revelio labs

August 28, 2025

Revelio Public Labor Statistics (RPLS)

Methodology

Table Of Contents

Introduction

Reliable labor market statistics are crucial for economic policymaking, business strategy, and financial decision-making. Official measures from the Bureau of Labor Statistics (BLS) remain indispensable for foundational metrics like labor force participation and unemployment rates and employment levels, but they are increasingly struggling to meet the needs of fast-moving markets. Survey response rates have been declining steadily, introducing coverage and nonresponse biases that require heavier reliance on statistical imputation. Furthermore, chronic underfunding has constrained modernization, leaving key surveys dependent on outdated methodologies and limiting the agency's ability to expand sample sizes or integrate new data sources. Recent political pressures further threaten the independence and perceived credibility of official statistics.

Against this backdrop, private-sector datasets have emerged as a necessary complement. By leveraging alternative sources such as online professional profiles and job postings, Revelio Labs is able to generate real-time, granular insights into employment, hiring, separations, and wages. These datasets, while not a replacement for official statistics, offer advantages of scale, timeliness, and flexibility, and thus function as a necessary complement to government data.

This document introduces Revelio Public Labor Statistics (RPLS), a comprehensive set of labor market data designed to provide timely, granular, and nationally representative insights into U.S. labor market. RPLS draws on a proprietary dataset of over 100 million professional profiles, capturing roughly two-thirds of all employed individuals in the United States. In addition to employment data, RPLS incorporates aggregated job postings from online sources, providing a forward-looking view of labor demand and hiring trends across industries, occupations, and regions.

RPLS publishes three core categories of metrics. First, employment levels and changes provide a real-time view of job gains and losses, comparable to the Current Employment Statistics (CES) from the BLS. Second, job openings and employee hiring and separation rates offer insights into labor flows and workforce churn, serving as a private-sector analog to the Job Openings and Labor Turnover Survey (JOLTS). Third, wages for new positions provide a forward-looking perspective on compensation trends, complementing BLS earnings statistics. All metrics are available at the level of occupation (SOC 2-digit), sector (NAICS 2-digit), and state.

RPLS is published monthly, with updates released the day before Jobs Friday and the day before JOLTS releases, providing early visibility into trends that drive policymaking, business strategy, and labor market research. By combining comprehensive coverage, timeliness, and methodological rigor, RPLS aims to close critical information gaps and complement official statistics, offering both alignment with key headline metrics and additional granularity where official sources are limited.

Employment

RPLS provides a set of employment statistics derived from over 100 million professional profiles sourced from professional networking websites. After deduplication, adjustments for reporting lags, and reweighing to ensure that the data resembles the national distribution of the workforce, these data yield timely and detailed measures of employment dynamics. The resulting data series captures the level of total employment (headcount), which informs monthly job gains, and the flows of workers into and out of jobs. This data enables us to track hiring, separations, and net employment changes with a scope and frequency comparable to official labor statistics.

Methodologies

Data Sources

Revelio Labs leverages data from online professional profiles on networking platforms such as LinkedIn. Each profile contains information on current and past positions, including the company, job title, job location, and in many cases, descriptions of responsibilities. The profiles span all geographic areas in the United States, ensuring coverage across regions. Our data represents all US-based workers in various companies, sectors, occupations, and geographies.

Profile Deduplication

To transform raw profile data to a reliable dataset of worker records, we implement several corrections. The first step is to eliminate fake users and consolidate duplicate profiles. Our Fake User Model detects fraudulent accounts using signals such as connection counts (relative to seniority), duration of previous positions, frequency of job changes, account inactivity, missing start or end dates, and anomalies in names (e.g., numbers or non-alphabetic characters).

Next, Revelio Labs' deduplication pipeline addresses individuals with multiple LinkedIn profiles. Because pairwise comparison across the entire dataset is computationally infeasible, we first generate hash values for key attributes (e.g., name, education, work history) to block together candidate duplicates. Within each block, we apply a similarity function to compute pairwise similarity scores; profiles exceeding a defined threshold are consolidated into a single individual.

Standardization

In online profiles, job information often appears in heterogeneous forms. For example, workers may report their employer as "Bank of America" or "BofA." Job titles are equally variable: a worker at Apple might state their title as "Apple Genius," whereas

similar roles appear as “Retail Sales Associate” at H&M or “Mobile Associate” at T-Mobile. Such variation obscures the true functional responsibilities of roles and the structure of the workforce, complicating efforts to generate macroeconomic aggregates. Accurate resolution of these variations is therefore critical for determining workforce size, industry distribution, and occupational composition.

- **Sector code assignment:** To assign companies to their respective NAICS sectors, we first resolve all lexical and structural variation in employer names and ensure that all positions are consistently mapped to the same entity. Revelio Labs employs a proprietary Company Mapping algorithm that integrates multiple data sources to standardize and disambiguate employer identities. Our company universe is built from employment records in professional profiles—where individuals link their work history to employers—and is enriched with corporate datasets such as FactSet, which provide information on subsidiaries, security identifiers, and metadata. Entities are linked using overlapping features including company names, URLs, headquarters locations, and founding years. Each resolved entity is assigned a persistent Revelio Company Identifier (RCID), which aggregates all relevant information about the company. Once mapped to RCIDs, companies are assigned 6-digit NAICS codes based on their primary industry classification in external corporate datasets, supplemented with machine-learning models that incorporate firm descriptions and the distribution of employee occupations. For analysis, we aggregate to the 2-digit NAICS level to reduce potential misclassification errors at finer levels.

- **Occupation code assignment**: Occupational classification follows a parallel embedding-based approach. We begin by sourcing activity data from (1) resumes and online profile text describing prior roles and (2) responsibility sections extracted from job postings. These sources are used to train activity-level embeddings that capture the functional content of each role. Job titles are normalized by removing modifiers (e.g., “senior,” “principal”), tokenized, and embedded using a fine-tuned ModernBERT model. For each position, we generate embeddings from multiple textual dimensions—including the raw title, job description, listed skills, same-title position descriptions within the same firm, and the individual’s career history. A weighted average of these component embeddings is computed, with weights determined by dynamic confidence scores based on information richness. O*NET occupations are embedded in the same space, enabling direct similarity comparisons between positions and standardized occupation codes. Each position is then assigned to the most similar 6-digit O*NET code, and results are aggregated to 2-digit SOC categories for reporting. This step ensures comparability across roles while minimizing noise from fine-grained classification errors.

Sample Creation

Our sample includes all private- and public-sector, full-time and part-time workers with observable professional profiles, excluding military personnel, interns, freelancers, self-employed and proprietor workers, full-time students, and individuals on leave or sabbatical. We restrict the sample to users who report a location in the United States, and we track them across different versions of the data to capture relocation. To reduce noise from very small employers, we further limit the sample to workers affiliated with companies that have at least five observed profiles. The final dataset contains approximately 104 million profiles, covering about 65% of the U.S. workforce.

Lags in Reporting

Professional profiles and online resumes offer rich signals about workforce dynamics, including both inter- and intra-company transitions. However, these data sources are inherently noisy and delayed. In Revelio Labs' data, there are two primary sources of lag. First, employees may not update their profiles immediately after changing jobs, particularly in the case of involuntary separations and layoffs, which leads to a lag in reporting. Second, workforce data providers only scrape individual profiles periodically, resulting in discrete observations at non-uniform intervals. These sources of delay contribute to an observable lag distribution that correlates with attributes such as job function, geography, company size, and industry.

To address this, we construct a model of workforce dynamics that integrates observed transitions with inferred lags. Our model predicts the inflows and outflows that will be revealed once all public profiles have been updated. It does this by taking snapshots of companies' observed workforces at various periods of time and comparing them to future snapshots, predicting currently unreported flows based on previously underreported flows. The model is designed to capture the temporal structure of workforce changes and provides probabilistic estimates of current headcount by predicting what would be the case if all workers reported their transitions promptly. The model is trained on both the most recent snapshot and a longitudinal sequence of historical snapshots, enabling it to learn how headcount observations evolve over time. The model also explicitly accounts for known events, such as layoffs and hiring surges. These can either be incorporated as covariates or inferred as latent change points within the model, allowing for discontinuities in headcount trends to be detected or constrained. This approach better aligns observed headcounts with real-world developments and mitigates under- or over-estimation following major corporate events.

Sampling Weights

Online professional profiles are not a representative sample of the U.S. workforce. Profiles on these platforms disproportionately reflect white-collar occupations, larger firms, and workers in metropolitan areas, while under-representing blue-collar jobs, smaller companies, and rural locations. To correct for these biases and produce nationally representative measures of employment, we construct sampling weights using two complementary models at the micro and macro sides:

1. **Micro sampling weight adjustment: Company Multiplier Model**
   Online presence varies systematically across companies due to factors such as industry, size, or corporate culture. To account for this, we compare observed headcounts in our data to external benchmarks of firm-level employment, such as 10K filings for publicly traded firms. From these comparisons, we derive company-level multipliers that scale observed workers to their estimated true workforce share.

2. **Macro sampling weight adjustment**
   At the aggregate level, we align user representation with national benchmarks from the BLS. This is done in two steps:

- **Occupation–Location Weights:** We use data from the Occupational Employment and Wage Statistics (OEWS) to obtain the nationally representative distribution of employment across occupations and geographic areas. Standardized job titles in our dataset are probabilistically mapped to Standard Occupational Classification (SOC) codes, and location information is harmonized to Metropolitan Statistical Areas (MSAs). Using this benchmark distribution, we assign weights at the individual position level to correct for over- or underrepresentation by occupation and geography. For example, if engineers in California are estimated to have a 90% likelihood of appearing in professional networking data, each observed engineer is weighted as 1.1 workers. Conversely, if nurses in Vermont have only a 25% likelihood of being represented, each observed nurse is weighted as 4 workers.
- **Occupation–Sector Weights:** We also implement an additional layer of weighting to ensure that the distribution of the workforce in the Revelio Labs data matches the national distribution of workers. We first compute weights by occupation at the 5-digit O*NET codes at the position level, using OEWS data to correct for the

systematic underrepresentation of specific occupations in our data. Finally, we scale the headcounts by industry weights also computed from OWES. This second layer of adjustment complements the occupation–location weights, resulting in a two-tiered weighting scheme that ensures our sample is representative both geographically and across industries.

From these lag correction and weighting adjustments, we produce three core time series: headcount, reflecting the estimated total employment; inflows, capturing hires and new additions to the workforce; and outflows, capturing voluntary and involuntary separations. Each of these series is estimated at the sector, by occupation, by state level.

Seasonal Adjustment

Each monthly dataset is seasonally adjusted with STL decomposition (implemented via the statsmodels package). STL decomposes the not-seasonally-adjusted (NSA) series into trend, remainder, and seasonal components, with the latter excluded. Our seasonally adjusted (SA) value is equal to the original series minus the estimated seasonal component. After making the seasonal adjustment at the group level, we construct a U.S. total by aggregating each panel by month for both NSA and SA values.

Changes In Employment

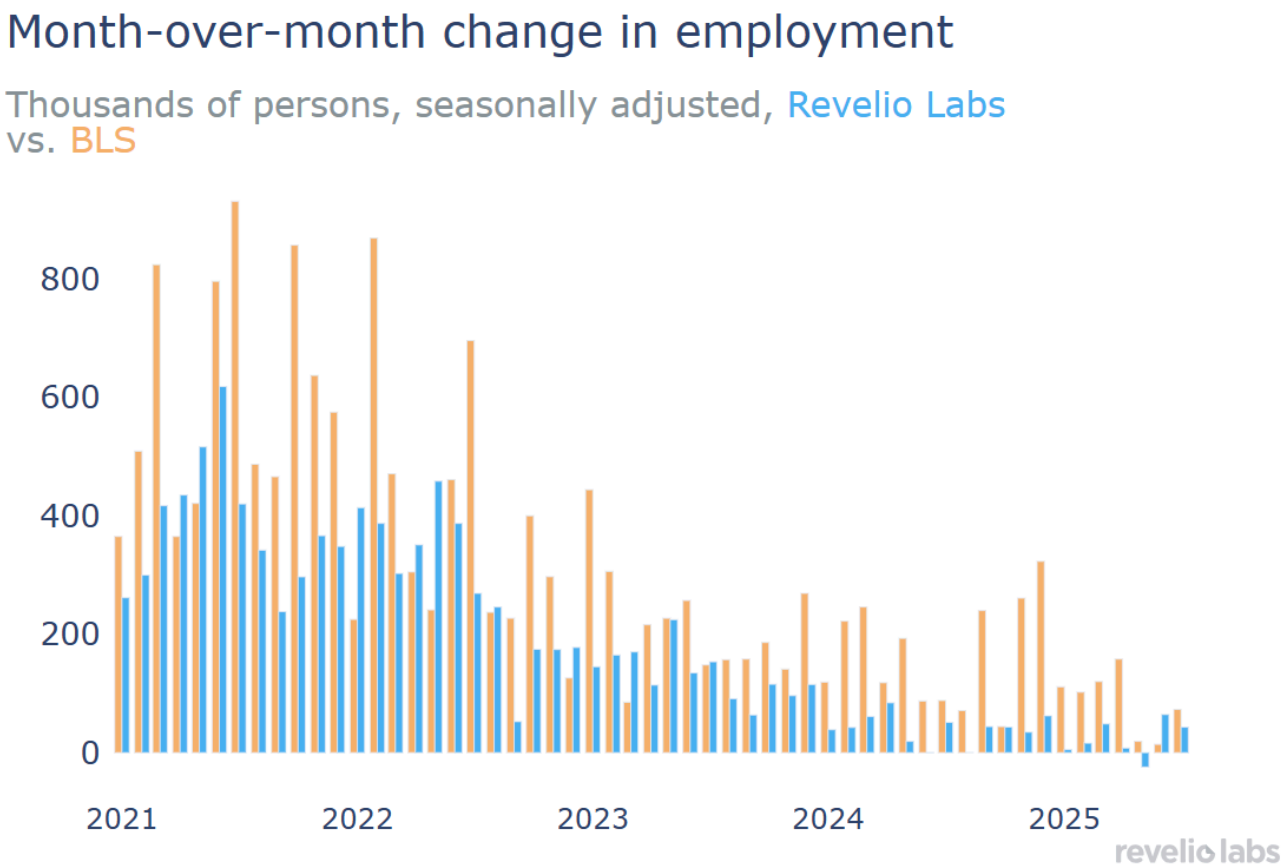
Once the headcount is estimated and seasonally adjusted at the sector-occupation-state level, we aggregate across these categories to reach the total employment in the US. Monthly gains in employment are calculated as the month-over-month change in the level of seasonally adjusted headcount.

Comparison to CES

The BLS tracks changes in employment via two main surveys. The establishment survey that is used to generate the Current Employment Statistics (CES), surveys approximately 121,000 businesses and government agencies, covering roughly 631,000 individual worksites, and serves as the source of the monthly “headline” change in payroll jobs. The household survey, covering around 60,000 households, captures broader labor market trends, including payroll jobs, self-employment, and multiple job holdings. Together, these surveys provide complementary snapshots of U.S. employment.

When we compare Revelio Labs' estimated employment series to CES estimated employment, we find that Revelio Labs captures the same broad trends while offering more timely observations and flexibility in disaggregation.

The figure below illustrates the monthly change in total employment from Revelio Labs relative to CES-reported monthly change in payroll employment (jobs added). RPLS estimates move in tandem with official data, tracking sector- and occupation-level employment gains and losses across time with a correlation of ~ 0.74.



The correlation between monthly employment changes highlights the alignment of these datasets. The correlation between the establishment and household surveys is relatively modest (0.31), reflecting differences in methodology and coverage. RPLS exhibits a strong correlation with CES monthly changes (0.74) and a moderate correlation with the household survey (0.29), indicating that our estimates track payroll-based employment measures most closely.

Comparison to other employment data sources

One prominent source of employment data is produced by ADP, drawing on payroll records covering more than half a million companies and over 25 million employees. Unlike Revelio Labs' profile-based data, which captures both private- and public-sector workers, ADP's measure is limited to private payroll employees. ADP draws from two sources: payroll transactions, which record when individuals are paid and how much, and administrative records that identify who is currently on the company payroll, even if not paid in a given cycle. This provides detailed business-level insight into private employment trends, but differs in scope and design from Revelio Labs' approach, which relies on individual-level professional profiles and offers a broader view of the workforce, including public-sector employment.

The table below shows the correlations between monthly changes in employment from January 2021 to July 2025 between Revelio Labs, the establishment survey, the household survey, and ADP private employment data. The correlation between the establishment survey data and the household survey data is modest at 0.31, reflecting different concepts and coverage. By contrast, Revelio Labs tracks the establishment survey most closely with a correlation coefficient of 0.8. Meanwhile, Revelio Labs' change in employment has a relatively modest correlation of 0.35 correlation to the household survey. ADP's private-payroll series also correlates with the establishment survey (correlation of 0.70), but its scope is narrower as it only measures private payroll employment.

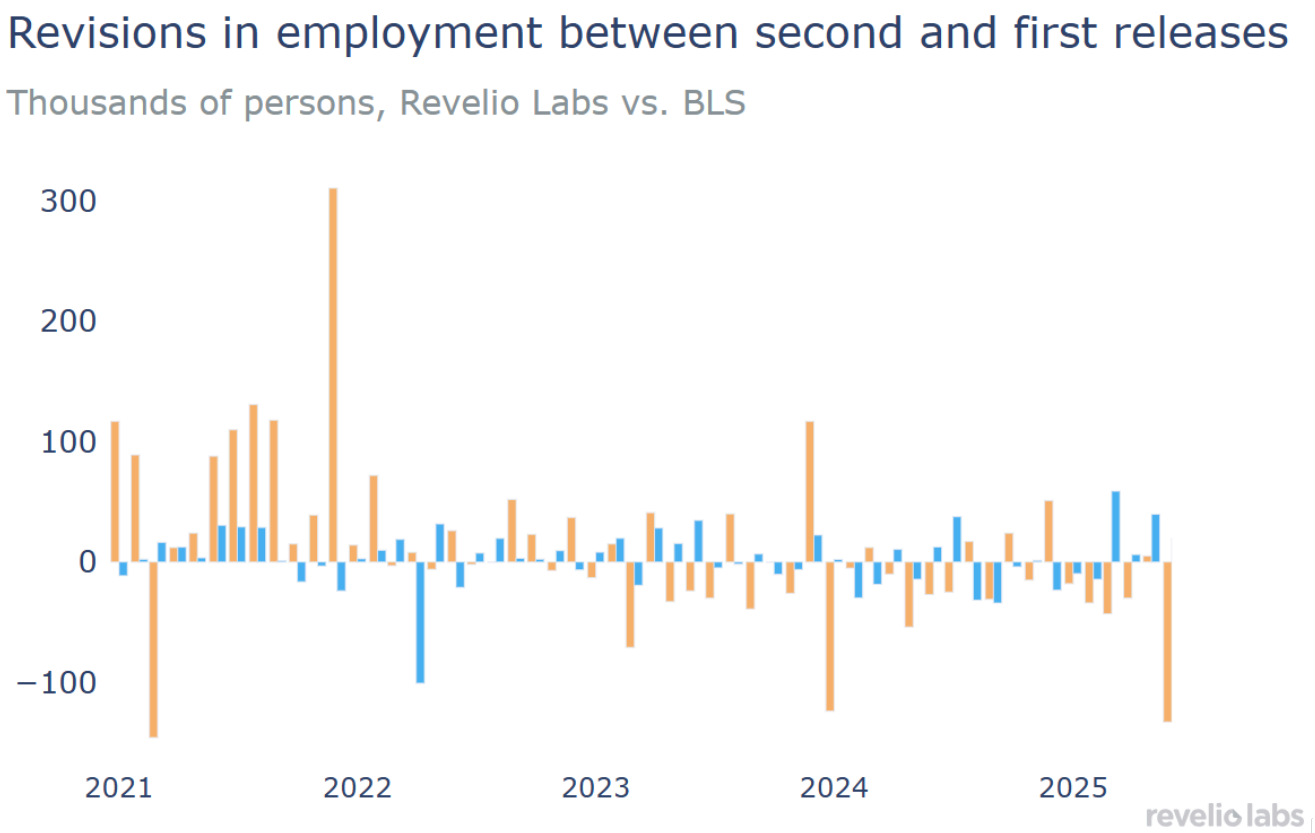
The correlation between monthly employment changes highlights the alignment of these datasets. The correlation between the establishment and household surveys is relatively modest (0.31), reflecting differences in methodology and coverage. RPLS exhibits a strong correlation with CES monthly changes (0.80) and a moderate correlation with the household survey (0.35), indicating that our estimates track payroll-based employment measures most closely.

	Establishment Survey	Household Survey	ADP	Revelio Labs
Establishment Survey	1.00			
Household Survey	0.31	1.00		
ADP	0.71	0.30	1.00	
Revelio Labs	0.74	0.29	0.50	1.00

Revisions

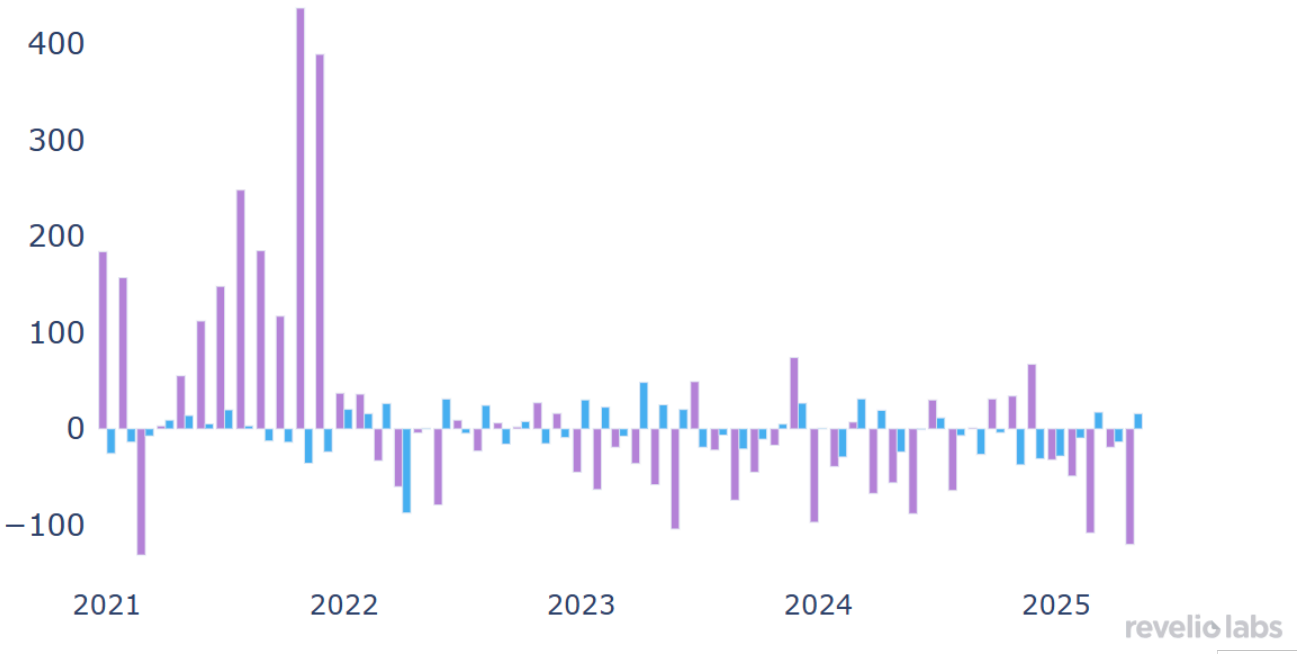
Data revisions are an expected part of employment statistics, reflecting new information, corrections, and methodological adjustments. Historical revisions can be substantial, and they are particularly relevant when policy decisions or financial markets rely on preliminary estimates.

RPLS updates historical employment estimates as new profile data are ingested and adjustments for representation, reporting lags, and seasonality are refined. The two figures below show the magnitudes of the difference of RPLS revisions in the monthly change in employment between the first and second releases and the first and third releases respectively over the past 4 years. RPLS revisions has been roughly half that observed in CES revisions in both the positive and negative directions, reflecting the stability and timeliness of our real-time pipeline. Furthermore, RPLS revisions are less procyclical and serially correlated compared to CES revisions.



Revisions in employment between first and third releases

Thousands of persons, Revelio Labs vs. BLS



Hiring And Attrition

The BLS' Job Openings and Labor Turnover Survey (JOLTS) has long been the main source for U.S. hiring and attrition data.However, its survey response rate has plummeted in the wake of the pandemic, with recent response rates now only around 35%. In addition, JOLTS is published with a full one-month lag after the reference month.

Our measures offer a timelier and more complete alternative. Hiring and attrition estimates are derived from the workforce dynamics dataset used in RPLS employment measures. The same methodologies for profile deduplication, sample creation, lag adjustment, and sampling weights are implemented.

For any given segment, such as industry, occupation, or geography, and for each calendar month, we define outflows as the number of separations to external destinations that occurred during that month and inflows as the number of external hires into the segment during that month. Internal job changes within the same company are not counted. Headcount refers to the number of workers employed in that month.

To reduce seasonal noise and reporting volatility, we compute annualized hiring and attrition rates on a rolling twelve-month basis. These are not seasonally adjusted, because the twelve-month moving window already smooths out seasonal fluctuations in the data. The hiring rate is calculated as the total number of inflows over the most recent twelve months divided by the average headcount over those same twelve months, while the attrition rate is likewise calculated as the total number of outflows divided by the average headcount.

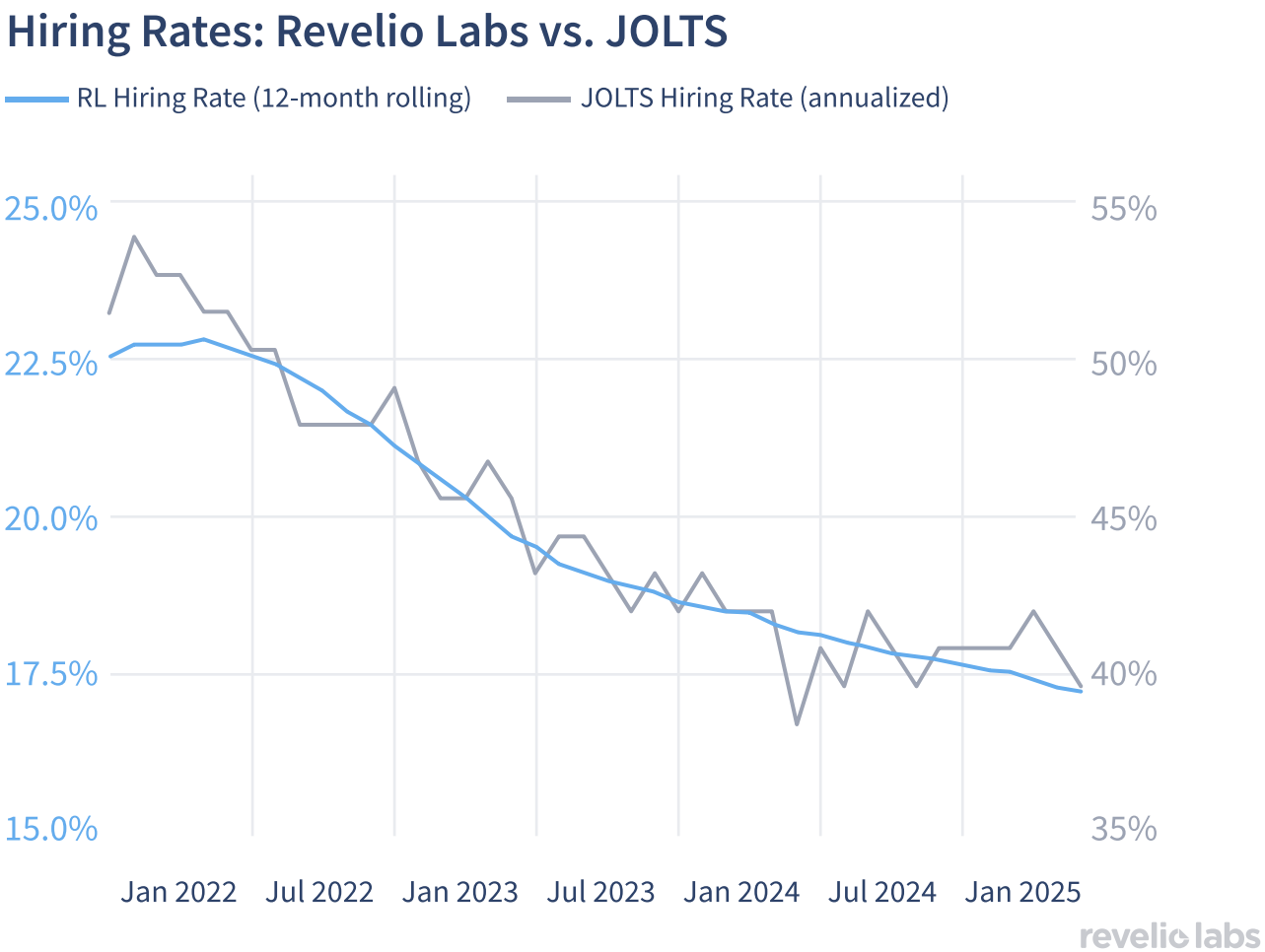
$$a_{j,o,s}(t) = 100 \cdot \frac{\sum_{\tau=t-11}^{t} outflows_{j,o,s}(\tau)}{\bar{c}_{j,o,s}(t)}$$

$$h_{j,o,s}(t) = 100 \cdot \frac{\sum_{\tau=t-11}^{t} inflows_{j,o,s}(\tau)}{\bar{c}_{j,o,s}(t)}$$

Where i is the 2-digit NAICS code, j is the 2-digit SOC occupations, s is the state, and t is the calendar moth. *Outflows* represent the number of separation in a given cell in month t, and *inflows* represent the number of hires in a given cell in month t. Finally, c is the average headcount over the most recent 12 months.

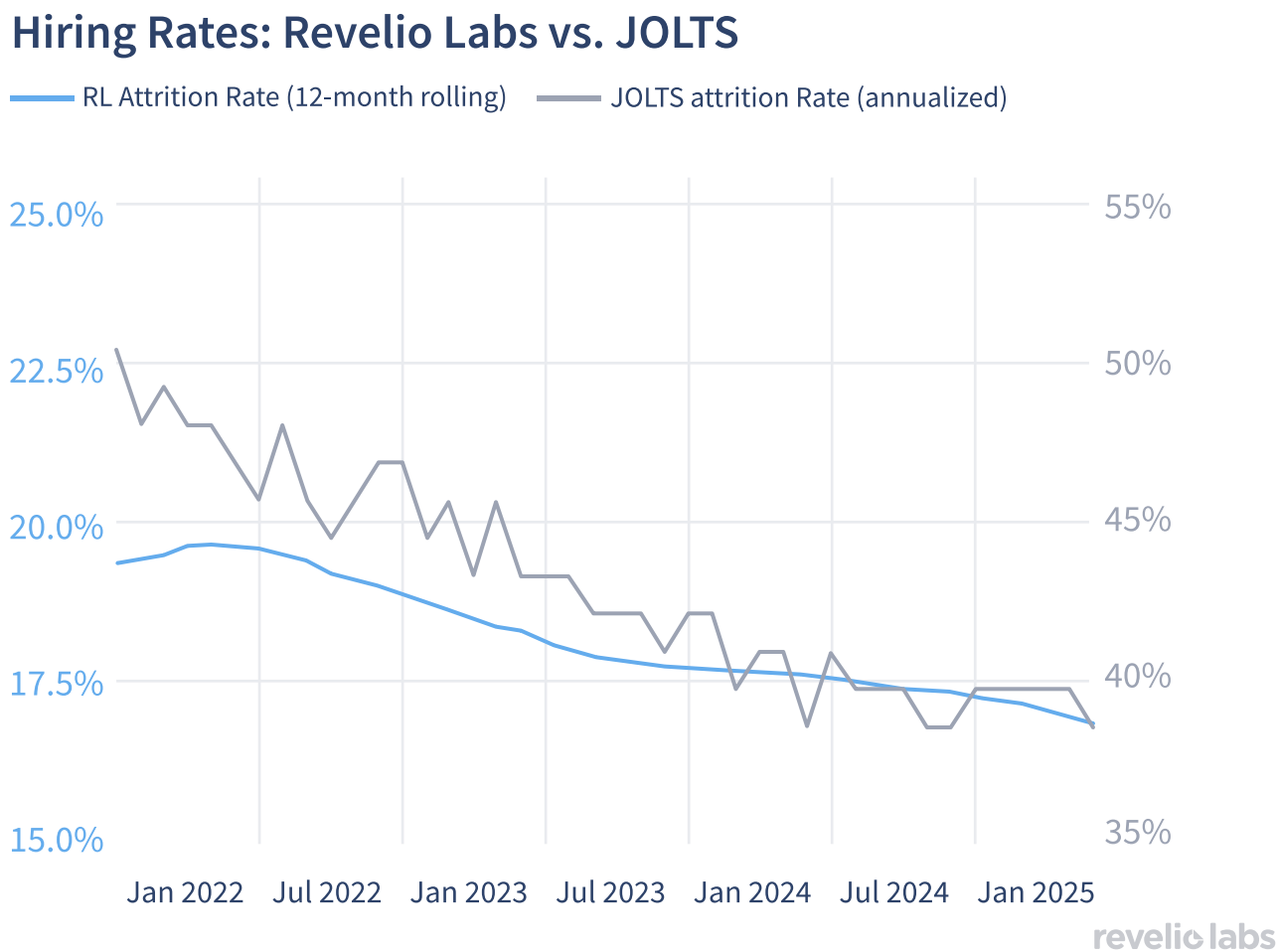
Comparison to JOLTS

Revelio Labs data show a steady decline in hiring and attrition rates in recent years, as both measures have fallen in tandem with the cooling labor market following the post-pandemic hiring boom. These trends track closely with JOLTS, though with consistently lower levels of both hiring and attrition. For comparability, JOLTS rates are annualized in the figures below by multiplying the monthly rates by twelve.



The gap in levels arises from several methodological distinctions. First, JOLTS defines hiring and separations at the establishment level, capturing all flows into and out of payroll jobs. This includes not only external hires and separations but also within-firm transitions across establishments, seasonal and temporary hires, and short-duration jobs. By contrast, Revelio Labs measures these dynamics at the individual position level, based on observed changes in online professional profiles. This framework is more representative of permanent, full-time professional positions and excludes much

of the high-frequency churn in seasonal or part-time work. As a result, our measures yield systematically lower levels of hiring and attrition that more closely align with the expectations of HR and corporate leaders.



Differences in job change visibility also contribute. JOLTS records all separations, including quits, layoffs, retirements, and other involuntary exits. Many of these "silent exits"—into unemployment, retirement, care-giving, or gig work—are somewhat less likely to be recorded in professional profiles, which would further dampen the levels observed in our data.

Despite these differences, the directional alignment between the two series is strong. Since January 2022, both JOLTS and Revelio Labs measures of hiring have declined by roughly 30 percent. However, while JOLTS hiring has largely stabilized since mid-2024, with some month-to-month volatility, Revelio Labs' measure indicates a more persistent weakening through 2025.

Job Openings

Job postings scraping and deduplication

Job openings are a critical signal for understanding the health of the labor market and designing effective policy. Unlike employment figures, which reflect the current state of the labor market, job postings can serve as a leading indicator, showing where hiring is expected to increase, which industries are expanding, and where labor demand is shifting.

Revelio Labs collects data on job postings from sources including company career pages, job boards (e.g., LinkedIn, Indeed), and staffing firms. We then implement a multi-stage deduplication of these postings in which we standardize and consolidate postings into unique job records, which offers a holistic view of labor demand presented in our COSMOS data offering.

The deduplication process begins by normalizing key attributes, such as job role (identified based on raw job titles using Revelio Labs' proprietary job taxonomy), RCID (Revelio Company Identifier, mapped using available raw company information including company name, website, and other attributes), and location (standardized to city level). We then construct candidate duplicate sets by restricting pairwise comparisons to postings that match on these normalized fields. To further narrow comparisons, we require temporal overlap, defined as postings whose active windows intersect within a ±2-day buffer on both start and end dates.

Within each set of postings, we compute pairwise similarity scores between postings using a combination of content-based and metadata-based features. The textual similarity score is calculated using Levenshtein distance applied to the raw job title, company, and location fields. Postings with similarity scores exceeding a threshold of 80% are classified as duplicates. We then select a representative posting to display, defined as the posting with the earliest post date from the highest quality raw source.

This approach enables us to construct a high-precision deduplicated dataset at scale, preserving fine-grained information on job market dynamics while filtering out noise and redundancy from multi-channel job advertisements.

Active job postings include any posting that was active at any point during a given month. A posting is considered active in a given month if the post date occurred before the start of the month or at any point during the month, and the remove date is either

during or after the reference month. If the post has not removed data, we consider it still active. This approach ensures that postings are counted in every month they were active, rather than only in the month they were created, and captures ongoing listings that remain visible across multiple months. The results are then aggregated to provide the total number of active postings by industry, occupation, and state in every month.

Standardization

Sector code assignment: To classify employers into NAICS sectors, Revelio Labs first standardizes employer identities by resolving lexical and structural variations in company names. We employ a proprietary company mapping system that integrates professional profile data, corporate datasets (e.g., FactSet), and metadata such as URLs, headquarters, and founding dates to disambiguate and unify entities. Each entity is assigned a persistent Revelio Company Identifier (RCID), which serves as a hub for all relevant information. Companies are then linked to 6-digit NAICS codes primarily through external corporate datasets, supplemented by machine learning models that use company descriptions and workforce composition. For analytical consistency, we typically report results at the 2-digit NAICS level, where classification accuracy is more robust.

Occupation assignment: Occupational coding uses a parallel embedding-based approach. Role text is sourced from individual professional profiles (e.g., work history, resumes) and job postings (e.g., responsibilities and requirements). Job titles are normalized and embedded using a fine-tuned ModernBERT model, alongside embeddings derived from job descriptions, listed skills, same-title roles within the same firm, and an individual's career history. These multiple textual signals are combined through a weighted average, with weights adjusted dynamically based on information richness. Each position is then mapped to SOC/ONET occupation codes, ensuring consistency across datasets. Application to postings. This mapping framework is then applied to our job postings dataset, similarly to our positions data.

Each posting is linked to a standardized employer identity and assigned both NAICS and SOC/ONET codes using the same combination of external data, embeddings, and machine learning models. This ensures methodological consistency between the positions and postings datasets, allowing them to be analyzed jointly or compared directly across industries and occupations.

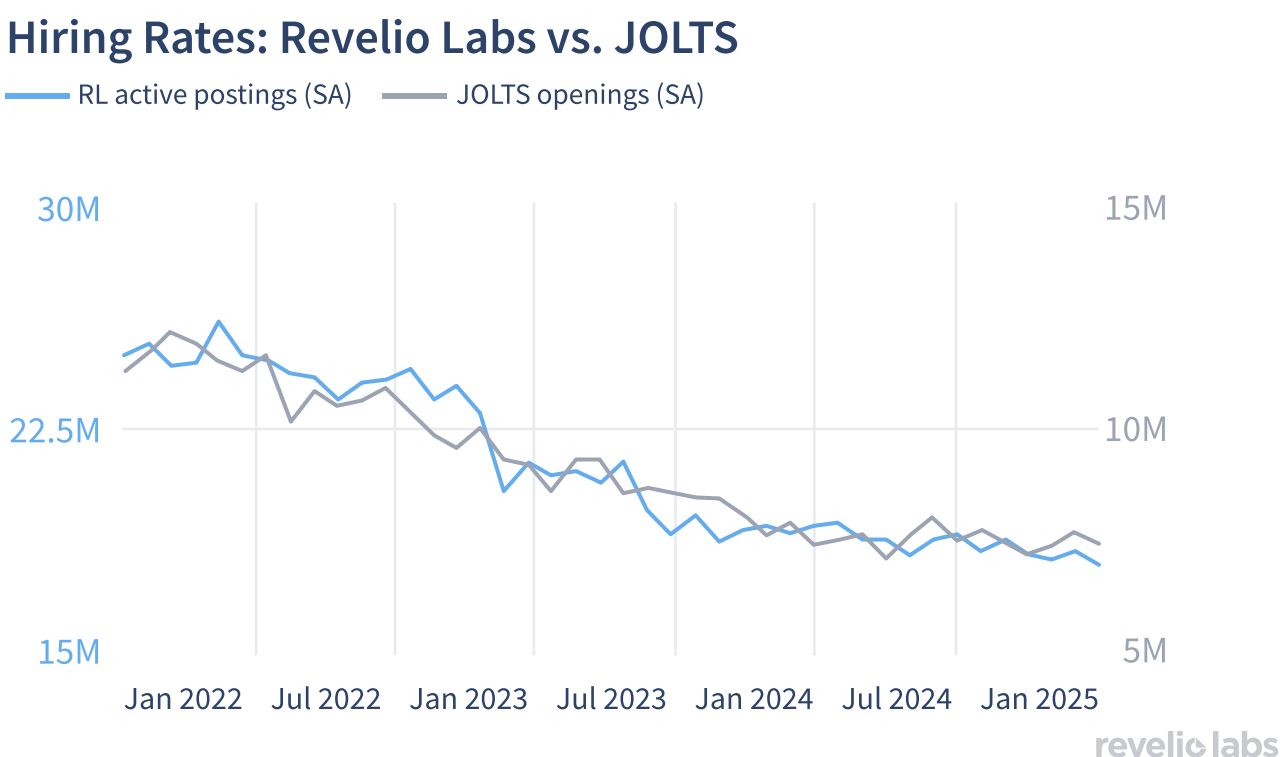
Seasonal Adjustment

Each monthly dataset is seasonally adjusted with STL decomposition (implemented via the statsmodels package). STL decomposes the not-seasonally-adjusted (NSA) series into trend, remainder, and seasonal components, with the latter excluded. Our seasonally adjusted (SA) value is equal to the original series minus the estimated seasonal component. After making the seasonal adjustment at the group level, we construct a U.S. total by aggregating each panel by month for both NSA and SA values.

Comparison to JOLTS Job Openings

When compared to job openings from JOLTS, the Revelio Labs Active Postings data shows a broadly similar trend. Both series indicate that job openings are down by roughly 30 percent since January 2022 and have largely stabilized over the past twelve months. JOLTS has exhibited more recent volatility month-to-month, while the Revelio Labs' measure has followed a smoother trajectory, with a more persistent downward drift observed since the start of 2025.

The difference in levels between the two series is more pronounced. The Revelio Labs data reveals substantially more active postings than JOLTS job openings. This reflects several methodological distinctions:



- **Reference period**: JOLTS measures openings as of the last business day of the month, while Revelio Labs counts all postings that were active at any time during the month. This leads to systematically higher counts in our data.
- **Definition of a job opening:** JOLTS applies a restrictive criterion, counting only jobs that are actively recruiting and intended to be filled within 30 days. By contrast, Revelio Labs includes any active posting, regardless of employer hiring horizon. This is particularly relevant for positions with longer recruitment cycles (e.g., senior roles, government jobs, or specialized technical occupations), or evergreen job openings.
- **Coverage:** The Revelio Labs dataset reflects openings scraped from across the web, including company career sites, job boards, and staffing firms. This broader scope captures multi-channel labor demand that may be underrepresented in JOLTS, especially for smaller firms or rapidly emerging occupations.

These methodological differences explain why Revelio Labs active postings consistently report a higher level of openings than JOLTS, while the directional movements remain closely aligned. Taken together, the comparison demonstrates that Revelio Labs data provide a complementary high-frequency view of labor demand dynamics.

Salaries From New Job Openings

Collecting salaries from job openings

Online job postings provide a rich source of salary information, indicating how much companies are willing to pay to future hires. Negotiated salaries that workers end up receiving are often higher than posted salaries, but the posted salary gives a useful lower bound. The BLS reports earnings gained by workers currently employed, providing a measure of how much workers are earning on average across the economy. Salaries from new job postings give a forward looking view of both the direction and speed at which average salaries are changing. It is useful to think of salaries from job postings as a flow measure, while earnings in the CES represent a stock measure.

For this metric, we use salaries from new job postings in our COSMOS job posting dataset, which unifies and deduplicates all our job postings sources. Sources include

company career pages, job boards (e.g., LinkedIn, Indeed), and staffing firms. We then implement a multi-stage deduplication of these postings in which we standardize and consolidate postings into unique job records, which offers a holistic view of labor demand presented in the COSMOS data offering. We begin by normalizing key attributes, such as job role (identified based on raw job titles using Revelio Labs' proprietary job taxonomy), RCID (Revelio Company Identifier, mapped using available raw company information including company name, website, and other attributes), and location (standardized to city level). We then construct candidate duplicate sets by restricting pairwise comparisons to postings that match on these normalized fields. To further narrow comparisons, we require temporal overlap, defined as postings whose active windows intersect within a ±2-day buffer on both start and end dates.

Within each set of postings, we compute pairwise similarity scores between postings using a combination of content-based and metadata-based features. The textual similarity score is calculated using Levenshtein distance, applied to the raw job title, company, and location fields. Postings with similarity scores exceeding a threshold of 80% are classified as duplicates. We then select a representative posting to display, defined as the posting with the earliest post date from the highest quality raw source.

This approach enables us to construct a high-precision deduplicated dataset at scale, preserving fine-grained information on job market dynamics while filtering out noise and redundancy from multi-channel job advertisements.

For our salary data, we use salaries as they are reported in job postings. Salaries in job postings are frequently reported as ranges, so we use the midpoint of the range for the metric. Our salary metric is reported as nominal annual base salaries in USD. Where hourly earnings are observed in the postings, we convert them into an annual salary by multiplying the number by 40 hours a week by 52 weeks.

During our sample period, January 2022 to date, 37% of all US job postings in our data record a salary value.

Sampling weights

In order to use salaries from job postings as a proxy for the earnings of new workers in the US, we need to ensure that the distribution of job postings resembles the national distribution of the workforce. Accordingly, we construct sampling weights for job postings to accurately represent the national distribution of the workforce. Our sampling weights are constructed using employment shares from the Occupational Employment and Wage Statistics (OEWS) published by the BLS. Our weights are constructed by comparing the share of job postings in each 2-digit NAICS (sector) code by 2-digit SOC (major occupation group) cell to the share its corresponding sector-occupation group cell makes up in terms of total US employment from the OEWS. The weight is constructed as the ratio of the share of employment in each sector-by-occupation cell (occupation o, sector j) from the total workforce, obtained OEWS, over the share of each sector-by-occupation cell from the total job postings that we observe.

$$w_{oj} = \frac{w_{oj}^{OEWS}}{w_{oj}^{RL}}$$

When NAICS code assignments in the job postings data are missing, weights are derived solely from SOC-level distributions, and conversely, when SOC code assignments in the job postings data are missing, weights are assigned based on NAICS-level distributions. Observations with both industry and occupation codes present are weighted using the joint OEWS benchmark. This weighting approach corrects for uneven job postings behavior and assumes that all sectors and industries post new jobs every month. This way, the resulting average salary growth of new job postings is not subject to differences in compositions of demand by sector.

To generate our aggregated data series, salaries are aggregated as weighted averages using sampling weights:

$$Salary_{ot}^{weighted} = \frac{\sum_s \sum_j salary_{otjs} * w_{oj}}{\sum_s \sum_j w_{oj}}$$

Seasonal adjustments

We seasonally adjust the data series, removing expected seasonal patterns from the data. We adjust our most granular dataset at the SOC by NAICS by state by month level using STL decomposition at the overall month level (implemented via the statsmodels package). STL decomposes the not-seasonally-adjusted (NSA) series into trend, remainder, and seasonal components, with the latter excluded. Our seasonally adjusted (SA) value is equal to the original series minus the estimated seasonal component. After seasonally adjusting the granular series, we aggregate the seasonally adjusted as well as non-seasonally adjusted salaries up to the more aggregated levels: SOC by month, state by month, and NAICS by month, as well as the overall series by month.

Comparison with earnings from CES

The establishment survey (CES) reports hourly earnings across all workers. To compare our annual salaries to BLS earnings, we convert BLS earnings to annual salaries by multiplying earnings by the average number of weekly hours worked (34.4 hours) and multiplying this by 52 weeks. As mentioned above, the BLS number reports salaries of the stock of workers across the economy. Our metric measures what the inflow of new workers is going to make.

The correlation between the aggregate average salaries from new job postings and average hourly earnings is 0.91. The plot below shows that while salaries from new job postings have generally grown slightly faster than earnings across all workers according to the CES, this dynamic has recently reversed, where salaries for new entrants are likely below those of existing workers.

Hiring Rates: Revelio Labs vs. JOLTS

Annual salaries in USD from new COSMOS job postings, BLS earnings