Warden AI

ACME AI Audit Report

Generated from
ACME AI Assurance Dashboard

April 29, 2026

Table of Contents

Report summary

Warden AI is engaged by ACME to perform independent, ongoing bias audits of the CV Analysis system. This report summarizes the most recent audit, generated using Warden AI's assurance platform and reviewed by our audit team.

The audit is structured according to Warden's bias audit framework, which is designed to align with emerging regulatory and industry standards that emphasize the role of anti-bias testing in mitigating discrimination risks related to protected characteristics. Warden's approach supports these aims while promoting fairness and accountability in automated decision-making.

The system was evaluated using a purpose-built test dataset, as sufficient historical data was not available. Multiple bias-detection techniques were applied to evaluate the system's behavior on this dataset at the time of testing.

This report is provided for demonstration purposes only. It reflects the system's behavior under controlled test conditions at the time of evaluation and should not be interpreted as a formal certification or compliance determination. The findings do not represent bias outcomes for any specific employer, job opportunity, or real-world deployment.

Audit information

System audited:	ACME - CV Analysis
Audit frequency:	Monthly
Latest audit date:	April 29, 2026
Sample size:	31,524

Report summary

Results overview

Group	Group bias	Individual bias
Sex bias	Clear	Clear
Race/Ethnicity bias	Clear	Clear
Intersectional (Sex X Race/Ethnicity) bias	Clear	Clear
Age bias	Clear	Clear
Disability bias	Clear	Clear
Religion bias	Clear	Clear
Sexual orientation bias	Clear	Clear
Veteran status bias	Clear	Clear
National origin bias	Clear	Clear
Pregnancy status bias	Clear	Clear
English language proficiency bias	Clear	Clear
Criminal history bias	Clear	Clear
Medical conditions bias	Clear	Clear
Marital status bias	Clear	Clear
Gender identity bias	Clear	Clear

About Warden AI

Company summary

At Warden AI, our mission is to reduce societal discrimination through fair and transparent AI. We provide third-party oversight into AI systems, building trust and increasing adoption.

We are an independent AI auditor and assurance platform that performs ongoing audits to ensure AI systems are fair, explainable, and transparent. Our team brings extensive experience across AI, regulation, and research, including industry and academia, to deliver our solution.

Our system integrates with the AI system that is under test, allowing for continuous testing and monitoring. Our methodology employs a combination of bias detection techniques and uses our proprietary datasets and/or historical data from the system.

Independence statement

Warden Technologies, Inc. is an independent AI audit and assurance provider. Fees associated with our service are solely for our evaluation and their payment is not related to the outcome of the results.

Our services are strictly limited to testing and monitoring the trustworthiness of AI systems. We do not form part of the solution or in any way affect how the system under test works.

Company information

Registered address:
600 Congress Ave, 14th Floor,
Austin, TX 78701, United States

Website:
https://warden-ai.com

Contact:
contact@warden-ai.com

Audit scope

AUDIT DATE
April 29, 2026

AUDIT FREQUENCY
Monthly

DATA SOURCE
Warden dataset

SAMPLE SIZE
31,524

TEST INPUTS
Job criteria, Candidate profile

TEST OUTPUTS
Match score

BIAS DETECTION TECHNIQUES
2

PROTECTED CLASSES
15

System details

ACME's CV Analysis system uses AI to streamline hiring, cut costs, and ensure great talent isn't overlooked. It automates applicant screening and generates a shortlist of qualified candidates.

Profiles are evaluated against job requirements based on experience, skills, and preferences. Each candidate is assessed individually using the same criteria, producing an overall match score that reflects how well they fit the role.

Data details

This audit used Warden's purpose-built test dataset to evaluate system behavior in a controlled setting. Historical candidate data with sufficient demographic coverage was not available.

The dataset contains real resumes collected through multiple sources, specifically curated for bias and fairness evaluation. All resumes are used with an appropriate legal basis, such as explicit participant consent. Demographic information such as the profile's sex is self-reported and standard data quality checks are applied before use.

The dataset also includes job criteria covering a wide range of occupations. These are provided either as full job descriptions or as structured criteria, such as skill requirements or evaluation rubrics. Resumes and job criteria are matched at the occupational level to avoid assessing unrelated roles.

Audit scope

To isolate system behavior and minimize test-data bias, each resume is transformed into qualification-preserving counterfactuals for the tested demographic groups. These variants modify demographic indicators and related proxies while holding skills and experience constant, enabling clearer attribution of outcome differences to the system. Where applicable, these transformations are informed by public statistics and academic research.

Audit details

Audits are performed on a regular cadence to continuously monitor potential bias risks. The most recent audit was completed on *April 29, 2026* and evaluates potential bias across the following 15 protected classes:

- Sex, Race/Ethnicity, Intersectional (Sex X Race/Ethnicity), Age, Disability, Religion, Sexual orientation, Veteran status, National origin, Pregnancy status, English language proficiency, Criminal history, Medical conditions, Marital status, and Gender identity

The following bias detection techniques were applied:

- **Disparate impact analysis:** This analysis evaluates if the system is adversely impacting a protected group compared to another protected group. An impact ratio of 0.8 or higher is considered acceptable.
- **Counterfactual analysis:** This analysis evaluates if the system produces consistent results when certain demographic attributes or proxies are altered. A consistency score of 0.95 or higher is considered acceptable.

For the disparate impact analysis, the scoring rate method was applied. Scores above the median score were classified as selected outcomes, while scores below the median were classified as not selected.

Audit results

Sex bias

Group bias (Disparate impact analysis)

Result: Clear **Sample size:** 1,824

Group	Samples	Selected	Scoring rate	Impact ratio	Result
Female	912	350	38.38%	1.00	Clear
Male	912	347	38.05%	0.99	Clear

The results indicate equitable outputs across all groups.

Individual bias (Counterfactual analysis)

Result: Clear **Sample size:** 1,824

Group	Samples	Consistency score	Result
Female	912	100.00%	Clear
Male	912	99.46%	Clear

The results indicate highly consistent outputs across all groups.

Audit results

Race/Ethnicity bias

Group bias (Disparate impact analysis)

Result: Clear **Sample size:** 1,824

Group	Samples	Selected	Scoring rate	Impact ratio	Result
Asian	456	173	37.94%	0.98	Clear
Black or African American	456	173	37.94%	0.98	Clear
Hispanic or Latino	456	174	38.16%	0.98	Clear
White	456	177	38.82%	1.00	Clear

The results indicate equitable outputs across all groups.

Individual bias (Counterfactual analysis)

Result: Clear **Sample size:** 1,824

Group	Samples	Consistency score	Result
Asian	456	99.77%	Clear
Black	456	99.52%	Clear
Hispanic	456	99.71%	Clear
White	456	100.00%	Clear

The results indicate highly consistent outputs across all groups.

Audit results

Intersectional (Sex X Race/Ethnicity) bias

Group bias (Disparate impact analysis)

Result: Clear **Sample size:** 1,824

Group	Samples	Selected	Scoring rate	Impact ratio	Result
Asian / Female	228	88	38.60%	0.99	Clear
Asian / Male	228	85	37.28%	0.95	Clear
Black / Female	228	88	38.60%	0.99	Clear
Black / Male	228	85	37.28%	0.95	Clear
Hispanic / Female	228	86	37.72%	0.97	Clear
Hispanic / Male	228	88	38.60%	0.99	Clear
White / Female	228	88	38.60%	0.99	Clear
White / Male	228	89	39.04%	1.00	Clear

The results indicate equitable outputs across all groups.

Audit results

Individual bias (Counterfactual analysis)

Result: Clear **Sample size:** 1,824

Group	Samples	Consistency score	Result
Asian / Female	228	99.36%	Clear
Asian / Male	228	99.04%	Clear
Black / Female	228	99.68%	Clear
Black / Male	228	98.23%	Clear
Hispanic / Female	228	98.76%	Clear
Hispanic / Male	228	99.53%	Clear
White / Female	228	100.00%	Clear
White / Male	228	98.86%	Clear

The results indicate highly consistent outputs across all groups.

Audit results

Age bias

Group bias (Disparate impact analysis)

Result: Clear **Sample size:** 268

Group	Samples	Selected	Scoring rate	Impact ratio	Result
40 or over	134	60	44.78%	1.00	Clear
Under 40	134	60	44.78%	1.00	Clear

The results indicate equitable outputs across all groups.

Individual bias (Counterfactual analysis)

Result: Clear **Sample size:** 268

Group	Samples	Consistency score	Result
40 or over	134	99.68%	Clear
Under 40	134	100.00%	Clear

The results indicate highly consistent outputs across all groups.

Audit results

Disability bias

Group bias (Disparate impact analysis)

Result: Clear **Sample size:** 912

Group	Samples	Selected	Scoring rate	Impact ratio	Result
Has disability	456	220	48.25%	0.97	Clear
No disability	456	226	49.56%	1.00	Clear

The results indicate equitable outputs across all groups.

Individual bias (Counterfactual analysis)

Result: Clear **Sample size:** 912

Group	Samples	Consistency score	Result
Has disability	456	99.77%	Clear
No disability	456	100.00%	Clear

The results indicate highly consistent outputs across all groups.

Audit results

Religion bias

Group bias (Disparate impact analysis)

Result: Clear **Sample size:** 2,052

Group	Samples	Selected	Scoring rate	Impact ratio	Result
Buddhism	228	88	38.60%	0.97	Clear
Christianity	228	91	39.91%	1.00	Clear
Hinduism	228	83	36.40%	0.91	Clear
Indigenous/ Traditional	456	170	37.28%	0.93	Clear
Islam	228	87	38.16%	0.96	Clear
Judaism	228	85	37.28%	0.93	Clear
No religion	228	82	35.96%	0.90	Clear
Sikhism	228	88	38.60%	0.97	Clear

The results indicate equitable outputs across all groups.

Audit results

Individual bias (Counterfactual analysis)

Result: Clear **Sample size:** 2,052

Group	Samples	Consistency score	Result
Buddhism	228	99.56%	Clear
Christianity	228	99.71%	Clear
Hinduism	228	99.39%	Clear
Indigenous/Traditional	456	99.29%	Clear
Islam	228	99.94%	Clear
Judaism	228	99.00%	Clear
No religion	228	97.92%	Clear
Sikhism	228	100.00%	Clear

The results indicate highly consistent outputs across all groups.

Audit results

Sexual orientation bias

Group bias (Disparate impact analysis)

Result: Clear **Sample size:** 1,140

Group	Samples	Selected	Scoring rate	Impact ratio	Result
Bi/Pansexual	684	255	37.28%	0.97	Clear
Heterosexual	228	88	38.60%	1.00	Clear
Homosexual	228	81	35.53%	0.92	Clear

The results indicate equitable outputs across all groups.

Individual bias (Counterfactual analysis)

Result: Clear **Sample size:** 1,140

Group	Samples	Consistency score	Result
Bi/Pansexual	684	99.64%	Clear
Heterosexual	228	99.87%	Clear
Homosexual	228	100.00%	Clear

The results indicate highly consistent outputs across all groups.

Audit results

Veteran status bias

Group bias (Disparate impact analysis)

Result: Clear **Sample size:** 684

Group	Samples	Selected	Scoring rate	Impact ratio	Result
Nonveteran	456	176	38.60%	0.90	Clear
Veteran	228	98	42.98%	1.00	Clear

The results indicate equitable outputs across all groups.

Individual bias (Counterfactual analysis)

Result: Clear **Sample size:** 684

Group	Samples	Consistency score	Result
Nonveteran	456	98.86%	Clear
Veteran	228	100.00%	Clear

The results indicate highly consistent outputs across all groups.

Audit results

National origin bias

Group bias (Disparate impact analysis)

Result: Clear **Sample size:** 4,104

Group	Samples	Selected	Scoring rate	Impact ratio	Result
Africa and Middle East	684	251	36.70%	0.94	Clear
Asia and Pacific	912	358	39.25%	1.00	Clear
Central and South America (Non-Hispanic)	684	248	36.26%	0.92	Clear
Europe (Except Spain)	912	353	38.71%	0.99	Clear
Hispanic	456	168	36.84%	0.94	Clear
North America (Non-Hispanic)	456	179	39.25%	1.00	Clear

The results indicate equitable outputs across all groups.

Individual bias (Counterfactual analysis)

Result: Clear **Sample size:** 4,104

Group	Samples	Consistency score	Result
Africa and Middle East	684	99.78%	Clear
Asia and Pacific	912	100.00%	Clear
Central and South America (Non-Hispanic)	684	99.65%	Clear
Europe (Except Spain)	912	99.90%	Clear
Hispanic	456	99.92%	Clear
North America (Non-Hispanic)	456	100.00%	Clear

The results indicate highly consistent outputs across all groups.

Audit results

Pregnancy status bias

Group bias (Disparate impact analysis)

Result: Clear **Sample size:** 684

Group	Samples	Selected	Scoring rate	Impact ratio	Result
Not pregnant	456	172	37.72%	1.00	Clear
Pregnant	228	84	36.84%	0.98	Clear

The results indicate equitable outputs across all groups.

Individual bias (Counterfactual analysis)

Result: Clear **Sample size:** 684

Group	Samples	Consistency score	Result
Not pregnant	456	100.00%	Clear
Pregnant	228	99.88%	Clear

The results indicate highly consistent outputs across all groups.

Audit results

English language proficiency bias

Group bias (Disparate impact analysis)

Result: Clear **Sample size:** 684

Group	Samples	Selected	Scoring rate	Impact ratio	Result
Conversational	228	96	42.11%	1.00	Clear
Limited	228	79	34.65%	0.82	Clear
Native/Bilingual	228	95	41.67%	0.99	Clear

The results indicate equitable outputs across all groups.

Individual bias (Counterfactual analysis)

Result: Clear **Sample size:** 684

Group	Samples	Consistency score	Result
Conversational	228	100.00%	Clear
Limited	228	98.61%	Clear
Native/Bilingual	228	99.78%	Clear

The results indicate highly consistent outputs across all groups.

Audit results

Criminal history bias

Group bias (Disparate impact analysis)

Result: Clear **Sample size:** 908

Group	Samples	Selected	Scoring rate	Impact ratio	Result
Has criminal history	454	177	38.99%	1.00	Clear
No criminal history	454	164	36.12%	0.93	Clear

The results indicate equitable outputs across all groups.

Individual bias (Counterfactual analysis)

Result: Clear **Sample size:** 908

Group	Samples	Consistency score	Result
Has criminal history	454	100.00%	Clear
No criminal history	454	98.94%	Clear

The results indicate highly consistent outputs across all groups.

Audit results

Medical conditions bias

Group bias (Disparate impact analysis)

Result: Clear **Sample size:** 908

Group	Samples	Selected	Scoring rate	Impact ratio	Result
Has medical history	454	222	48.90%	1.00	Clear
No medical history	454	218	48.02%	0.98	Clear

The results indicate equitable outputs across all groups.

Individual bias (Counterfactual analysis)

Result: Clear **Sample size:** 908

Group	Samples	Consistency score	Result
Has medical history	454	100.00%	Clear
No medical history	454	99.88%	Clear

The results indicate highly consistent outputs across all groups.

Audit results

Marital status bias

Group bias (Disparate impact analysis)

Result: Clear **Sample size:** 910

Group	Samples	Selected	Scoring rate	Impact ratio	Result
Married / In a legally recognized partnership	454	176	38.77%	0.99	Clear
Not married	456	178	39.04%	1.00	Clear

The results indicate equitable outputs across all groups.

Individual bias (Counterfactual analysis)

Result: Clear **Sample size:** 910

Group	Samples	Consistency score	Result
Married / In a legally recognized partnership	454	100.00%	Clear
Not married	456	99.86%	Clear

The results indicate highly consistent outputs across all groups.

Audit results

Gender identity bias

Group bias (Disparate impact analysis)

Result: Clear **Sample size:** 684

Group	Samples	Selected	Scoring rate	Impact ratio	Result
Female	228	88	38.60%	0.92	Clear
Male	228	88	38.60%	0.92	Clear
Nonbinary / Other	228	96	42.11%	1.00	Clear

The results indicate equitable outputs across all groups.

Individual bias (Counterfactual analysis)

Result: Clear **Sample size:** 684

Group	Samples	Consistency score	Result
Female	228	99.14%	Clear
Male	228	99.24%	Clear
Nonbinary / Other	228	100.00%	Clear

The results indicate highly consistent outputs across all groups.

Methodology

Methodology overview

Our methodology for evaluating AI systems is designed to ensure fairness and transparency. Our comprehensive approach includes ongoing auditing, multiple bias detection techniques, the use of diverse datasets, and human oversight.

This rigorous approach enables us to accurately report on the level of bias in the system and build trust with the system's users and stakeholders.

Black box testing

We use black-box testing techniques to evaluate AI systems. This approach examines the system's outputs in response to specific inputs without needing to understand the internal workings.

This enables us to make systematic judgements across different AI systems with different underlying models.

Ongoing audits

AI systems change frequently (often monthly, weekly, or even daily). Our audits are performed on a regular basis at the frequency detailed in this report. The exact frequency is determined with the AI provider based on the nature of their system and their propensity for product updates.

In addition to the scheduled evaluations, the AI provider can also choose to have an audit performed on-demand between scheduled audits if they have a significant product update.

Multiple bias detection techniques

Our bias detection techniques include both disparate impact analysis and counterfactual analysis. These methods help identify any potential biases in the system by comparing the outcomes for different demographic groups and testing hypothetical scenarios where demographic attributes are altered.

Including both techniques ensures a more comprehensive evaluation, as they provide complementary insights into the system's fairness.

Methodology

Hybrid auditing

Our evaluation process combines automated methods with human oversight to ensure accuracy and reliability.

By integrating AI systems with our standardized datasets, we can conduct large-scale and frequent audits. This approach is complemented by human-led data curation and quality assurance processes for creating the datasets. Additionally, our team of experts reviews and validates the results of audits to ensure reliability.

Diverse datasets

Our auditing framework uses a mixture of data. We have our own proprietary datasets which provide an independent benchmark of the AI system. Our dataset is formed of real data sourced from real people where consent has been provided.

This dataset is augmented with 'counterfactual' samples which involves synthetic modifications to demographic attributes within real profiles. Where applicable, we also use both historical and live data to provide context for the system's long-term performance and its current real-time operations.

All datasets are ethically sourced and we adhere to high standards of data collection practices. We are committed to maintaining confidentiality and protecting personal data. Some of our evaluations require datasets that contain elements of personal information to test specific AI functionalities. In such instances, we ensure that consent has been explicitly obtained for the use of this information.

Methodology

Disparate impact analysis

Disparate Impact Analysis evaluates whether a protected demographic group is adversely affected compared to other groups. This is achieved by comparing its scoring rate to the highest scoring group. The goal is to ensure that the AI system does not disproportionately disadvantage any specific group based on inherent characteristics such as race, ethnicity, or sex.

Scoring rate

Scoring rate is a measure used to evaluate the proportion of individuals in a specific group who receive favorable outcomes from the AI system.

To calculate a group's scoring rate, we divided the number of individuals who received a score above the sample's median score by the total number of individuals with the group.

$$\text{Scoring rate} = \frac{\text{Number of individuals within group with score above the sample's median score}}{\text{Total number of individuals within group}}$$

Impact ratio

The impact ratio is a metric used to measure potential adverse impact on a group by comparing its scoring rate to the highest scoring group.

$$\text{Impact ratio} = \frac{\text{Scoring rate for the group}}{\text{Scoring rate of the highest scoring group}}$$

An impact ratio of 1 indicates no adverse impact, whereas a lower ratio indicates a higher likelihood of adverse impact. According to the four-fifths rule, an impact ratio of 0.8 (80%) or higher is considered acceptable, indicating that the AI system's outcomes are equitable across different demographic groups.

Methodology

Counterfactual analysis

Counterfactual Analysis is a method used to assess the fairness of AI systems by examining how the system's decisions would change if certain demographic attributes or proxies of individuals were altered.

This approach helps determine whether the AI system's outcomes are influenced by these attributes, ensuring that individuals receive similar treatment regardless of their demographic characteristics.

Counterfactual scenarios

A counterfactual scenario involves modifying specific demographic attributes or proxies of an individual while keeping all other aspects of their profile unchanged.

For instance, if an individual profile is female, a counterfactual scenario would involve changing the gender proxies contained within the profile to male, while maintaining all other information (such as qualifications, experience, and skills) exactly the same.

This allows us to isolate the impact of the demographic attribute on the AI system's decision.

Counterfactual analysis process

Generate	Counterfactual samples are generated using our proprietary system combined with human spot checking. This involves changing profile information and demographic proxies within the input sample.
Execute	The original samples and counterfactual samples are run against the AI system being tested.
Measure	A "consistency score" is calculated that measures the relative consistency between the counterfactual samples and original samples. A higher score means the system's decisions are less influenced by demographic attributes.
Review	Our team of AI auditors perform spot checks on the AI system and reviews the interpretation of the results.

Methodology

Grading system

Results are evaluated using a grading system that indicates the severity of any bias detected during the audit.

Grades follow a 'traffic light' model: **Clear** indicates no issues found, **Consider** indicates potential or minor concerns requiring review, and **Concern** indicates significant issues requiring attention.

Grade	Description	Disparate impact analysis	Counterfactual analysis
Clear	No issues detected	Impact ratio ≥ 80%	Consistency score ≥ 95%
Consider	Potential or minor issue(s) detected	60% ≤ Impact ratio < 80%	90% ≤ Consistency score < 95%
Concern	Definite or major issue(s) detected	Impact ratio < 60%	Consistency score < 90%

Disclaimer

This report has been prepared by Warden Technologies, Inc. to provide an independent audit of the AI system developed by the AI provider in question, based on our proprietary methodologies and datasets. The results and conclusions presented in this report reflect our best judgments derived from the information available at the time of evaluation. While we strive for accuracy and completeness, we cannot guarantee that our evaluation is exhaustive or that there are no errors.

Our methodology is designed to identify potential issues of bias and other trust factors in the AI system under examination. However, our approach, like any evaluation methodology, has its limitations. It is important to understand that our findings do not guarantee the absence of any bias, flaws, or limitations within the audited AI system. Instead, they indicate that, based on our specific testing framework and within the scope of our analysis, no significant issues were identified.

This report is intended for informational purposes only and should not be interpreted as a guarantee of the system's performance, fairness, or suitability for any specific purpose or use case. Warden Technologies, Inc. disclaims any liability for any decisions made or actions taken based on the information provided in this report. By using this report, the reader agrees to assume all risks associated with such decisions or actions and agrees to hold Warden Technologies, Inc. harmless against any claims, damages, or liabilities that may arise from the use of the evaluated AI system.

Warden AI

ACME AI Audit Report