upwage

Summary of Bias Audit Results

Official Document
Last Update:
5/25/2026

Report on Upwage AI recruiting platform and on the suitability of its design and operating effectiveness relevant to automated employment decision tool bias and associated risks

Pursuant to Reporting on the New York City Local Law 2021/144 – Title: A Local Law to amend the administrative code of the city of New York, in relation to automated employment decision tools.

Audit Checklist *0 flagged, 100%*

Dates Conducted:	February 1, 2026 - April 30, 2026	Conducted By:	Sameer Saadi
Audit Scope:	AI Analyst	Report Issued On / By:	May 25, 2026 / Santiago Leon

Bias audits are performed quarterly, including disparate impact assessments & a review of all bias mitigation policies. See Bias Audit	YES
We closely monitor the evolving legislative landscape around AI to ensure our practices remain compliant with new regulations. See Legislation Tracker	YES
We stay informed about the latest developments and best practices around AI bias mitigation & governance See New Research Findings	YES
We track and report incidents to inform customers of performance and to inform our product development. See Incidences Tracker	YES

Table of Contents

Assertion of Management

Purpose

We (“Upwage”) have conducted an impartial audit (“Bias Audit”) to identify bias and sources of bias in our processes, AI and software design and resulting products and reporting (“Products and Services”). In accordance, we provide here a Bias Audit report (this document) including a Summary of Results and, if needed, any recommendations.

Scope

We confirm, to the best of our knowledge and belief, there is no relevant audit information in our possession, custody, or control that we did not subject to thorough review and reflects how our AI systems operated from February 2026 to April 2026.

Santiago Leon
Full Name

Chief Information Security Officer
Title

San Diego May 25, 2026
Location and Date

Santiago J Leon
Signature

Please direct questions to: santiago@upwage.com

Executive Summary

Purpose

We (“Upwage”) have conducted an impartial audit (“Bias Audit”) to identify bias and sources of bias in our processes, AI and software design and resulting products and reporting. In accordance, we provide here a Bias Audit report (this document) including a Summary of Results and, if needed, any recommendations.

Scope

We confirm, to the best of our knowledge and belief, there is no relevant audit information in our possession, custody, or control that we did not subject to thorough review and reflects how our AI systems operated from February 2026 to April 2026.

Key Findings

No adverse impact found across overall standalone and intersectional calculable impact ratios in regard to the application of the four-fifths rule.

Non-intersectional, Gender, sorted by Scoring rate

Gender	Candidates	Selection Rate	EEOC Impact RatioThreshold	Impact Ratio (Error)
Female	26,644	0.81	0.80	1.00 (0.00)
Male	26,710	0.79	0.80	0.98 (0.00)
Unknown	5,363	0.83	0.80	1.03 (0.01)

Non-intersectional, Race/Ethnicity

Race/Ethnicity	Candidates	Selection Rate	EEOC Impact RatioThreshold	Impact Ratio (Error)
Asian	3,641	0.88	0.80	1.00 (0.01)
Black	10,389	0.81	0.80	0.92 (0.00)
Hispanic	6,777	0.79	0.80	0.90 (0.01)
Other	103	0.82	0.80	**N/A***
Unknown	943	0.86	0.80	**N/A***
White	36,864	0.80	0.80	0.91 (0.00)

Intersectional, Gender and Race/Ethnicity

Gender	Race/ Ethnicity	Candidates	Selection Rate	EEOC Impact Ratio Threshold	Impact Ratio (Error)
Female	Asian	1,373	0.88	0.80	1.00 (0.01)
Female	Black	5,250	0.83	0.80	0.95 (0.01)
Female	Hispanic	3,057	0.80	0.80	0.91 (0.01)
Female	Other	30	0.83	0.80	N/A
Female	Unknown	76	0.92	0.80	N/A
Female	White	16,858	0.80	0.80	0.91 (0.00)
Male	Asian	1,506	0.87	0.80	0.99 (0.01)
Male	Black	3,802	0.78	0.80	0.89 (0.01)
Male	Hispanic	3,240	0.78	0.80	0.89 (0.01)
Male	Other	51	0.76	0.80	N/A
Male	Unknown	129	0.84	0.80	N/A
Male	White	17,982	0.79	0.80	0.90 (0.00)
Unknown	Asian	762	0.89	0.80	N/A
Unknown	Black	1,337	0.81	0.80	0.93 (0.01)
Unknown	Hispanic	480	0.83	0.80	N/A
Unknown	Other	22	0.91	0.80	N/A
Unknown	Unknown	738	0.85	0.80	N/A
Unknown	White	2,024	0.82	0.80	0.93 (0.01)

* Data from this category constitutes less than 2% of the data and therefore was excluded from impact ratio calculations.

Recommendations

Quarterly audit with the next date set at July 31, 2026.

Demographic data collection: It is recommended to explore collecting candidate demographic data (gender and/or race/ethnicity) directly from employers and/or during the interview screening process. This would enable the collection of the most accurate and robust candidate demographic data, reducing reliance on name-based inference methods. The current quarter still relies on inferred demographic data, which remains a methodological limitation.

Underrepresented group analysis: The "Other" race category remains too small (103 candidates) to evaluate for adverse impact. It is recommended to investigate alternative approaches such as synthetic data generation, alternative demographic data providers, or cumulative multi-quarter analysis to enable meaningful assessment of these groups.

Bias Audit

Objective

The purpose of this Bias Audit is to audit the AI Products and Services of Upwage, specifically the AI Analyst product. The AI Interviewer conducts behavioral interviews to gather answers to a set of interview questions defined in collaboration with Upwage's customers to assess a set of professional competencies that are required for a given role. The AI Analyst assesses candidate interview transcripts, produced by the AI Interviewer, for the competencies defined by each customer.

The AI Interviewer produces a transcript of a text-based interview between the AI Interviewer and a given candidate. For each candidate transcript, the AI Analyst produces an assessment ranking for each competency of "Low", "Medium", or "High" as well as an overall ranking, also defined as "Low", "Medium" or "High". Both the AI Interviewer and AI Analyst are designed for human-centric use and not as automated decision-making tools. Human recruiters calibrate both the inputs (interview questions, competencies) and outputs (determining order of examination of candidate transcripts and AI Analyst results, outreach to candidates and other actions according to recruiter and business judgment and preferences).

We aim to obtain reasonable assurance that:

- The output of our AI systems and models are causing no adverse impact to our candidate populations
- Our processes and policies for identifying and mitigating bias remain at the forefront of the latest developments in the field
- When opportunities to further mitigate bias arise, we are tracking and acting on these opportunities in a timely manner.

Methodology

The Bias Audit is comprised of four major stages:

1. Scoping - Relevant materials and data are collected. Data is correctly formatted. (Week 1)

   During the Scoping phase, Upwage reviews the requirements of the audit as well as identifying the key systems and begins extracting the associated data for conducting the audit. This includes reviewing relevant definitions and requirements of existing and new AI legislation (Table 1).

Table 1. Relevant AI Legislation

Legislation	Definitions & Requirements	Source
NYC Local Law 144 in effect 7/5/23	Annual Bias Audit	NY Local Law 144
Colorado AI Act to be in effect 2/1/26	Annual Impact Assessment	Senate Bill 24-205

2. Data Cleaning and Analysis - Data is cleaned, if and where needed, and analyzed using the appropriate metrics for the type of data provided. (Week 1-2)

   During the Data Cleaning and Analysis phase, Upwage aims to methodologically understand any changes to the system that may impact the key information it uses and outputs it produces. Any changes are reflected in this report in the system description and Bias Audit objective and methodology.

3. Report - The Bias Audit report is curated on the audited system that includes updated system descriptions (if necessary), updated descriptions to Bias Audit objectives or scope (if necessary), impact ratios and recommendations. (Week 3-4)

   The Bias Audit report (this document) is generated outlining the details of the Products and Services being audited, the methodology used to conduct the audit, key findings, and recommendations. Findings are presented in terms of impact ratios. Upwage's auditing team then performs a review of the outcome of the audit using the four- fifths threshold[1] as a guide to determine whether exceptions should be further analyzed.

   https://www.eeoc.gov/laws/guidance/select-issues-assessing-adverse-impact-software-al gorithms-and-artificial

   Per the Equal Employment Opportunity Commission Guidelines :

   A selection rate for any race, sex, or ethnic group which is less than four-fifths (4/5) (or eighty percent) of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact, while a greater than four-fifths rate will generally not be regarded by Federal enforcement agencies as evidence of adverse impact. (29 CFR § 1607.4 - Information on impact.)[2]

Any analyses based on small sample sizes are indicated with an asterisk.

4. Ongoing Monitoring - The output of Upwage's AI systems and models are monitored regularly and re-audited at least quarterly or after major changes. (Week 4+)

   Given that AI and other automated systems and the legislation that regulates them regularly change and are updated as additional data and insights become available, Upwage audits its systems quarterly and as needed following any major updates or modifications or following the implementation of relevant mitigation procedures to examine their effectiveness at reducing bias or the risk of other relevant verticals. In the case of additional audits, the audit report and summary of results are reproduced to reflect the latest changes.

Data Extraction and Quality Assurance

The data and results herein disclosed are based on the following time periods and data samples. The data includes a diversity of hiring cycles, job roles, industries and geographical spread. This diversity is consistent with the model's total historical output. All outputs were captured during the current review period and reflect the current state and performance of the AI Analyst product. This data was selected to ensure a comprehensive and unbiased view of the model's performance over time while respecting customer confidentiality.

Time Period:	2026-02-01 to 2026-04-30

[2] https://www.law.cornell.edu/cfr/text/29/1607.4

Product Scope:	AI Analyst
Model Versioning:	gpt-4o and prompt version v3
Data Sample:	The data was generated by candidates interacting with our AI Interviewer which produces candidate transcripts. Those transcripts are then assessed using our AI Analyst product. The sample includes all production data produced from the AI Analyst during the time period, including a diversity of hiring cycles, job roles, industries and geographical spread. This diversity is consistent with the model's total historical output. All outputs were captured during the current review period and reflect the current state and performance of the tool. All test / non-candidate sessions were removed. Only sessions where the candidate completed the interview (sessionStatus = "completed") were included. Competency groups with fewer than 100 sessions were excluded due to lack of statistics. Total sessions = 58,717
Data Modifications:	None

Demographics

Our data sample (n=58,717) did not include candidate demographic information – specifically, race and gender/ethnicity. Having complete demographic information allows for a thorough and comprehensive analysis of potential biases. Without this data, any conclusions drawn could be incomplete or skewed, undermining the integrity of the audit. In particular, missing demographic information can introduce bias if the missing data is not randomly distributed. To reduce the risk of this type of bias, leading to more accurate and reliable results, we derive demographic information where possible for each candidate.
Using established datasets (e.g., UC Irvine ML Repository's Gender by Name, Harvard

Dataverse's Race and Ethnicity data) and verified methods (such as LLMs followed by manual review) ensures that our approach is systematic and replicable. Further, by clearly documenting our methods for deriving demographic information, we maintain transparency in our auditing process, allowing for scrutiny and validation by stakeholders.

We derived gender using each candidate's first name according to UC Irvine ML Repository's Gender by Name dataset[3]. We flagged names that did not exist or did not have a close match in the dataset and therefore gender could not be inferred. We asked an LLM (Claude-3-Opus) to derive gender based on the first name and manually reviewed the results. This procedure resulted in additional candidates with gender derived and remaining candidates that could not be matched via either approach. These unmatched candidates are referred to as "Unknown" in the audit results.

To derive race and ethnicity, we used Harvard Dataverse's Race and Ethnicity data[4]. We chose the most likely ethnicity for each last name. For Last names that do not exist in the dataset we asked an LLM (Claude-3-Opus) to derive their race and ethnicity and then manually reviewed the results. This procedure resulted in additional candidates with race and ethnicity derived and remaining candidates that could not be matched via either approach. These unmatched candidates are referred to as "Unknown" in the audit results. The candidates whose indicator is "Other" consist of American Indian, Pacific Islander, and Alaskan Indian ethnicities.

Selection Rate

Candidate's overall fit ratings are categorized as "High", "Medium" or "Low", where "High" is the strongest fit with the competencies that employers have defined for the role. Practically, employers tend to use these categories as a prioritization tool, considering "High" and "Medium" candidates before evaluating "Low" candidates.

The "Selection Rate" is defined as the ratio of candidates scoring "High" or "Medium" out of the total candidate population for a given group. The grouping of the top two categories together reflects typical use amongst employers, who may prioritize both "High" and "Medium" candidates above "Low" candidates.

Impact Ratio

The Impact ratio is the ratio of each demographic group's selection rate to the largest statistically significant selection rate (greater than 2% of population: 1,175 people). If the group has less than 2% of the total population (too small to be statistically significant) then the impact ratio is NA.

[3] Gender by Name. (2020). UCI Machine Learning Repository. https://doi.org/10.24432/C55G7X.

[4] Rosenman, Evan; Olivella, Santiago; Imai, Kosuke, 2022, "Race and ethnicity data for first, middle, and last names", https://doi.org/10.7910/DVN/SGKW0K, Harvard Dataverse, V9, UNF:6:Z4OdPbRiTIYpwYm8CCktow== [fileUNF]

Bias Audit Results

The "Unknown" category includes candidates where gender and/or race/ethnicity information could not be inferred. The "Other" category includes all other race categories, including Native Hawaiian or Other Pacific Islander, American Indian or Alaska Native and Two or more races.

Non-intersectional, Gender, sorted by Scoring rate

Gender	Candidates	Selection Rate	EEOC Impact RatioThreshold	Impact Ratio (Error)
Female	26,644	0.81	0.80	1.00 (0.00)
Male	26,710	0.79	0.80	0.98 (0.00)
Unknown	5,363	0.83	0.80	1.03 (0.01)

Non-intersectional, Race/Ethnicity

Race/Ethnicity	Candidates	Selection Rate	EEOC Impact RatioThreshold	Impact Ratio (Error)
Asian	3,641	0.88	0.80	1.00 (0.01)
Black	10,389	0.81	0.80	0.92 (0.00)
Hispanic	6,777	0.79	0.80	0.90 (0.01)
Other	103	0.82	0.80	**N/A***
Unknown	943	0.86	0.80	**N/A***
White	36,864	0.80	0.80	0.91 (0.00)

Intersectional, Gender and Race/Ethnicity

Gender	Race/ Ethnicity	Candidates	Selection Rate	EEOC Impact Ratio Threshold	Impact Ratio (Error)
Female	Asian	1,373	0.88	0.80	1.00 (0.01)
Female	Black	5,250	0.83	0.80	0.95 (0.01)
Female	Hispanic	3,057	0.80	0.80	0.91 (0.01)
Female	Other	30	0.83	0.80	N/A
Female	Unknown	76	0.92	0.80	N/A
Female	White	16,858	0.80	0.80	0.91 (0.00)
Male	Asian	1,506	0.87	0.80	0.99 (0.01)

Male	Black	3,802	0.78	0.80	0.89 (0.01)
Male	Hispanic	3,240	0.78	0.80	0.89 (0.01)
Male	Other	51	0.76	0.80	N/A
Male	Unknown	129	0.84	0.80	N/A
Male	White	17,982	0.79	0.80	0.90 (0.00)
Unknown	Asian	762	0.89	0.80	N/A
Unknown	Black	1,337	0.81	0.80	0.93 (0.01)
Unknown	Hispanic	480	0.83	0.80	N/A
Unknown	Other	22	0.91	0.80	N/A
Unknown	Unknown	738	0.85	0.80	N/A
Unknown	White	2,024	0.82	0.80	0.93 (0.01)

* Data from this category constitutes less than 2% of the data and therefore was excluded from impact ratio calculations.

Legislation Tracker

Our Legislation Tracker is regularly updated here.

New legislation identified:

Green = Legislation identified as relevant to Upwage's AI Products and Services; flagged for ongoing monitoring

Legislation	Status	Scope	Relevant Stipulations
Texas Responsible Artificial Intelligence Governance Act (TRAIGA), HB 14	In effect, Jan. 1, 2026	AI developers and deployers offering or using AI systems in Texas	Defines both **developers** and **deployers** of AI systems used in Texas; prohibits developing or deploying AI with the intent to unlawfully discriminate against a protected class; Texas AG enforcement; cure process; references recognized AI risk management frameworks.

California AB 2013, Generative AI Training Data Transparency	In effect, Jan. 1, 2026	Developers making generative AI systems or services available to Californians	Requires covered developers to post website documentation about training data before making a generative AI system, service, or substantial modification publicly available to Californians; required disclosures include dataset summaries, sources/owners, whether copyrighted data is included, whether personal information is included, and whether datasets were purchased or licensed.

New Research Findings

1. Evaluating LLM Behavior in Hiring: Implicit Weights, Fairness Across Groups, and Alignment with Human Preferences

(January 2026; Hoffmann, Jouffroy, Jouanneau, Palyart & Pebereau; Affiliation not listed on arXiv page; Source: arXiv; Article: arXiv:2601.11379)

This study suggests LLMs used in hiring may apply different implicit weights to candidate signals across demographic groups, even when average discrimination appears limited.

Using synthetic datasets built from real freelancer profiles and project descriptions from a major European online freelance marketplace, the authors estimated how an LLM weighed match-relevant criteria when evaluating freelancer-project fit. They found that the model emphasized core productivity signals such as skills and experience, but also interpreted some features beyond their explicit matching value. While average discrimination against minority groups appeared limited, the authors found intersectional effects, with productivity signals carrying different weights across demographic groups.

Aggregate fairness metrics may miss risks that only appear when you look at how the same qualification signals are valued across subgroups.

2. Small Changes, Big Impact: Demographic Bias in LLM-Based Hiring

Through Subtle Sociocultural Markers in Anonymised Resumes

(March 2026; Tan, Khoo, Doan, Liu, Chen & Lee; Affiliation not listed on arXiv page; Source: arXiv; Article: arXiv:2603.05189)

This study suggests LLM-based hiring systems may still produce demographic bias even when explicit personal identifiers are removed from resumes.

The authors introduced a hiring fairness stress-test using 100 neutral job-aligned resumes expanded into 4,100 variants spanning four ethnicities and two genders, differing only in job-irrelevant sociocultural markers. They evaluated 18 LLMs in direct comparison and score-and-shortlist settings and found that, even without explicit identifiers, models could infer demographic attributes and showed systematic disparities, including favoritism toward markers associated with Chinese and Caucasian male candidates. The study also found that prompting models to provide explanations could amplify bias rather than reduce it.

Removing explicit personal identifiers may not be enough to prevent bias if other seemingly neutral resume details continue to act as demographic proxies.

3. Human, Algorithm, or Both? Gender Bias in Human-Augmented Recruiting

(March 2026; Kaya & Bogers; Affiliation not listed on arXiv page; Source: arXiv; Article: arXiv:2603.06240)

This study suggests human oversight may improve fairness outcomes when AI is used in recruiting workflows.

The authors conducted a quantitative analysis of gender bias across three scenarios on a real-world recruitment platform: recruiters manually searching a CV database, AI-driven matching between candidates and jobs, and a combined human-plus-AI approach. They found that human recruiters produced fairer candidate lists than the AI-only solution, and that the hybrid approach produced the fairest candidate lists overall. In particular, interacting with AI recommendations first and then manually searching for additional candidates appeared to improve gender fairness in the candidates who were viewed, clicked, and contacted.

Human oversight may be an important control for reducing bias in AI-assisted recruiting, particularly when AI recommendations influence who is reviewed or contacted.

Incidents Tracker

We report incidents where Upwage's AI Products or Services has caused or is reasonably likely to have caused algorithmic discrimination within 90 days.

No incidents to report.

Revisions

Date	Editor	Changes
2025-10-20	Taylor McLoughlin	Initial draft
2026-02-06	Taylor McLoughlin	Initial publication
2026-04-03	Santiago Leon	Quarterly Update
2026-05-25	Santiago Leon	Quarterly Update