# Extracting Structured Labor Market Information from Job Postings with Generative AI

MARK HOWISON, Amazon.com Inc, Seattle, United States
WILLIAM O. ENSOR, Amazon.com Inc, Seattle, United States
SURAJ MAHARJAN, Amazon.com Inc, Seattle, United States
RAHIL PARIKH, Amazon.com Inc, Seattle, United States
SRINIVASAN H. SENGAMEDU, Amazon.com Inc, Seattle, United States
PAUL DANIELS, National Association of State Workforce Agencies, Washington, United States
AMBER GAITHER, National Association of State Workforce Agencies, Washington, United States
CARRIE YEATS, National Association of State Workforce Agencies, Washington, United States
CHANDAN K. REDDY, Virginia Polytechnic Institute and State University, Blacksburg, United States and Amazon.com Inc, Seattle, United States
JUSTINE S. HASTINGS, University of Washington, Seattle, United States and Amazon.com Inc, Seattle, United States

Labor market information is an important input to labor, workforce, education, and macroeconomic policy. However, granular and real-time data on labor market trends are lacking; publicly available data from survey samples are released with significant lags and miss critical information such as skills and benefits. We use generative Artificial Intelligence to automatically extract structured labor market information from unstructured online job postings for the entire U.S. labor market. To demonstrate our methodology, we construct a sample of 6,800 job postings stratified by 68 major occupational groups, extract structured information on educational requirements, remote-work flexibility, full-time availability, and benefits, and show how these job characteristics vary across occupations. As a validation, we compare frequencies of educational requirements by occupation from our sample to survey data and find no statistically significant difference. Finally, we discuss the scalability to collections of millions of job postings. Our results establish the feasibility of measuring labor market trends at scale from online job postings thanks to advances in generative AI techniques. Improved access to such insights at scale and in real-time could transform the ability of policy leaders, including federal and state agencies and education providers, to make data-informed decisions that better support the American workforce.

Authors' Contact Information: Mark Howison (Corresponding author), Amazon.com Inc, Seattle, Washington, United States; e-mail: mhowison@amazon.com; William O. Ensor, Amazon.com Inc, Seattle, Washington, United States; e-mail: ensorw@amazon.com; Suraj Maharjan, Amazon.com Inc, Seattle, Washington, United States; e-mail: mhjsuraj@amazon.com; Rahil Parikh, Amazon.com Inc, Seattle, Washington, United States; e-mail: parrahil@amazon.com; Srinivasan H. Sengamedu, Amazon.com Inc, Seattle, Washington, United States; e-mail: sengamed@amazon.com; Paul Daniels, National Association of State Workforce Agencies, Washington, District of Columbia, United States; e-mail: pdaniels@naswa.org; Amber Gaither, National Association of State Workforce Agencies, Washington, District of Columbia, United States; e-mail: agaither@naswa.org; Carrie Yeats, National Association of State Workforce Agencies, Washington, District of Columbia, United States; e-mail: cyeats@naswa.org; Chandan K. Reddy, Virginia Polytechnic Institute and State University, Blacksburg, Virginia, United States and Amazon.com Inc, Seattle, Washington, United States; e-mail: ckreddy@amazon.com; Justine S. Hastings, University of Washington, Seattle, Washington, United States and Amazon.com Inc, Seattle, Washington, United States; e-mail: jhastin@uw.edu.

ACM 2639-0175/2025/02-ART9
https://doi.org/10.1145/3674847

CCS Concepts: • **Applied computing** → **Economics;** • **Computing methodologies** → **Information extraction**;

Additional Key Words and Phrases: Workforce, education, policy, large language models, Amazon Bedrock

**ACM Reference Format:**
Mark Howison, William O. Ensor, Suraj Maharjan, Rahil Parikh, Srinivasan H. Sengamedu, Paul Daniels, Amber Gaither, Carrie Yeats, Chandan K. Reddy, and Justine S. Hastings. 2025. Extracting Structured Labor Market Information from Job Postings with Generative AI. *Digit. Gov. Res. Pract.* 6, 1, Article 9 (February 2025), 12 pages. https://doi.org/10.1145/3674847

## 1 Introduction

Technology continues to transform the labor market [1]. Policy leaders, workforce agencies, and education innovators in the U.S. need clear data and analysis to make the best investments in America's workforce to ensure they are skilled for today's jobs and well-prepared for the jobs of the future [2]. COVID further transformed the U.S. labor market, rapidly changing what skills are in demand [3] and even changing where and how Americans work [4].

Online job postings are a rich and novel source of real-time data on trends in the labor market that can help policy leaders plan investments in labor training and education. They provide a daily snapshot of employers' current hiring needs across the many employers and industries that advertise job openings online. Currently, it is challenging to use these data for labor market analysis, because specific details such as the qualifications that employers are seeking or the benefits they offer are embedded in unstructured text in the job posting.

Efforts to structure information from job postings have focused on understanding occupational structure, in particular the **Standard Occupational Classification (SOC)** code that is most likely associated with a job title or job description [5]. These approaches typically rely on human-annotated training datasets [6–8], which are challenging to scale. In addition to SOC codes, more recent studies have focused on identifying skills [9] and wages/salaries [10] in job postings. Additional details such as benefits, educational requirements, and remote-work flexibility are not as well studied yet are important for labor policy decisions. Moreover, by focusing on current taxonomies and ontologies based on occupational codes, analyses may miss the nuances that exist in the text of job postings, which represent real and direct employer demand. These nuances often point to new and emerging occupations and skills that standard taxonomies are slower to recognize.

Insights and trends from structured job postings could help policy, education, and industry leaders understand labor demand dynamics and make the right investments in skills and training for American workers to stay ahead of the curve. Examples include the following:

Data collection efforts by **federal agencies** could benefit from the larger sample and scale of online job postings compared to current survey approaches. Today, federal data on job openings come primarily from the **Job Openings and Labor Turnover Survey (JOLTS)**, which reports the number of vacancies at the end of each month, relying on a sample size of 21,000 (out of 8 million) establishments with a response rate of 30% [11]. JOLTS is released after a two-month delay and does not include information on wages or benefits. Accurate measures of the number of job openings and the wages posted in job openings are a leading indicator of labor market conditions and wage pressures [12]. For example, a scalable data resource built using our methods could help the Federal Reserve make more timely decisions about monetary policy through improved predictions of labor market conditions and labor costs. The granularity of job posting data could also help the Federal Reserve measure progress toward its maximum employment target.

**State and local governments** and **education providers** could benefit from measuring skills demand by geographic region, occupational cluster, industry, and time period. Gaps in demand and existing skills availability in regional workforces can guide investments in post-secondary education and reskilling and training programs

[13–16]. In addition, the ability to perform gap analysis between skills demand and skills development could help inform and refine the development of academic curriculums, apprenticeship opportunities, and non-degree credentials to better build the American workforce for the ever-evolving labor market. Similarly, local and state-provided job search assistance, especially in conjunction with the Unemployment Insurance program, could be better tailored to local labor market conditions [17]. Understanding the evolving dynamics of remote versus in-office work could help policy makers better target infrastructure projects, such as broadband internet access [18], to support the future of work [19].

**Employers** could benefit by applying scalable and comprehensive labor information to benchmarking their job openings within geographic regions. Benchmarking could help employers remain competitive and improve job offerings to attract workers more effectively in a tight labor market.

**Academic researchers** could benefit from access to structured data to study dynamics in skills demand, wages, and benefits by region, and the supply of remote vs. in-office work. Findings from such studies will benefit general knowledge of the labor market and the impacts of technological change and population dynamics, such as aging and declining workforce growth [20]. Access to this type of structured data would also allow for the creation of updated occupational taxonomies and ontologies that would provide a richer and more dynamic understanding of labor market trends.

The challenge with extracting structured information in a consistent and reliable way is the variation in language, organization, formatting, abbreviations, and conventions used in online job postings. Recent advances in **generative Artificial Intelligence (GenAI)**, however, have produced commercially available large language models that excel at summarizing this kind of variation in unstructured text and can follow specific instructions for output. We test these capabilities on a sample of job postings, demonstrate that comprehensive information can be extracted automatically and reliably, and discuss the next challenge of scaling these methods to deliver labor market insights for the broader U.S. labor market, both historically and in real-time.

## 2 Data and Methods

We analyzed publicly available job postings from the National Labor Exchange that were posted online between March and November 2023. We collected the postings in their original HTML format and extracted the job description text from the body using the BeautifulSoup library in Python. Additionally, we replaced new lines with spaces and removed repeated spaces from the job posting text. We did not perform any normalization beyond this, since additional variation in structure and content, such as misspellings, will be handled by the foundation model.

We constructed a stratified sample of 6,800 job postings by uniformly sampling 100 distinct and valid job postings for each of 68 "Minor Group" occupational codes from the 2018 **Standard Occupational Classification (SOC)** System [21]. Military-related "Minor Group" designations, "Fishing and Hunting Workers" (45–3000), and special groups for "supervisor" or "other" types of occupations were omitted because of low coverage. Valid job postings were defined as those having between 250 to 8,000 tokens, as determined by the "cl100k_base" encoding in the Python package tiktoken. These criteria eliminate short postings that have expired or contain error codes and long postings that contain extraneous information. Overall, 62.3% of the job postings considered as part of the sampling process met these filtering criteria.

Using the prompt shown in Procedure 1, we extracted structured information in JSON format from the raw unstructured job posting text. We queried each of the 6,800 sampled job postings using this prompt in Amazon Bedrock, which is a managed service we used to access GenAI foundation models. We called the foundation model with default settings except for temperature, which tunes the amount of randomness in generated output and which we lowered from 1.0 to 0.2 to reduce randomness. We concatenated the resulting JSON output into an analysis dataset (see Figure 1).

Large language models are well-suited to parsing unstructured data when queried with appropriate prompts. In designing our prompt, we followed existing best practices for specifying instructions and ensuring structured,

**(a)** Example job posting

**Software Engineering Manager - Compiler Technologies**

The documentation engineering is responsible for creating world-class developer tools and framework-level support for the documentation workflows...

**Key Qualification**

5+ years of professional software engineering experience
Experience shipping high quality user facing products and features
Experience with: Swift, compilers, macOS applications, or similar technologies
Experience fostering a collaborative and inclusive experience in an open source environment
Ability to understand problems and designs at a very low level
Ability to communicate with cross-functional teams about various engineering...

**Education & Experience**

BS/MS in Computer Science, Computer Engineering, Informatics (or equivalent experience)

**Pay & Benefits**

Base pay is one part of our total compensation package and is determined within a range. This provides the opportunity to progress as you grow and develop within a role. The base pay range for this role is between $173,000 and $299,000, and your base pay will depend on your skills, qualifications, experience, and location.

You'll also receive benefits including: Comprehensive medical and dental coverage, retirement benefits, a range of discounted products and free services, and for formal education related to advancing your career at [redacted], reimbursement for certain educational expenses, including tuition. Additionally, this role might be eligible for discretionary bonuses or commission payments as well as relocation.

[Redacted] is an equal opportunity employer that is committed to inclusion and diversity...

**(b)** Example of extracted JSON information


```
{
  "wage_range": null,
  "wage_allin": false,
  "wage_upto": false,
  "salary_range": [
    173000,
    299000
  ],
  "full_time": true,
  "part_time": false,
  "remote": false,
  "remote_str": null,
  "visa": false,
  "visa_str": null,
  "benefits": true,
  "benefits_str": "You'll also receive benefits including:
    Comprehensive medical and dental coverage, retirement
    benefits, a range of discounted products and free services, and
    for formal education related to advancing your career at
    [redacted], reimbursement for certain educational expenses,
    including tuition. Additionally, this role might be eligible for
    discretionary bonuses or commission payments as well as
    relocation.",
  "career": true,
  "career_str": "formal education related to advancing your career at
    [redacted], reimbursement for certain educational expenses,
    including tuition",
  "scheduling": false,
  "scheduling_str": null,
  "licensing": false,
  "licensing_str": null,
  "requirements": [
    {
      "education": "Bachelors degree",
      "major": "Computer Science, Computer Engineering,
        Informatics",
      "experience": 5.0
    }
  ]
}
```


Fig. 1. (a) An example job posting and (b) the structured information extracted by our method.

machine-readable output [22]. In particular, we started with a role prompting statement, "Your task is to read a job post and return relevant information." The prompt is then templated to insert the unstructured text of the job posting, and the parsing instructions are placed after the job posting text. To ensure machine-readable output, the instructions provide a JSON-formatted output template with predefined keys and value definitions, and the instructions, "Match the schema below exactly. Do not include a preamble in your response." This prompt design lies between a zero-shot and one-shot learning approach, as the output JSON structure is fully defined but no parsed values are provided as examples.

We performed descriptive analysis on the extracted information. As a validation, we tested whether the information on education level by occupation extracted from our sample differed from that reported in survey data from the U.S. Census Bureau's **Current Population Survey - Annual Social and Economic Supplement (CPS ASEC)** using a two-sided, paired Wilcoxon test.

Finally, we estimated the association between education level and remote-work availability, controlling for other job posting characteristics, using a logistic regression with specification:

$$r_i = \alpha + \beta e_i + \gamma b_i + \delta f_i + \omega O_i + \epsilon,$$

**PROCEDURE 1**: Generative AI Prompt for JSON-structured Data from Job Postings

```
Human: Your task is to read a job post and return relevant information. The job post is inside <text></text>XML tags.
<text>
{{text}}
</text>
Based on the job post, return the following information as a JSON blob.Match the schema below exactly. Do not include
a preamble in your response.
‘‘‘json
{
    ‘wage_range’: <>What is the wage range offered? Return a List of Float values if range. If single value, return
List with single Float. Otherwise, return Null.</>
    ‘wage_allin’: <>Does the wage include overtime or premium payments? If yes, True. Otherwise False.</>
    ‘wage_upto’: <>Is the wage described as up to the specified amount? For example, "up to $20." If yes, return True.
Otherwise False.</>
    ‘salary_range’: <>What is the annual salary range offered? Return as List of Float values if range. If single
value, return List with single Float. Otherwise, return Null.</>
    ‘full_time’ : <>Bool, True if full-time job, otherwise False</>
    ‘part_time’ : <>Bool, True if part-time job, otherwise False</>
    ‘remote’ : <>Does the job post offer remote work? If yes, True. Otherwise False.</>,
    ‘remote_str’ : <>If ’remote’ is True summarize the remote work policy as described in job post. Otherwise Null.</>
    ‘visa’ : <>Does the job post support visa sponsorship? If yes, True. Otherwise False.</>,
    ‘visa_str’ : <>If ’visa’ is True summarize what the job post says about visa sponsorship. OtherwiseNull.</>
    ‘benefits’ : <>Does the job post offer benefits? If yes, True. Otherwise False.</>
    ‘benefits_str’ : <>If ’benefits’ is True provide a list of the benefits in the job post. Otherwise Null.</>
    ‘career’ : <>Does the job post offer education, training, or professional development opportunities? If yes, True.
Otherwise False.</>
    ‘career_str’ : <>If ‘career’ is True summarize the education, training, or professional development opportunities.
Otherwise Null.</>
    ‘scheduling’ : <>Does the job post describe a particular shift or hours to work? If yes, True. Otherwise False.</>
    ‘scheduling_str’ : <>If ’scheduling’ is True summarize the shift or hours worked by this job. Otherwise Null.</>
    ‘licensing’ : <>Does the job post require licensing, such as a Commercial Drivers License or other licensing
requirement? If yes, True. Otherwise, False.</>
    ‘licensing_str’ : <>If ’licensing’ is True, summarize the licensing requirements in the job post. Otherwise Null.</>
    ‘requirements’ : <>List, acceptable combinations of education, major, and experience, according to job post. For
each combination, create a dictionary:
       {
           ‘education’: <education>String, one of these: [’No education’, ’High school or equivalent’, ’Associates
degree’, ’Bachelors degree’, ’Masters degree’, ’PhD’, ’MBA’, ’JD’]</education>,
           ‘major’: <major>String, major or area of study</major>,
           ‘experience’: <experience>Float, years of work experience</experience>
       }</>
}
‘‘‘
Assistant:
```

where $r_i$ is an indicator for whether job posting $i$ advertises remote-work availability, $e_i$ is an indicator for whether the job requires college or post-graduate education, $b_i$ is an indicator for whether the posting advertises benefits, $f_i$ is an indicator for whether the posting advertises full-time availability, $O_i$ are dummy variables for “Minor Group” occupation, and $\epsilon$ is an error term. The parameter of interest is $\beta$.

## 3 Results

Our descriptive analysis finds that the foundation models we tested from Amazon Bedrock can extract comprehensive structured information from job postings using the prompting strategy described above. First, we examined the frequency of remote-work and full-time availability in each “Minor Group” occupation (Figure 2). As expected, occupations that typically require physical presence at a work site, such as Moving (53–7,000), Food Preparation (35–2,000), Building Cleaning (37–2,000), and Grounds Maintenance (37–3,000), had no job postings in our sample that indicated remote-work availability. The occupations with the highest frequency of remote availability (at approximately 50% of postings) were Marketing/PR/Sales Management (11–2,000), Anal-

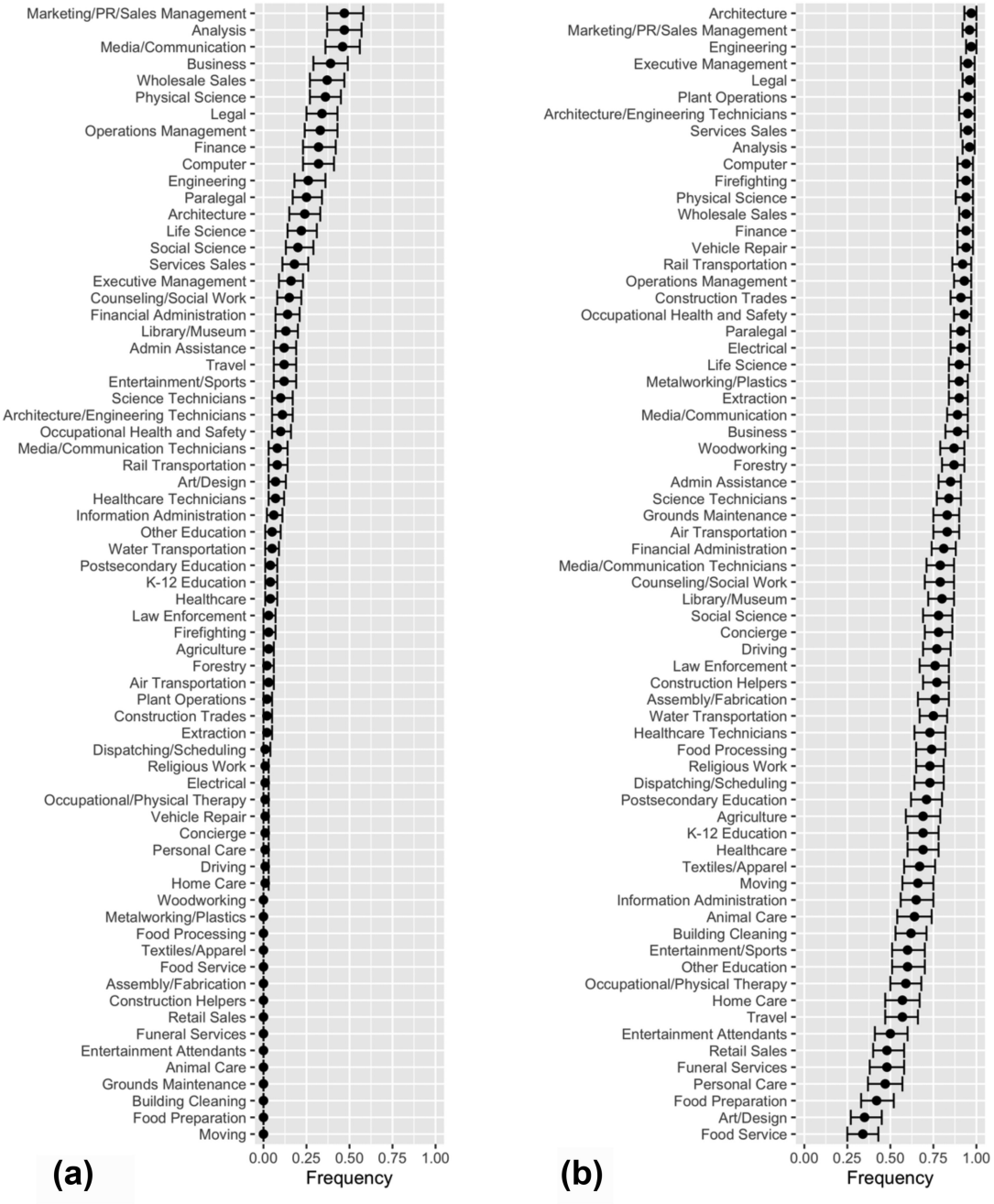


Fig. 2. The frequency of job postings in our sample that indicate remote-work availability (a) or full-time availability (b), by occupation. Error bars show bootstrapped 95% confidence intervals for 1,000 replicates.

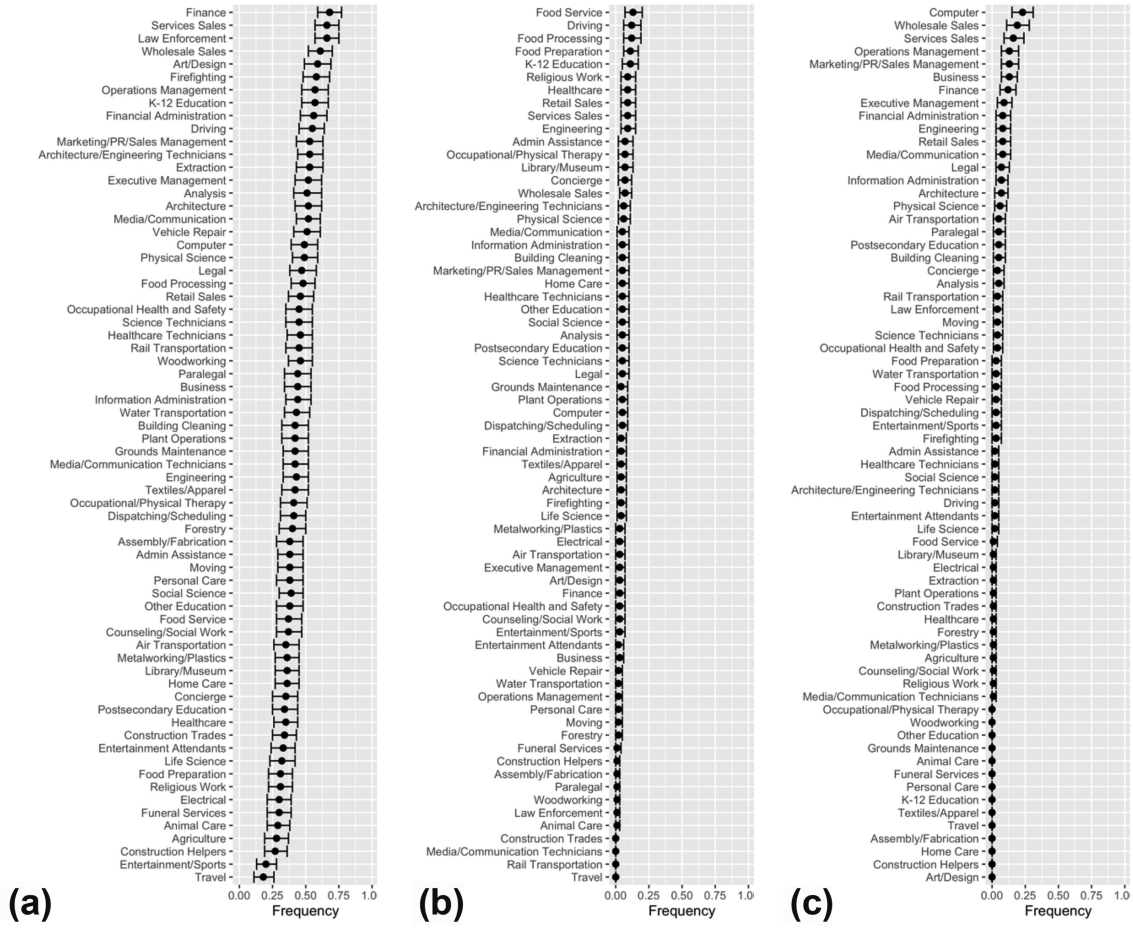


Fig. 3. The frequency of job postings in our sample that advertise benefits for retirement contributions (a), tuition reimbursement (b), or paid parental leave (c), by occupation. Error bars show bootstrapped 95% confidence intervals for 1,000 replicates.

ysis (15–2,000), and Media/Communication (27–3,000). Full-time availability was most frequent in Architecture (17–1,000), Legal (23–1,000), and Engineering (17–2,000) occupations and least frequent in Food Service (35–3,000), Art/Design (27–1,000), and Food Preparation (35–2,000).

Next, we examined the frequency of advertised benefits for retirement contributions, tuition reimbursement, and paid parental leave (Figure 3). Retirement benefits were the most commonly advertised across occupations, with the highest frequency in Finance (13–2,000) occupations and the least in Travel (39–7,000). Tuition reimbursement and paid parental leave benefits were less frequently advertised, less than 25% of the time even for the most frequent occupations.

To validate the accuracy of our frequency estimates against currently available data, we compared the frequency of each education level by occupation to the CPS ASEC (Figure 4). The results are visually similar, and to formalize the comparison, we tested and rejected the hypothesis that the samples come from different distributions (Wilcoxon test, $p = 0.13$).

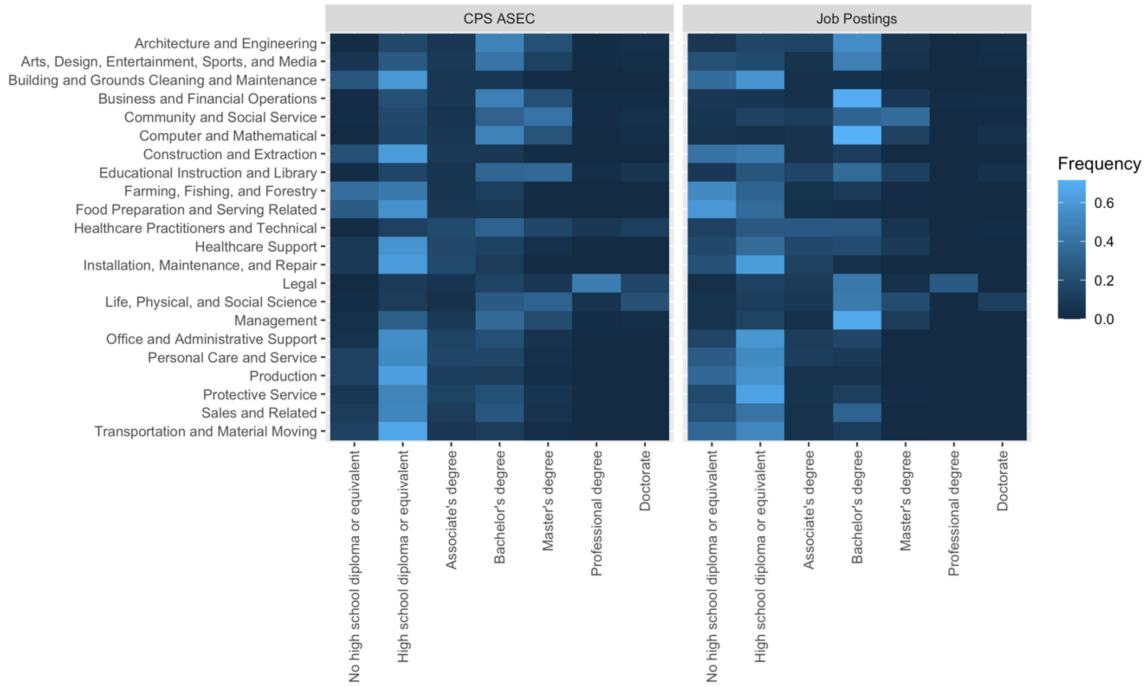


Fig. 4. The frequency of required education level extracted from our sample of job postings vs. the frequency reported in survey data from the U.S. Census Bureau's Current Population Survey - Annual Social and Economic Supplement (CPS ASEC), by major occupation group.

Finally, we estimated an odds ratio of 4.2 (1.429, se = 0.150, p = 0.000, odds ratio 95% C.I.: 3.1–5.6) for the association between job postings requiring at least a college education and advertising remote-work availability, after controlling for benefits, full-time availability, and occupation (full regression results are available in Table 1). In other words, job postings requiring higher education were 4.2 times more likely to advertise remote-work availability than those requiring less than a college education. Robustness checks that omit controls for occupation have higher estimates for the association (see Table 1), suggesting that educational requirements and remote-work availability are mediated by occupation, which is also supported by our descriptive analysis above.

## 4 Discussion

With GenAI, we are able to extract additional information from job postings, including educational requirements, remote-work flexibility, full-time availability, and benefits, compared to previous approaches that focused on occupations, skills, and wages/salaries [5–10]. Unlike in another recent study that found wage/salary information in job postings is skewed relative to survey data [23], our study found that education requirements extracted from job postings agree with survey data.

Another benefit of our GenAI-based approach is its improved scalability compared to existing approaches that require human-annotated training datasets to model occupational information in job postings [6–8] and to survey approaches for labor market information [11]. Our study required no human annotation and only the creation of a single GenAI prompt that can be re-used for any number of job postings. Our analysis of the resulting structured data shows that human annotation is not required to obtain accurate measures of labor market information that are on par with current surveys.

Table 1. Logistic Regression Table for Estimates of the Association between Higher Education Requirements (a Bachelor's Degree or Higher) and Remote-Work Availability (the Dependent Variable), with Heteroskedasticity-consistent Standard Errors

| | (1) | (2) | (3) |
|---|---|---|---|
| Intercept | −3.307*** (0.104) | −4.110*** (0.227) | −2.804*** (0.367) |
| higher_education | 2.253*** (0.115) | 2.159*** (0.113) | 1.429*** (0.150) |
| benefits | | −0.076 (0.125) | −0.226 (0.138) |
| full_time | | 1.029*** (0.194) | 0.601*** (0.206) |
| occupation: Agriculture | | | −1.227* (0.650) |
| occupation: Air Transportation | | | −1.770*** (0.673) |
| occupation: Analysis | | | 0.934** (0.390) |
| occupation: Architecture/Engineering Technicians | | | −0.408 (0.457) |
| occupation: Architecture | | | 0.129 (0.399) |
| occupation: Art/Design | | | −0.157 (0.517) |
| occupation: Business | | | 0.789** (0.386) |
| occupation: Computer | | | 0.350 (0.399) |
| occupation: Concierge | | | −2.007* (1.057) |
| occupation: Construction Trades | | | −1.498* (0.774) |
| occupation: Counseling/Social Work | | | −0.254 (0.428) |
| occupation: Dispatching/Scheduling | | | −2.187** (1.039) |
| occupation: Driving | | | −2.071** (1.054) |
| occupation: Electrical | | | −2.387** (1.059) |
| occupation: Engineering | | | 0.007 (0.402) |
| occupation: Entertainment/Sports | | | −0.612 (0.459) |
| occupation: Executive Management | | | −0.234 (0.420) |
| occupation: Extraction | | | −1.532* (0.791) |
| occupation: Finance | | | 0.419 (0.392) |
| occupation: Financial Administration | | | 0.003 (0.452) |
| occupation: Firefighting | | | −1.878** (0.787) |
| occupation: Forestry | | | −1.554** (0.760) |
| occupation: Healthcare Technicians | | | −0.546 (0.518) |
| occupation: Healthcare | | | −1.669*** (0.603) |
| occupation: Home Care | | | −2.095** (1.061) |
| occupation: Information Administration | | | −0.338 (0.537) |
| occupation: K-12 Education | | | −1.515** (0.596) |
| occupation: Law Enforcement | | | −1.453** (0.657) |
| occupation: Legal | | | 0.565 (0.398) |
| occupation: Library/Museum | | | −0.437 (0.446) |
| occupation: Life Science | | | −0.082 (0.410) |
| occupation: Marketing/PR/Sales Management | | | 0.968** (0.386) |
| occupation: Media/Communication Technicians | | | −0.706 (0.496) |
| occupation: Media/Communication | | | 0.958** (0.397) |
| occupation: Occupational Health and Safety | | | −0.753 (0.476) |
| occupation: Occupational/Physical Therapy | | | −3.211*** (1.078) |
| occupation: Operations Management | | | 0.393 (0.397) |
| occupation: Other Education | | | −1.303** (0.558) |
| occupation: Paralegal | | | 0.207 (0.410) |
| occupation: Personal Care | | | −1.950* (1.059) |
| occupation: Physical Science | | | 0.479 (0.393) |
| occupation: Plant Operations | | | −1.653** (0.778) |
| occupation: Postsecondary Education | | | −1.855*** (0.606) |
| occupation: Rail Transportation | | | −0.469 (0.486) |
| occupation: Religious Work | | | −3.342*** (1.055) |
| occupation: Science Technicians | | | −0.798* (0.474) |
| occupation: Services Sales | | | 0.001 (0.422) |
| occupation: Social Science | | | −0.145 (0.422) |
| occupation: Travel | | | −1.191* (0.634) |
| occupation: Vehicle Repair | | | −2.158** (1.054) |
| occupation: Water Transportation | | | −0.777 (0.584) |
| occupation: Wholesale Sales | | | 0.911** (0.385) |
| *Note:* | | | *p<0.1; **p<0.05; ***p<0.01 |

We tested a univariate model (1) and multivariate models with controls for benefits and full-time availability (2) and dummy variables for occupation (3).

The scalability of our method is determined by the number of job postings, the average number of input tokens per posting, and the average number of output tokens per posting. Real-time processing is not required for the National Labor Exchange, which is updated in nightly batches, but an automated pipeline would need to process tens of thousands of postings per day to stay up to date. Input tokens could be optimized through additional pre-processing of job description text to remove extraneous text. For example, many U.S. job postings contain an Equal Employment Opportunity statement that could be removed to reduce input tokens. Output tokens are driven by the scope and size of the extracted feature set.

The feature set extracted by our prompt in this study was designed in collaboration with NASWA and feedback from National Labor Exchange Research Hub users to address the highest priority use cases and demonstrate the extraction of features not found in previous methods. However, the prompting strategy we introduced is flexible, and we are exploring additional features, including alternative compensation (sign-on bonuses, moving expenses), years of experience, supervisory or travel requirements, visa sponsorship, and more. The flexibility of the prompt will allow the method to evolve to meet changing needs or priorities for labor market information. As an example, we identified remote-work availability as a priority feature for this study, but it may not have been identified as such four years ago before the COVID pandemic caused widespread changes in the availability of remote work across the labor market. Because GenAI costs generally scale more with the number of output tokens (e.g., the decoding step) than input tokens (e.g., the encoding step), new features can be added incrementally by reprocessing job postings to output only the additional features. The encoded job postings could also be cached to further optimize this incremental approach.

GenAI methods, including the foundation models we used, may have biases originating from the choice of training data, labeling, validation metrics, model fitting, and more. While we are unable to modify the foundation model, we can design our prompting and sampling strategies to mitigate bias and validate them by testing for potential differences in outcomes with external data. For example, we included instructions in our prompt to return only relevant information and to match the output schema exactly. We sampled job postings uniformly by occupation to ensure equal representation, since prior research has shown biases in representation of online job postings by occupation [9]. Finally, we tested for differences in required education levels by occupation and found no bias relative to survey data. This approach of bias mitigation through prompt engineering and sampling and validation against external data is broadly applicable to analyses of structured output from foundation models.

Another set of potential concerns surrounding GenAI methods involve the privacy of individuals' data processed by the method. In this study, we do not analyze any personally identifiable information or employer identifiers, and there are no privacy concerns with processing job descriptions, because they were created with the intent of public distribution. By structuring publicly available job descriptions, we avoid the privacy concerns inherent with other labor market data such as individual employment records or candidate resumes.

GenAI methods are potentially subject to misuse. These risks are greatest for methods that generate new content directly from user input in real time, which can be manipulated to achieve unanticipated or undesirable outputs. In our application, the job posting inputs do not come directly from users but indirectly via the National Labor Exchange, which has processes in place to verify every employer who contributes job postings. This reduces the risk of manipulation. The output generated by our method has a known and pre-specified structure, which reduces the risk of unanticipated or undesirable outputs.

Because labor market information extracted by our method could affect policy decisions, it is critical that the information is valid and can be inspected or audited by stakeholders. Currently, labor market information that is available for purchase from commercial providers is produced by closed-source models that cannot be inspected. The lack of transparency and cost of those data limit their availability and extensibility. Our goal is to broaden access to labor market information through open-source methods and publicly available data. In partnership with NASWA, we are convening a user group of stakeholders and policymakers to ensure that such a public resource is continuously improved and responsive to policy priorities.

## 5 Conclusion and Future Work

We have demonstrated a successful prompting strategy for automatically extracting labor market information from online job postings using GenAI techniques. With extracted information from a sample of 6,800 job postings, we compared the frequencies of required education levels by occupation to survey data from the U.S. Census Bureau and found no statistically significant differences. We also estimated that job postings requiring a college education are 4.2 times more likely to advertise remote-work availability.

While the proposed method works well at extracting information from job postings at the small scale of our sample, the next challenge is the cost and runtime limitations of scaling it to hundreds of thousands of job postings that are updated on a daily basis and to millions of historical job postings. One potential solution we are exploring is a knowledge-distillation approach. Using a sample of job postings with high-quality information extracted by our initial analysis, we can train a more efficient large language model or classification model for the specialized task of extracting information from job postings. This approach will incur the one-time cost of creating the training sample with the more expensive foundation model but will lower the ongoing operational costs of applying the distilled model to larger collections of job postings. Scaling our approach to historical job postings will also help address another limitation of our preliminary results, which is the compressed time frame from March to November 2023 that prevents us from analyzing longer time trends in the labor market, such as changes in job characteristics pre- and post-COVID.

In a collaboration with the **National Association of State Workforce Agencies (NASWA)** and DirectEmployers Association, we will be scaling this method to process tens of millions of historical job postings in the National Labor Exchange Research Hub. In addition to the structuring of historic data, the project will explore the feasibility of automating nightly batch processing of the Research Hub's database, which is refreshed every 24 hours to provide analysts with up-to-date data on labor demand. Publicly available labor market information resulting from this collaboration will empower NASWA's member agencies and a broader coalition of employers, partners, and researchers to use data-driven approaches to plan for the future of work.

## References


[1] James Manyika, Susan Lund, Michael Chui, Jacques Bughin, Jonathan Woetzel, Parul Batra, Ryan Ko, and Saurabh Sanghvi. 2017. Jobs lost, jobs gained: Workforce transitions in a time of automation. *McKinsey Global Institute*. Retrieved from https://www.mckinsey.com/featured-insights/future-of-work/jobs-lost-jobs-gained-what-the-future-of-work-will-mean-for-jobs-skills-and-wages

[2] National Association of State Workforce Agencies. 2023. Legislative Priorities: Data Infrastructure. Retrieved from https://www.naswa.org/advocacy/government-relations/2023-legislative-priorities

[3] R. M. del Rio-Chanona, P. Mealy, A. Pichler, F. Lafond, and J. D. Farmer. 2020. Supply and demand shocks in the COVID-19 pandemic: An industry and occupation perspective. *Oxf. Rev. Econ. Pol.* 36 (2020), S94–S137. DOI: https://doi.org/10.1093/oxrep/graa033

[4] J. I. Dingel and B. Neiman. 2020. How many jobs can be done at home? *J. Publ. Econ.* 189 (2020), 104235. DOI: https://doi.org/10.1016/j.jpubeco.2020.104235

[5] S. Buckner-Petty, A. M. Dale, and B. A. Evanoff. 2019. Efficiency of autocoding programs for converting job descriptors into standard occupational classification (SOC) codes. *Am. J. Ind. Med.* 62 (2019), 59–68. DOI: https://doi.org/10.1002/ajim.22928

[6] M. Schmitz and L. Forst. 2016. Industry and occupation in the electronic health record: An investigation of the National Institute for Occupational Safety and Health industry and occupation computerized coding system. *JMIR Med. Info.* 4 (2016), e5. DOI: https://doi.org/10.2196/medinform.4839

[7] Daniel E. Russ, Kwan-Yuet Ho, Joanne S. Colt, Karla R. Armenti, Dalsu Baris, Wong-Ho Chow, Faith Davis, Alison Johnson, Mark P. Purdue, Margaret R. Karagas, Kendra Schwartz, Molly Schwenn, Debra T. Silverman, Calvin A. Johnson, and Melissa C. Friesen. 2016. Computer-based coding of free-text job descriptions to efficiently identify occupations in epidemiological studies. *Occup. Environ. Med.* 73 (2016), 417–424. DOI: https://doi.org/10.1136/oemed-2015-103152

[8] S. De Matteis, D. Jarvis, H. Young, A. Young, N. Allen, J. Potts, A. Darnton, L. Rushton, and P. Cullinan. 2017. Occupational self-coding and automatic recording (OSCAR): A novel web-based tool to collect and code lifetime job histories in large population-based studies. *Scand. J. Work. Environ. Health* 43 (2017), 181–186. https://doi.org/10.5271/sjweh.3613

[9] Nile Dixon, Marcelle Goggins, Ethan Ho, Mark Howison, Joe Long, Emma Northcott, and Karen Shen. 2023. Occupational models from 42 million unstructured job postings. *Patterns.* 4 (2023), 100757. DOI: https://doi.org/10.1016/j.patter.2023.100757

[10] Sarah H. Bana. 2021. job2vec: Using language models to understand wage premia. Retrieved from https://conference.iza.org/conference_files/DATA_2021/bana_s26582.pdf
[11] Chris Anstey. 2023. "Garbage" surveys erode US data confidence. *Bloomberg*. Retrieved from https://www.bloomberg.com/news/newsletters/2023-02-15/us-economic-data-why-surveys-for-jolts-and-payroll-numbers
[12] Pawel Adrjan and Reamonn Lydon. 2023. What do wages in online job postings tell us about wage growth? SSRN. Retrieved from https://dx.doi.org/10.2139/ssrn.4451751
[13] Mintaka Angell, Samantha Gold, Justine S. Hastings, Mark Howison, Scott Jensen, Niall Keleher, Daniel Molitor, and Amelia Roberts. 2021. Estimating value-added returns to labor training programs with causal machine learning. *OSF Preprints*. DOI: https://doi.org/10.31219/osf.io/thg23
[14] White House Council of Economic Advisers. 2018. Addressing america's reskilling challenge. Retrieved from https://trumpwhitehouse.archives.gov/briefings-statements/cea-report-addressing-americas-reskilling-challenge/
[15] Martha Laboissiere and Mona Mourshed. 2017. Closing the skills gap: Creating workforce-development programs that work for everyone. *McKinsey & Company*. Retrieved from https://www.mckinsey.com/industries/education/our-insights/closing-the-skills-gap-creating-workforce-development-programs-that-work-for-everyone
[16] Melanie Zaber, Lynn Karoly, and Katie Whipkey. 2019. Reimagining the workforce development and employment system for the 21st century and beyond. *RAND Corporation*, Research Report RR-2768-RC. DOI: https://doi.org/10.7249/RR2768
[17] Mark Howison, Joe Long, and Justine S. Hastings. 2023. Recommending career transitions to job seekers using earnings estimates, skills similarity, and occupational demand. *SSRN*. Retrieved from https://dx.doi.org/10.2139/ssrn.4371445
[18] Catherine Isley and Sarah A. Low. 2022. Broadband adoption and availability: Impacts on rural employment during COVID-19. *Telecommun. Polic.* 46, 7 (2022), 102310. DOI: https://doi.org/10.1016/j.telpol.2022.102310
[19] Korn Ferry. 2018. Future of work: The global talent crunch. Retrieved from https://www.kornferry.com/content/dam/kornferry/docs/pdfs/KF-Future-of-Work-Talent-Crunch-Report.pdf
[20] Nicole Maestas, Kathleen J. Mullen, and David Powell. 2023. The effect of population aging on economic growth, the labor force, and productivity. *Amer. Econ. J.: Macroecon.* 15, 2 (2023), 306–332. DOI: https://doi.org/10.1257/mac.20190196
[21] U.S. Bureau of Labor Statistics. 2018. Standard occupational classification (SOC) system. Retrieved from https://www.bls.gov/soc/2018/home.htm
[22] Amazon Web Services. 2023. Amazon bedrock user guide: Prompt engineering guidelines. Retrieved from https://docs.aws.amazon.com/bedrock/latest/userguide/prompt-engineering-guidelines.html
[23] Honey Batra, Amanda Michaud, and Simon Mongey. 2023. Online job posts contain very little wage information. *NBER Working Paper 31984*. DOI: https://doi.org/10.3386/w31984

Received 9 January 2024; revised 1 April 2024; accepted 26 May 2024