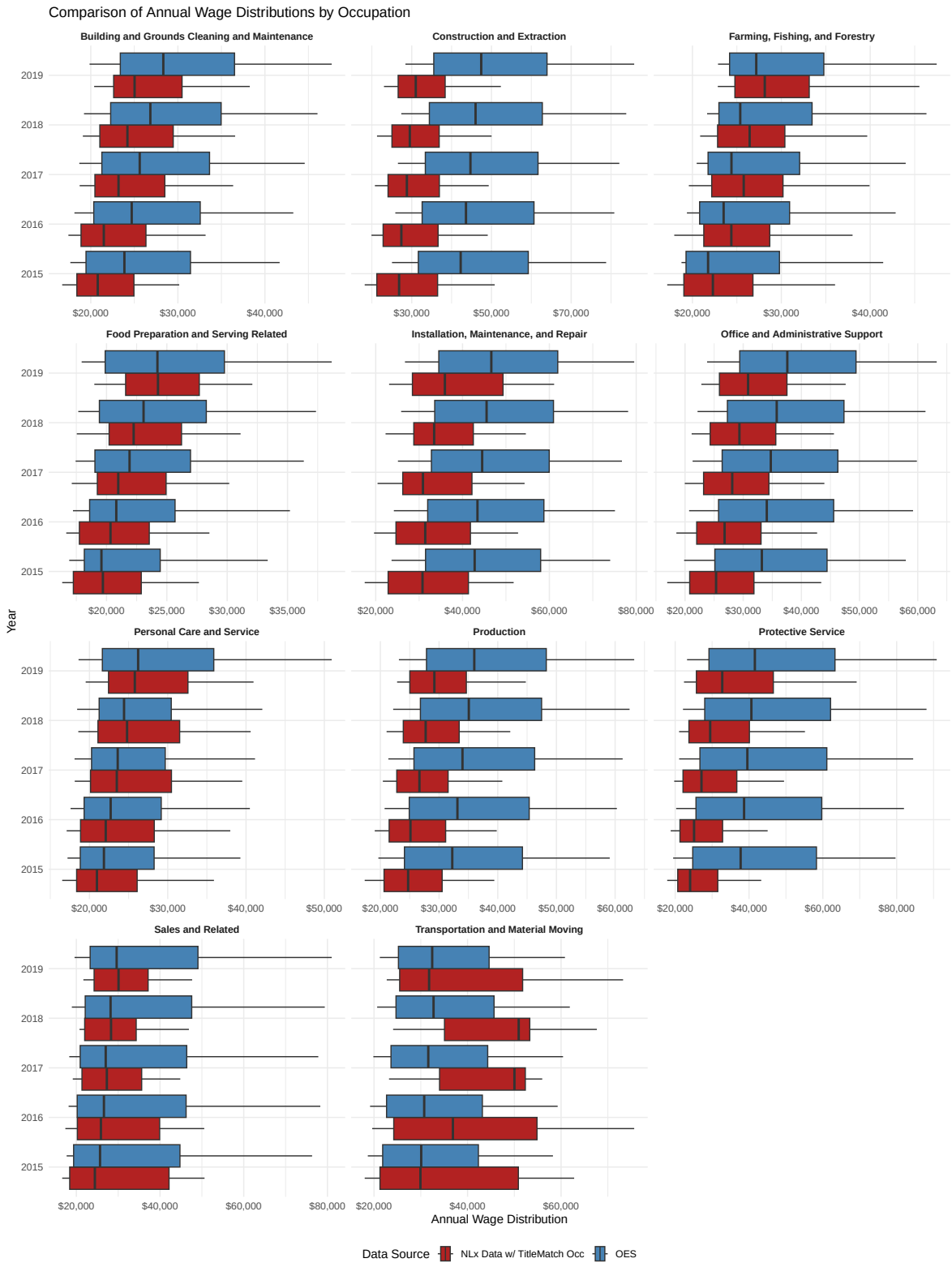



Figure B.13: Wage Distributions, 2015-2019, for SOC codes above 32.

# C Detailed Methodology and Validation Steps

The harmonic mean of precision and recall (or F1-score) is frequently used to evaluate model performance. When using this metric, a high score (i.e., towards 1) is important to instill confidence that a trained model has not only learned the training data well, but can also reasonably generalize beyond it. We report F1 scores in Table 1, but do not rely solely on them.

The basis for all models we develop are "dictionaries" or lists of terms with standardized labels. In cases where no taxonomy exists or existing taxonomies have insufficient detail, we follow in the tradition of interpretive text analysis (Gephart, 1997): we engage in inductive and deductive sampling of text, qualitatively analyze real job ad language, label it manually, and use machines to scale unseen instances with embeddings. Validity in the content analysis tradition typically starts with keyword searches, and including strategic and random audits of labeled output to ensure high content validity (Neuendorf, 2017). Oversampling and manually annotating the most frequently occurring text patterns, random sampling from bins across the frequency distribution of match scores, and random sampling of text chunks labeled by models illustrates where problems occur and enhances the ultimate model performance.

We first describe the approach undertaken to building and validating dictionaries, and then describe how we bridge standard or novel dictionaries with embeddings and ML methods to transform text into structured data in the job ad domain, and provide additional details of the model training procedures. We then provide additional detail on procedures undertaken for several of the key models.

### C.0.1 Dictionaries

We adopt standard dictionaries from O*NET and build novel dictionaries, but validate both by auditing binned results by frequency of appearance of features of a large sample of text. We focus effort on manual reviews of the most frequently appearing observations. For each dictionary, we also audit small random samples of the results to check accuracy and are satisfied when over 90% are accurate. For example, we find 98/100 identifications of a tool or technology were true positives. A limitation of reliance on benchmark / standard dictionaries is that emerging tools and technologies, e.g., 'ChatGPT', will not have a commodity code or entry in the O*NET. Another is that stemmed words may not always match the desired concept. We manually augment standard dictionaries, build custom dictionaries, and make direct edits to the results until satisfied.

### C.0.2 Bridging Dictionaries and Machine Learning

Dictionaries lack context and require matching an exact text string. We initially attempted to split the job advertisements into sections for targeted extraction, e.g., 'skill requirements' or 'task' sections.

Had that been successful, dictionaries operating within specific sections could have been a fruitful approach to context-specific results. However, job advertisements are free form and do not obey a uniform structure.

For tasks and skills, to prevent false positives, we build two-stage models in which we first identify sentences as 'task' or 'skill' sentences with a binary ML classifier trained to recognize sentences containing task or skill descriptions relevant to the worker. We built initial lists of task and skill sentences with content from O*NET and ESCO, respectively. We then pass only the sentences that contain task or skill content on to a similarity search embedding model.

To augment dictionaries and rule-based approaches, we use word and sentence embeddings (Li, Larsen and Abbasi, 2020). We thereby generalize from a dictionary to capture the occurrence of a concept. For embedding models, we typically begin with an existing dictionary, augment it with highly similar real job ad language, and embed the results to complete a search for similar text in the job ads. In some cases, we add supervised ML built with training data created from the dictionary and boosted by real job ad text identified by embeddings with a high match score following audits of the training data.

### C.0.3 Model Training Procedures

All procedures involving the fine-tuning of pretrained Language Models were performed using the HuggingFace Trainer library[4], using default parameters – including a learning rate of 5e-5 and the Adam optimizer. Training was performed for one epoch on the 90% train split, which was always obtained using a random seed of 42. All training was run on a single Nvidia RTX A6000 GPU, and the utilized batch size was tailored to the maximum sequence length and pretrained model size.

Models requiring the entire job ad text as context (i.e., FirmExtract) were limited to 1024 tokens, whereas models trained on sentences (TaskMatch, SkillMatch, binary WageExtract model) were trained with a maximum length of 64 (roughly 3x the average English sentence length), or 128 in the case of the two larger WageExtract models. TitleMatch models were similarly restricted to 32 tokens, since only the title text is given as input to these models.

### C.0.4 LLM-as-a-Judge Validation Prompts

In Table C.1, we provide the prompt used to validation the binary classification stage of SkillMatch. Few-shot examples were curated from the human-coded dataset used to train the classification model.

Similarly, in Table C.2, we provide the prompt used for the second part of the LLM-as-a-Judge validation, namely on the matching results for Skill Match.

[4] https://huggingface.co/docs/transformers/en/main_classes/trainer

Table C.1: Prompt for LLM-as-a-Judge Validation of Binary Skill Classification.

| **Prompt** |
|---|
| You will be given a sentence. Your task is to decide whether the given sentence contains a skill statement or not. A skill is the ability to perform a specific task and apply knowledge, particularly in the work context. Answer simply with SKILL or NOT_SKILL, denoting that the sentence contains a skill statement or not, respectively.<br>Provide your feedback as follows:<br>Output:::<br>Classification: (SKILL or NOT_SKILL)<br>Here are some examples:<br>sentence: listens to what other people are saying and asks questions as appropriate.<br>Output:::<br>Classification: SKILL<br>sentence: *position summary: * customer service - country code top-level domain (cctld) specialist's work within the client service organization supporting our client base of primarily fortune 100 companies.<br>Output:::<br>Classification: NOT_SKILL<br>sentence: responsible for delivering and serving food and beverage to guests in a friendly, prompt and efficient manner.<br>Output:::<br>Classification: SKILL<br>sentence: this position will travel onsite to the client's charlotte, nc location as needed for client meetings.<br>Output:::<br>Classification: NOT_SKILL<br>Now here is the sentence.<br>sentence: [INPUT SENTENCE]<br>Output:::<br>Classification: |

Table C.2: Prompt for LLM-as-a-Judge Validation of Skill Matching.

| **Prompt** |
|---|
| You will be given a SENTENCE and a CANDIDATE. Your task is to decide whether the given CANDIDATE is an appropriate match for the given SENTENCE. To be an appropriate match, the CANDIDATE should accurately represent what skills are described in the SENTENCE. A skill is the ability to perform a specific task and apply knowledge, particularly in the work context. Answer simply with MATCH or NOT_MATCH, denoting that the CANDIDATE matches the skill statement in the SENTENCE or not, respectively.<br>Provide your feedback as follows:<br>Output:::<br>Classification: (MATCH or NOT_MATCH)<br>Here are some examples:<br>SENTENCE: operate, calibrate, and maintain lab equipment.<br>CANDIDATE: operating scientific and laboratory equipment<br>Output:::<br>Classification: MATCH<br>SENTENCE: workers will be exposed to all types of weather conditions.<br>CANDIDATE: operating lifting or moving equipment<br>Output:::<br>Classification: NOT_MATCH<br>SENTENCE: problem solving .<br>CANDIDATE: solving problems<br>Output:::<br>Classification: MATCH<br>SENTENCE: uses time effectively to manage workload/tasks.<br>CANDIDATE: documenting technical designs, procedures, problems or activities<br>Output:::<br>Classification: NOT_MATCH<br>Now here are the actual texts.<br>SENTENCE: TEST<br>CANDIDATE: TEST<br>Output:::<br>Classification: |

## C.1 TaskMatch

We first attempted a dictionary of tasks based on analysis of parts of speech in the O*NET task statements. Identifying unique noun-verb pairs within the text of O*NET task statements, and authoring rules requiring pairs appear within a narrow context window, proved unsatisfying in both recall and precision. To boost recall, we augmented task verbs and nouns from the O*NET with highly similar verbs and nouns suggested by embeddings. However, many positive identifications proved to be false as the sentences in context did not truly contain information relevant to the duties to be performed by the worker, i.e., were not sentences that contained relevant task information. We used labeled data from this analysis in the next step.

To identify which sentences in a job ad contain task information, we built a binary BERT-based classifier to identify task sentences – sentences that contains description of tasks a worker might perform on the job. To build this classifier, we used the 20,000 task statements from O'NET as a basis for training data. Next, we used an augmentation approach, where a random sample of 100,000 job ads was segmented into sentences, and using an Sentence Transformer embedding model (Reimers and Gurevych, 2019), specifically GTE-SMALL (Li et al., 2023), we matched the most and least semantically similar sentences, thus augmenting the dataset to contain other plausible task statements, as well as clear examples of "not-task" statements.

A manual examination of the augmented data suggested by embeddings confirmed that task statements were relevant, and the non-task sentences were irrelevant to task classification. In an additional step, a random audit of 1000 training data samples was performed to identify clear misclassifications resulting from the matching. This validation feedback was used to re-run the augmentation process, resulting in the final training dataset of nearly 150,000 texts (44k task, 106k not task). Finally, an efficient, compact version of a BERT model (BERT-TINY) was fine-tuned on the training dataset for one epoch to produce a binary task classification model. This model was chosen due to its compactness and ability to be run efficiently (even on CPU). Only statements identified by the binary text classification model to be task statements were considered in the semantic matching process described below.

To build the second stage of TaskMatch, we embedded O*NET's task statements and searched for similar task sentences identified by the binary classifier from a large random sample of NLx job ads. In validation, we randomly audited within bins of the similarity score. We identified that above an embedding match score of 0.90 (i.e., cosine similarity), we obtained excellent precision scores (7 false positives / 165 reviewed), and the quality deteriorated below that level (65 false positives / 90). Discarding the results below 0.90 meant dropping 60% of the results. We observed that while many of the false positives were false in a strict sense, the embedding model often caught interesting

similarities between tasks and suggest general concepts that are false only in that they differ from the precise meaning of the O*NET task statements. In interactively testing the model on a small sample of individual job ads from different firms and sectors, while recall was at times an issue, the accuracy of results obtained was impressive. As we sought a highly precise model, we deployed TaskMatch with the 0.90 threshold.

## C.2 FirmExtract

This training dataset was prepared by randomly sampling 200,000 job ads from the NLx corpus in which the self-report firm name was present. An algorithm was created to match all instances of this firm name in the text (lower case, shortened title, etc.), thereby creating quality tagged NER data. The firmNER model was trained for one epoch on a 90% split of our training dataset, with a sequence classification objective (all tokens are labeled as either part of a firm name or not), and this model achieved an F1 score of 94.5 on the validation set.

We first standardize Data Axle firm names using the same cleaners discussed above, and then calculate the Levenshtein Distance similarity ratio between standardized firm names. We accept a match when the match score is above 0.8. If no match is found at the zip code level, we seek the highest-scoring match within state, and if no match above 0.8 is found at the state level, we retrieve the highest-scoring match nationally above 0.8. We developed this process after auditing three rounds of firm-matching results, focusing on accurate identification of the largest firms.

We leverage the availability of zip code and state information in both job ads and Data Axle to match firms to establishments at the zip code level if possible. Based on a manual audit, we use 0.8 as a threshold for fuzzy matching. If no match is above 0.8 at the zip code level, we match firm names at the state level. If no match is above 0.8 at the state level, we match against the national list of all firm names, and expedite the search by looking only at firms with the first character in common with the job ad firm name. Although we retain the best match and match score, in the dataset we construct, we treat all results below 0.8 as missing industry.

## C.3 TitleMatch

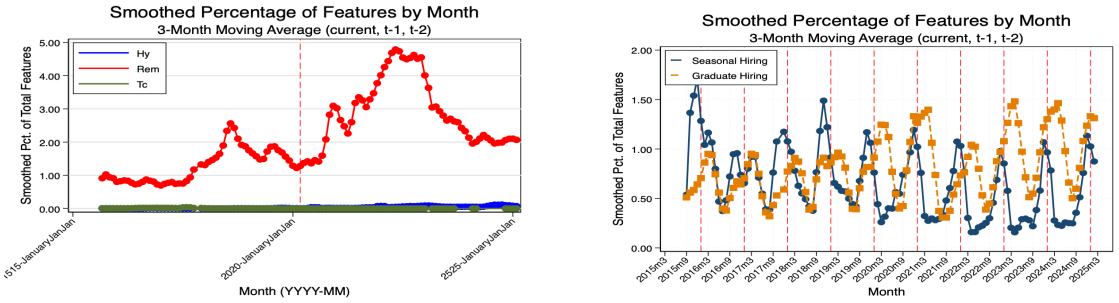
There is no accepted validation resource to match job titles to occupation codes at scale. We take multiple approaches to confirm the final model. After an initial round of exact matching between reported job ad titles from O*NET and job ad titles, we augmented the training data of reported titles with high-match scoring and manually-audited high-frequency titles that could not be exact matched. We manually deleted high-frequency false positives from the sample of reported titles in O*NET, and corrected high-frequency false positives in the training data when multiple occupation

codes were matched based on a dictionary approach. For example, "delivery" in isolation is often correctly associated with the courier occupation, but the co-appearance of "delivery" and "nurse" in a title is corrected to associate with the nursing occupation. We compared results between Version 1 (sample of reported titles embeddings) and Version 2 (sample of reported titles augmented by real job titles suggested by TitleMatchV1). We find an exact match of 89.5%, and manually audit a strategic sample of 100 job titles from across the distribution of match scores.

We also compared TitleMatch's output to results form publicly available research. Drawing from a newspaper corpus of help wanted ads from the 1950s through 2000, Atalay et al. construct job titles, and using a bag of words model, predict standard occupation codes. The underlying job titles in newspapers have challenges not faced in online job ad data: drawn from OCR scans of newspapers, it is subject to transcription errors, and titles could be misidentified. Newspaper job ad titles are also much shorter than online job ad titles, and the job ad titles are much older in newspapers, reflecting an earlier time perido (Marinescu and Wolthoff, 2020). Nevertheless, in a 1% sample of job titles, for 1,124 titles where the Atalay et al. model returns a confidence level of 1, TitleMatch retrieves the same major occupational group in 91% of observations and the same title or six-digit occupational code for 74% of the observed job titles. This suggests a high degree of out of sample correspondence.

We developed a list of features in titles by examining a random sample of job ad titles and noticing words that were not indicators of hierarchy or occupation. For example, we notice urgent hiring, sign-on bonuses, travel, remote work, seasonal, part-time, and full-time work. In Appendix Table **??**, we list the unique features we extract and the associated codes. Figure C.1 illustrates trends in the presence of remote work in job titles in Panel A, and cyclical hiring trends for positions with titles that indicate seasonal and college graduate hiring. To extract such features from job titles, we fine-tuned an additional DeBERTa-v3-base model in a multi-label setting, and the resulting model achieves an 81.4% accuracy in correctly identifying the labels present in a title.

Figure C.1: TitleMatch Features: Remote Work and Seasonal Hiring




(a) Remote work before and after the Pandemic (March 2020 vertical line)

(b) Seasonal and Graduate Hiring (January vertical line).

To extract the hierarchy values, we built a knowledge map of terms that indicate hierarchy. We exhaustively searched for but do not display in the excerpts all the variant keywords we include in the production-use map. Null results default to zero. A second 'stepper' knowledge map returns a value ranging from -7 to +4. Words such as "assistant" and "vice" decrement the base level, and words such as "senior" or "chief" increment the base level. The sum of the base and the 'stepper' maps is equal to the hierarchy level value, and we construct a training dataset of nearly 245k text examples from NLx using this base knowledge map and stepper logic.

The training dataset maps title texts to hierarchy values, and we use this data to train BERT-based regression model, i.e., to predict the hierarchy value given a text input. Due to the semantic intricacy of the task, we choose to fine-tune a DeBERTa-v3-base model (He et al., 2021) for regression, which we perform on a 90% train split for one epoch.

Table C.3: Mapping Hierarchy to Job Titles

| Keyword | Value | Label |
|---|---|---|
| Internship | -10 | Intern level |
| Trainee | -10 | Intern level |
| Entry-Level | 0 | Base level |
| Manager | 10 | First-Level Supervisor |
| Supervisor | 10 | First-Level Supervisor |
| Team Leader | 10 | First-Level Supervisor |
| Territory Manager | 20 | Second-Level Supervisor |
| Division Leader | 30 | Third-Level Supervisor |
| General Manager | 30 | Third-Level Supervisor |
| Director | 40 | First-level Executive |
| CHRO | 50 | Senior Executive |
| CEO | 60 | Top Management |

1A. Excerpt of Base Hierarchy Map.

| | | |
|---|---|---|
| Helper | -7 | Helper |
| Junior | -6 | Junior |
| Asst | -5 | Assistant |
| Associate | -3 | Associate |
| Vice | -2 | Vice |
| Deputy | -1 | Deputy |
| Lead | 1 | Lead |
| Leader | 1 | Leader |
| Sr | 2 | Senior |
| Exec | 3 | Executive |
| Chief | 4 | Executive |

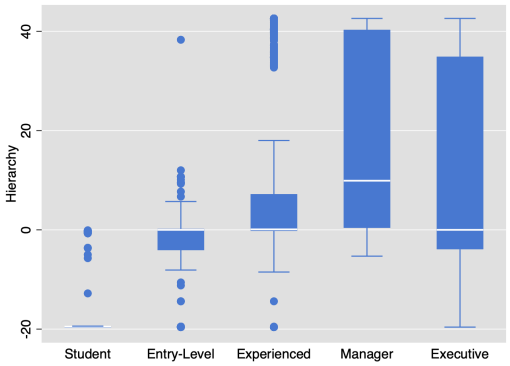
1B. Excerpt of Hierarchy Stepper Map

We assess the accuracy of the hierarchy match by running TitleMatch on 3,219 New York City job ads downloaded on March 24, 2025 (of New York, 2025). NYC job ad metadata includes five career levels (Student, Entry-Level, Experienced, Manager, and Executive). Figure C.2 illustrates the results. With the exception of executive titles, Panel A's boxplot illustrates that the distribution of the model's predicted hierarchal level corresponds to student, entry-level, experienced, and managerial positions. Inspection shows that 'commissioner' appears frequently in NYC executive rank postings, but was not in our hierarchy coding model and note this for future improvements. Panel B's density plot shows that the distribution of TitleMatch's hierarchy results correspond to the internal hierarchical scheme of NYC's personnel management system. For Panel B, we map TitleMatch predictions to the Career Levels used by NYC Job Ads - we divide by 10, round to the nearest whole number, and truncate the hierarchy level returned by TitleMatch to match the five levels of the NYC system.

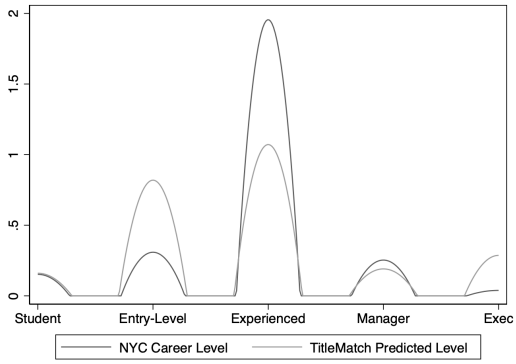
Overall, the correlations between TitleMatch's hierarchy level and the minimum salary range (0.41) and the top salary range (0.49) for positions, and the career level (0.48), are consistent with a moderate

positive association between this measure and important characteristics of the job. In many cases where wage information is unavailable, this measure can be an informative indicator of a job's status within an organization.

Figure C.2: TitleMatch Hierarchy Prediction and NYC Job Ad Career Levels




(a) Boxplot: TitleMatch Hierarchy Prediction Matched to NYC Career Levels

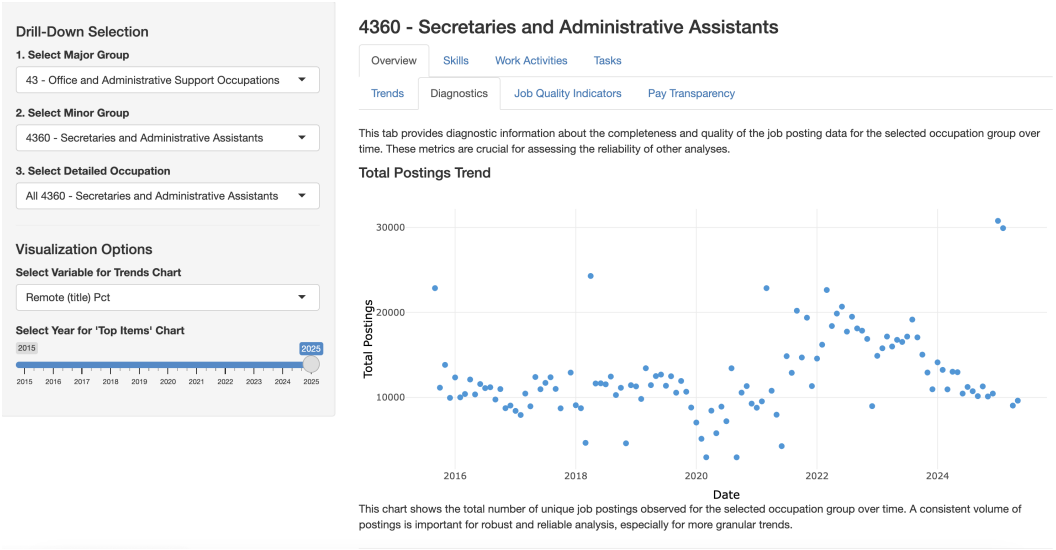



(b) Density Plot: TitleMatch Hierarchy Prediction Similar to NYC Career Levels

# D Appendix Sample Occupational Report for Secretaries and Administrative Assistants.

A Shiny App enables inspections of the data. This report illustrates some of the data available in a report focused on Secretaries and Administrative Assistants, and includes SOC codes 43-6011, 43-6012, 43-6013, and 43-6014.

Figure D.1: Detailed Occupation Selected with Total Job Postings per Month Displayed

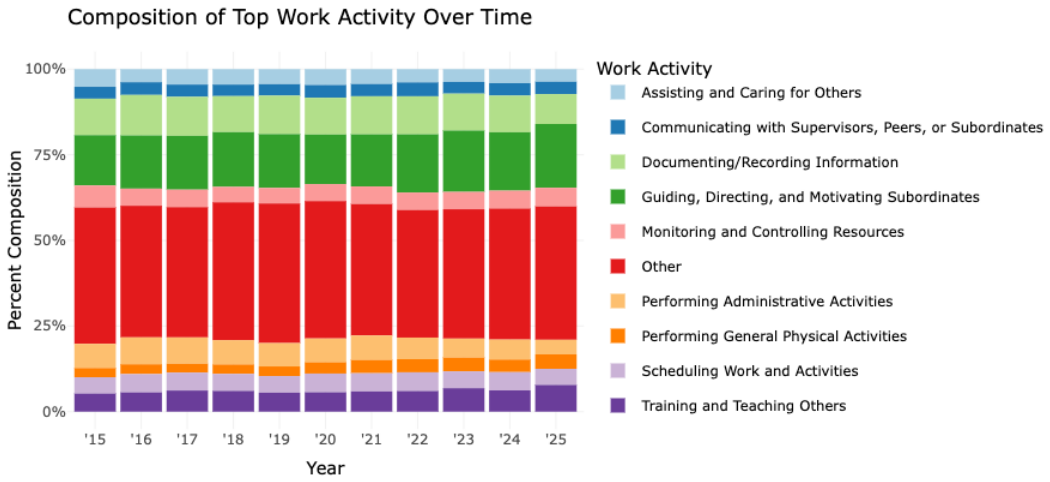



Based on an analysis of NLx Research Hub job ad data. Occupations matched with TitleMatch.

Table D.1: Top 10 Tasks in 2025

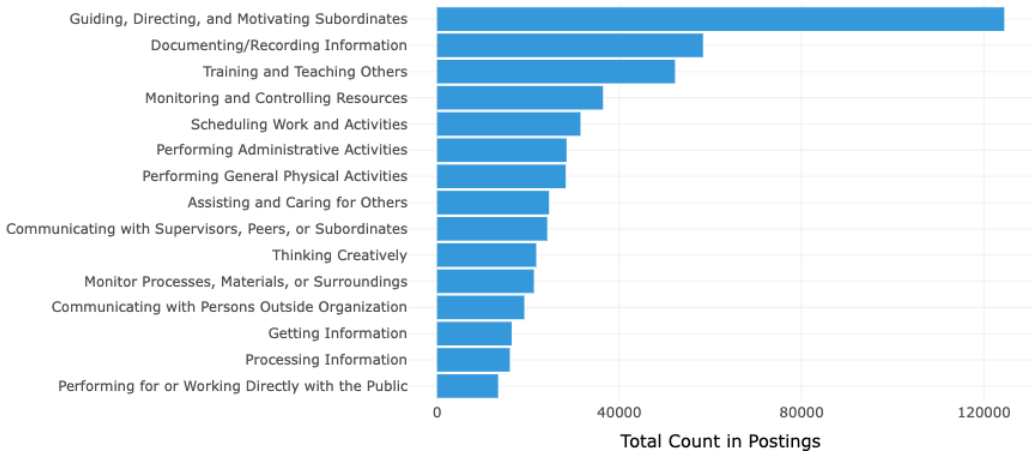
| Rank | Task Description | Value |
|---|---|---|
| 1 | Assign duties to other staff and give instructions regarding work methods and routines. | 7746 |
| 2 | Assign employees to specific duties. | 7537 |
| 3 | Supervise and instruct assistants. | 4706 |
| 4 | Refer guest problems or complaints to supervisors. | 4562 |
| 5 | Assign duties to workers, and schedule shifts. | 4376 |
| 6 | Maintain cleanliness of work area. | 3527 |
| 7 | Teach library patrons basic computer skills, such as searching computerized databases. | 3470 |
| 8 | Adhere to local, state, and federal laws, regulations, and statutes. | 3469 |
| 9 | Oversee training programs. | 3247 |
| 10 | Instruct patients, parents, teachers, or employers in communication strategies to maximize effective receptive communication. | 2964 |

Figure D.2: Change in Work Activity Composition over Time



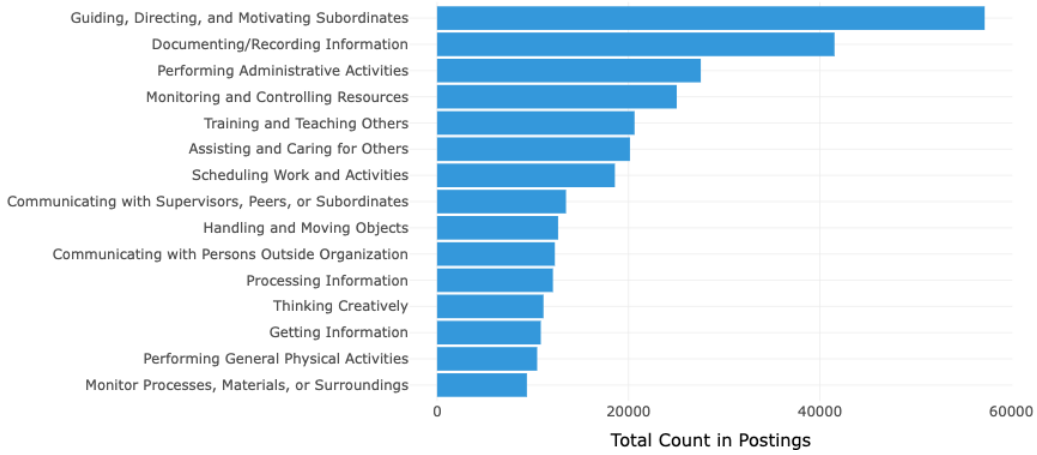

Based on an analysis of tasks extracted with TaskMatch.

Figure D.3: Top Work Activities in 2025



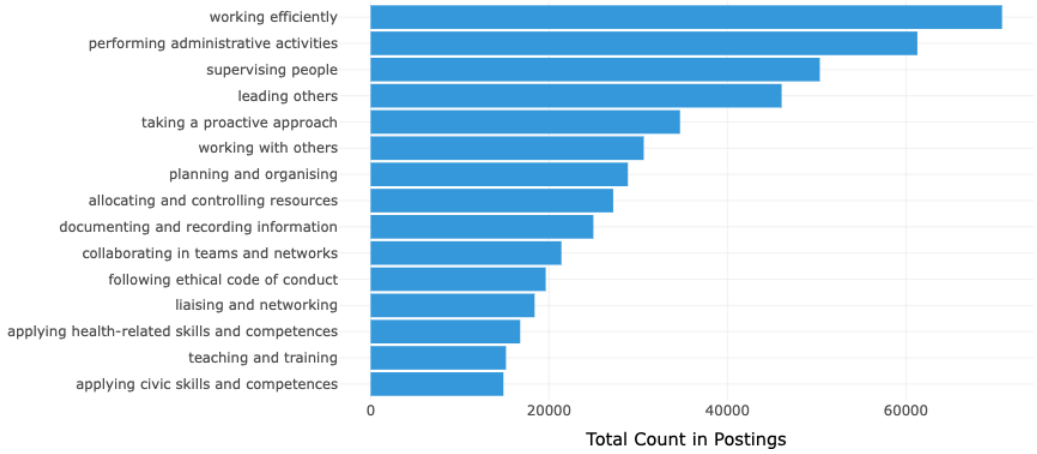

Based on an analysis of tasks extracted with TaskMatch.

Figure D.4: Top Work Activities in 2015



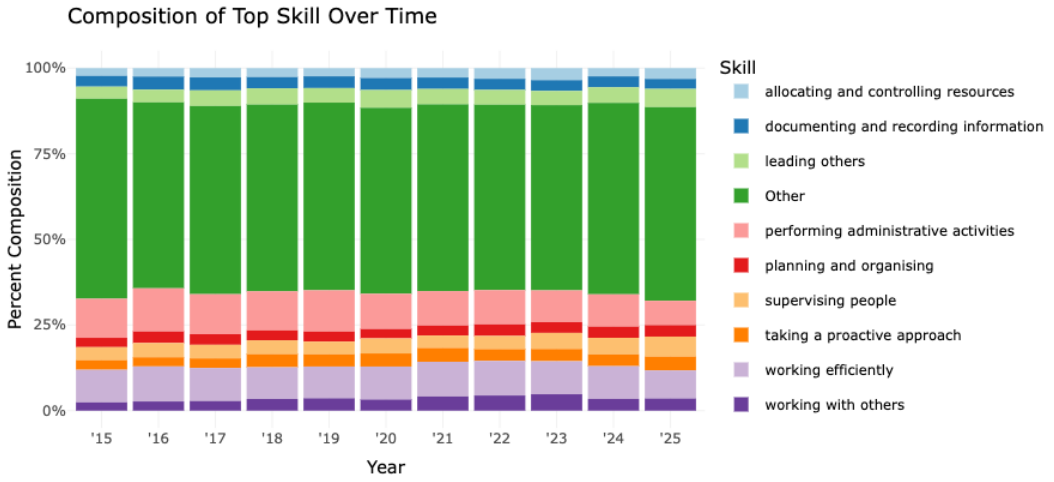

Based on an analysis of tasks extracted with TaskMatch.

Figure D.5: Top Skills in May 2025




Based on an analysis of skills extracted with SkillMatch.

Figure D.6: Change in Skill Composition Over time




Based on an analysis of skills extracted with SkillMatch.

# E Appendix Custom and Standard Dictionaries

This section provides one illustrative term for each label inside a custom dictionary that we build. In addition to the illustrative terms, the full dictionary is available.

## E.1 Features from Job Titles

Table E.1: Custom Dictionary of Features Extracted from Job Titles

| Illustrative Term | Value | Label |
|---|---|---|
| Allowance | All | Allowance |
| Benefits | Bnf | Benefits |
| Bilingual | Bi | Bilingual |
| Bonus | Bns | Bonus |
| Commission | Cm | Commission-Based |
| Contract | Ctr | Contract |
| Contract-to-Hire | C2H | Contract-to-Hire |
| Entry-Level | EL | Entry-Level |
| Experienced | Ex | Experienced |
| Flexible | Flx | Flexible Hours |
| F/T | FT | Full-Time |
| Grad | Grd | Graduate Hiring |
| Holiday | Holiday | Holiday |
| Home-Based | Hm | Home-Based |
| Hourly | hrly | Hourly |
| Hybrid | Hy | Hybrid |
| Immediate | Imm | Immediate Start |
| Incentive | Inc | Incentive |
| Intern | Int | Internship |
| Job Fair | JF | Job Fair |
| Evening | ON | Nights |
| On Call | OC | On Call |
| Onsite | OS | Onsite |
| On-Site | OS | On-Site |
| Overnight | ON | Overnight |
| Paid | Pd | Paid |
| Paid Per | PPS | Paid Per Service |
| P/T | PT | Part-time |
| Peak Time | Peak | Peak Time |
| Per Diem | PD | Per Diem |
| Relocation | Rel | Relocation |
| Remote | Rem | Remote |
| Rotational | Rot | Rotational |
| Salary | Slry | Salary |
| Seasonal | Ssn | Seasonal |
| Security Clearance | Cl | Security Clearance |
| Shift | SW | Shift Work |
| Sign on | SO | Sign-on Bonus |
| Spanish | ES | Spanish language |
| Subcontract | Sub | Subcontract |
| Telecommute | Tc | Telecommute |
| Temporary | Temp | Temporary |
| Training | TP | Training Provided |
| Travel | Trv | Travel Required |
| Undergraduate | Ugrd | Undergraduate |
| Virtual Hiring | VJF | Virtual Job Fair |
| Volunteer | Vol | Volunteer |
| Weekend | WE | Weekend |

## E.2 Custom Dictionary of Benefits

Table E.2: Dictionary of Benefits Features Extracted from Job Ads

| Item | Label | Major Group Label |
|---|---|---|
| % discount | Discount | Additional˙Compensation |
| 401 (k) | 401K | Retirement |
| 401(k) plan with acompany match | 401K Match | Retirement |
| 403 (b) | 403b | Retirement |
| 529 college | collegesavingsplan | Saving |
| Accident and Disability Insurance | Accident | Insurance |
| Adoption Assistance | Family | Benefits |
| Annual equity | Equity | Additional˙Compensation |
| arcade games | arcade | Amenity |
| based on performance | payforperformance | Incentives |
| basketball hoops | basketball | Amentiy |
| BBQ | BBQ | Amenity |
| benefit for public transportation | transit | Benefits |
| BENEFITS Health Insurance | Health | Insurance |
| Benefits Information | Benefits | Benefits |
| career coaches | Coaching | Education |
| catered breakfast | breakfasts | Amenity |
| catered lunches | lunches | Food |
| company match | matching contribution | Additional˙Compensation |
| Company Vehicle | Auto | Amenity |
| competitive base pay | MarketPay | Pay |
| comprehensive training | Training | Education |
| Computer and Cell Phone | Computer/Cell | Expenses |
| Continuing Education Programs | Continuing Education | Education |
| deferred compensation | deferredcomp | Saving |
| dental and vision insurance | dental and Vision | Insurance |
| Dependent Care | FSA | Tax Shelter |
| Detention Pay | Special Pay | Additional˙Compensation |
| differential is paid | pay˙differential | Additional˙Compensation |
| Disability and Life Insurance | DisabilityLife | Insurance |
| Discounted tuition benefits | Tuition Discount | Education |
| Employee Assistance Program (EAP)* | EAP | Benefits |
| Employee Stock Ownership Plan | ESOP | Additional˙Compensation |
| Employee Stock Purchase Plan (ESPP) | ESPP | Additional˙Compensation |
| equity plan | Equity | Compensation |
| FMLA | Disability | Unpaid Leave |
| Full-time and Part-time benefits | Benefits | Benfits |
| Fun Fridays | Fun Fridays | Amenity |
| Group Auto Insurance | Auto | Insurance |
| Group Home Insurance | Home | Insurance |
| Identity Theft Protection | Theft | Benfits |
| income protection | Income | Insurance |
| Legal Assistance | Assistance | Legal |
| Life Insurance | Life | Insurance |
| Medical and Dental Insurance | HealthandDental | Insurance |
| military leave | Military Leave | Job Protected Leave |
| Onboarding Bonus | SignonBonus | Bonuses |
| opportunity for growth | Growth | Career |
| paid Dental | dental | Insurance |
| Paid Holidays | Holidays | Vacation |
| paid leave | PTO | Paid Leave |
| Paid parental leave | parental leave | Job Protected Leave |
| pension | pension | Retirement |
| Pet Insurance | Pet | Insurance |
| ping pong | ping pong | Amenity |
| prescription drug | Rx | Health |
| Relocation | Relocation | Expenses |
| retirement | retirement | Retirement |
| retirement savings plan | retirement savings plan | Retirement |
| savings | savings | Saving |
| Series A stage startup | Startup | Startup |
| short/long term disability | Disability | Insurance |
| sick leave | sick leave | Vacation |
| soccer | soccer | Recreational |
| stocked micro kitchens | kitchen | Amenity |
| student loan repayment | college loan repayment | Education |
| tax benefit | tax benefit | Tax Shelter |
| tax-free benefit for public transportation or parking expenses | transit | Tax Shelter |
| Tuition Assistance | Tuition Reimbursement | Education |
| Vacation | Vacation | Paid Leave |
| Vision and Life Insurance | Vision and Life | Insurance |
| wage will increase | progression˙in˙job | payprogression |
| wage will increase through training | training˙progression | payprogression |
| Week Home | Time Off Benefit | Vacation |
| workplace culture | culture | Culture |

## E.3 Custom Dictionary of Education

Table E.3: Dictionary of Education Features Extracted from Job Ads

| rule | Label | Description |
|---|---|---|
| qualifications high school | HS | High School or GED |
| high school degree ged | HSGED | High School or GED |
| ged equivalent | GED | High School or GED |
| associates degree | Associates | Associates |
| accredited college or university | Bachelors | College or University |
| bachelors masters | Masters | Postgraduate |
| bachelors masters or phd | Doctorate | Doctorate |

## E.4 Custom Dictionary of Drug, Background, and Criminal Checks

Table E.4: Dictionary of Background, Drug, and Criminal Check Features Extracted from Job Ads

| feature | value | label |
|---|---|---|
| background | background | background |
| successfully complete a background | backgroundcheck | background |
| asked questions regarding any felony | criminal noask | background |
| criminal conviction | criminalbackground | background |
| maintaining a satisfactory criminal and credit record | criminalbackground | ongoing |
| drug | drug | drug |
| no drug or alcohol related conviction | drugbackground | background |
| drug free workplace | drugfree | drug |
| pre employment drug | drugtest | drug |
| fair chance employer | Fair Chance Emp[loyer | Law |
| fair chance hiring | Fair Chance Employer | Law |
| Fair Chance Ordinance | Fair Chance Law | will˙hire |
| Criminal history will not automatically disqualify | Fair Chance Law | Law |
| dwi convictions | no˙dwi | no˙dwi |
| not be obligated to disclose sealed or expunged records of conviction | nobackgroundcriminal | nobackground |
| Applicants with a felony conviction or pending/unresolved court cases may not qualify for licenses in all required states. | nohire˙crime | will˙not˙hire |
| offer pre employment | preemployment | preemployment |
| screening | screening | screening |
| Second Chance Act | Second Chance Act | Law |
| recognizes the value in second chances | Second Chance Employer | will˙hire |
| test | test | test |
| second chance | will˙hire˙formerly˙incarcerated | will˙hire |

## E.5 Custom Dictionary of Shift Features

Table E.5: Dictionary of Shift Features Extracted from Job Ads

| Phrase | Category | Specific Category |
|---|---|---|
| Activity Schedule | Activity schedule | |
| Friday schedule | Weekly Schedule | |
| daily schedule | Daily Schedule | |
| Flexibility as to the hours and schedule of work | Flexible Schedule | Flexible for Employer |
| Flexibility as to the hours | Flexible Hours | Flexible for Employer |
| flexible scheduling | Flexible Schedule | |
| schedule allows | Flexible Schedule | Flexible for Employee |
| Special shift | Specific Shifts | Flexible for Employer |
| 40 hour | Predictable Hours | Full Time |
| holiday schedule | Holiday Shifts | |
| Hybrid Schedule | Hybrid Work Schedule | |
| 3rd Shift | Night Shift | |
| night work | Night Shift | |
| Regular Schedule | Predictable Schedule | |
| overtime | Overtime | |
| project schedule | Project Schedule | |
| rotating schedule | Rotational Schedule | |
| Retention Schedule for a minimum of three years | Salary Schedule | |
| Schedule | Schedule | |
| schedule appointments | Schedule Others | |
| schedule of appointments | Schedule others | Responsibility |
| Shift | Shift | |
| 12 Hour Shift | Specific Shift | Long |
| Schedule Per Diem | Shift Pay | |
| Shift Day | Specific shift | Day |
| schedule including early morning | Specific Shift | First Shift |
| Schedule Type | Specific Shfit | |
| during a shift | Specific Shift | |
| end of the day shift | Specific Shift | Second Shift |
| 2nd Shift | Specific Shift | Second Shift |
| end of shift | Specific Shift | |
| Entry Level Shift | Specific Shift | Seniority based shifts |
| Overnight shift | Specific Shift | Third Shift |
| supervision of a shift | Supervise Shfits | |
| Shift Lead | Supervise Shifts | |
| temp to hire position | Temp to Hire | |
| Weekend Shift | Weekend Shift | |
| weekly duties and schedule | Predictable Schedule | Weekly |
| shift pay | Shift Pay | Additional˙Compensation |
| closed most holidays | No Holidays | Scheduling |

### E.6 Standard Dictionaries from the Management Literature

Standard dictionaries from the management literature run on the job ads include the following:

- Recruiting signals (Banks et al., 2018): Achievement Oriented Culture, Creative Culture, Development Early, Development, Inclusive Military, Inclusiveness, Corporate Social Responsibility, Team Based Culture, Values Based Culture, Well Being.
- Long Term Orientation (Brigham et al., 2014): Continuity, Futurity, Perseverance.
- Time Horizon Words: Long term, Neutral, Short (Brochet, Loumioti and Serafeim, 2015); Long, Short (DesJardine and Bansal, 2019).
- Training Feedback: Utility, Satisfaction Glerum et al. In Press
- Management Innovation (Heyden, Sidhu and Volberda, 2015): Change, Practices, Processes, Structures.
- Digital Orientation (Kindermann et al., 2021): All dimensions, Architecture configuration, Capabilities, Digital ecosystem coordination, Digital technology scope.
- CreativitySpeak (Haselhuhn, Wong and Ormiston, 2022).
- TMT Attention (Levy, 2005): Board of Directors, Competitors, Computer Ind. Customer, Healthcare Ind. Semiconductor Ind. Dealers, Employees, Owners, Strategic Partners, Management. Foreign Operations and Markets, including: Africa, Asia Pacific, Europe, Latin America, Middle East, North America.
- Ambidexterity (McKenny et al., 2018*a*): Exploitation, Exploration.
- Entrepreneurial Orientation (McKenny et al., 2018*b*): Autonomy, Aggressiveness, Inductive, Innovativeness, Proactiveness, Risk Taking, Competitor Orientation, Customer Orientation, Interfunctional Coordination, Long Term Focus, Profitability. (Short et al., 2009): Autonomy, Aggressiveness, Inductive, Innovativeness, Proactiveness, Risk Taking
- Value Orientation (Moss et al., 2018): Economic Value, Social Value, Environmental Value
- Task Environment (Olsen, Sofka and Grimpe, 2016): Stakeholder.
- Corporate Social Responsibility (CSR) (Pencle and Mălăescu, 2016): Employee, Environment, Human Rights, Social and Community
- Sustainability Orientation (Vaupel et al., 2023): Addition, Environmental, Social

- Market Orientation (Zachary et al., 2011): Customer Orientation, Competitor Orientation, Interfunctional Coordination, Long Term Focus, Profitability.