Warden AI

# Popp AI Trust Assurance Report

**Generated from**
Popp AI Trust Center

March 6, 2024

# Table of Contents

# Report Summary

This report is generated from the Warden AI assurance platform, through which we serve as an independent evaluator of AI trust and safety. Through continuous verification processes, our platform assesses AI systems to ensure their reliability and integrity.

Contained within are the outcomes of our recent evaluations. For the most up-to-date results, please visit the Popp AI Trust Center.

## Evaluation information

| | |
|---|---|
| **System tested:** | *Popp AI - CV Analysis* |
| **Evaluation frequency:** | *Monthly* |
| **Latest evaluation date:** | *March 06 2024* |
| **Next evaluation date:** | *April 06, 2024* |
| **Test samples:** | *5,900* |


## Results summary

### Robustness

| *Check* | *Grade* |
|---|---|
| **Idempotency** | **Good** |
| **Paraphrasing sensitivity** | **Excellent** |
| **Reordering sensitivity** | **Good** |
| **Irrelevant information** | **Excellent** |

### Bias and Fairness

| *Check* | *Grade* |
|---|---|
| **Gender bias** | **Excellent** |
| **Ethnicity bias** | **Excellent** |

# About Warden AI

## Company summary

At Warden AI, our mission is to safeguard the deployment of AI systems to ensure they are fair and trustworthy to use. We provide an AI assurance platform that continuously evaluates AI systems to ensure they are fair, robust, and explainable.

Our evaluations are independent. We use our own proprietary datasets and evaluation techniques to conduct our testing framework. In contrast to other assurance and audit techniques in the market, we do not evaluate whether the AI developer has followed processes correctly. Instead, our sole concern is whether the end-product (i.e. the AI system under test) demonstrates trustworthy behaviour or not.

Our team brings extensive experience across AI, regulation, and research, including industry and academia, to build our innovative and trustworthy solution.

There is more information about our methodology in another section below.

## Company information

**Registered address:**
Warden AI Ltd, 71-75 Shelton Street, London WC2H 9JQ, United Kingdom

**Registered company number:**
15321282

**Website:**
https://warden-ai.com

**Contact:**
contact@warden-ai.com

# About Warden AI

## Independence statement

Warden AI Ltd is an independent AI safety evaluator. 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 evaluation and monitoring of AI safety within AI systems. We do not form part of the solution or in any way affect how the system under test works.

The nature of our evaluation methods are the same for all systems of the same use-case that we test, and we do not customise our service for each system that we evaluate.

# System and Evaluation Details

## System tested

**Name:**
*Popp AI - CV analysis*

**Description:**
*Popp AI's CV analysis* product is an AI system that assesses the fit of a list of job candidates against specified job criteria. The candidates are ranked in order of the score they have received from the system.

**Inputs:**
- Candidate profile: either CV file or LinkedIn user profile
- Job description

**Outputs:**
- Score (0 to 100)

## Evaluation details

| | |
|---|---|
| **Recurring evaluation** | Evaluations are performed on a regular basis because AI systems are subject to frequent changes. |
| **Evaluation frequency** | Monthly |
| **Latest evaluation** | March 6th, 2024 |
| **Next evaluation** | April 6th, 2024 |
| **Integration** | Bulk upload/download of Warden's testing dataset to the system's production environment. |

# Results

**Verifications performed**:
*Robustness, Bias and Fairness*

**Test samples:**
*5,900*

## Robustness verification

Evaluates the system's ability to handle various inputs consistently, ensuring reliability across different scenarios. It measures how sensitive the system is to minor variations in input.

### Idempotency check

Measures how consistently the system behaves when the same input is given multiple times, indicating the overall level of noise / randomness within the system.

| Grade: | Score: | Test samples: |
|---|---|---|
| *Good* | *90.1%* | *590* |

*The system demonstrates some noise, with a minor level of variance in output scores for the same input.*

### Paraphrasing sensitivity check

Evaluates the system's ability to understand and process variations in language, such as different phrasings or synonyms, without significant changes in output.

| Grade: | Score: | Test samples: |
|---|---|---|
| *Excellent* | *96.0%* | *590* |

*The system is largely resilient to linguistic variations, ensuring reliable user experiences.*

### Reordering sensitivity check

Evaluates how the system responds to differences in the order of content, reflecting its consistency in handling varying inputs .

| Grade: | Score: | Test samples: |
|---|---|---|
| *Good* | *94.2%* | *590* |

*The system is largely unaffected by changes in the sequence of input, indicating robustness to ordering.*

# Results

## Irrelevant information check

Evaluates the system's capacity to disregard non-essential or unrelated information in inputs, focusing on relevant data to generate its output.

| Grade: | Score: | Test samples: |
|---|---|---|
| *Excellent* | *97.2%* | *590* |

*The system effectively filters out irrelevant information, maintaining accuracy in its outputs.*

# Bias and Fairness Verification

Evaluates the system's consistency across different demographic groups. It measures how sensitive the system is to changes in demographic identifiers contained within the input.

## Gender bias check

Evaluates the system's consistency across different genders by modifying gender identifiers within profiles.

| Grade: | Score: | Test samples: |
|---|---|---|
| *Excellent* | *95.7%* | *1180* |

| | |
|---|---|
| **Female** | Reference group |
| **Male** | 95.7% |

*The test results indicate mostly consistent outputs across genders, potentially with a very minor bias in favour of female gender profiles.*

# Results

## Bias and Fairness Verification

### Ethnicity bias check

Evaluates the system's consistency across different ethnicities by modifying ethnic identifiers within profiles.

| Grade: | Score: | Test samples: |
|---|---|---|
| *Excellent* | *96.3%* | *2360* |

| | |
|---|---|
| **Asian** | 96.0% |
| **Black** | Reference group |
| **Hispanic** | 96.4% |
| **White** | 96.4% |

*The test results indicate mostly consistent outputs across ethnicities.*

# Methodology

## AI Assurance

AI assurance is the process of measuring, evaluating, and communicating the trustworthiness of AI systems. By building trust in AI systems, AI assurance plays a crucial role in enabling the responsible development and deployment of AI, unlocking both the economic and social benefits of AI systems.

We evaluate AI systems and verify that they are trustworthy across a range of trust factors. The trust factors we currently verify are robustness and bias.

## Black box testing

We use black-box testing techniques to perform our verifications. Black-box testing examines how a system responds to inputs without delving into internal mechanisms. This enables us to make systematic judgements about whether the system under test is trustworthy across the trust factors we evaluate in our role as an independent evaluator.

This approach also means that our framework is model-agnostic: it applies equally well across all types of AI or algorithmic decision-making systems.

## Application testing

An AI system is typically composed of one or multiple AI models, pre- and post-processing modules, and sits within a context-specific use-case that has end-user impact.

Our verifications are performed on the AI application (or system) as a whole; we do not just evaluate specific AI models. It is important to test the AI application and not just the model because risks with AI systems manifest in real-world use-cases within the application layer.

# Methodology

## On-going evaluation

AI systems change frequently (often monthly, weekly, or even daily). This is particularly important if the system uses a foundational model (such as a public Large Language Model like ChatGPT) because the underlying models can change without the developer intending them to.

Therefore, regular, on-going evaluation of an AI system is essential to ensure its trusthworthiness. Annual evaluations or audits are not sufficient to achieve this.

Our evaluations are performed on a recurring basis at the frequency detailed in this report. The exact frequency of our evaluations 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 evaluation performed on-demand between scheduled evaluations if they have a significant product update.

# Methodology

## Evaluation technique

We have a suite of evaluation techniques that are used for different systems and use-cases. The main technique used for this system is *counterfactual analysis* (also known as *counterfactual fairness* in the context of bias/fairness).

### Counterfactual fairness

Counterfactual fairness assumes a decision is fair towards an individual if the outcome is the same in reality as it would be in a 'counterfactual' world, in which the individual belongs to a different demographic group.

In this way, counterfactual fairness is a form of individual fairness: it evaluates whether an individual received a certain outcome from the system based on their merit, not because of their demographic group.

For further reading, there are many academic publications on the topic of counterfactual fairness, most notably this seminal paper by Matt Kusner et al.

### Implementing counterfactual fairness

The following process is followed to apply counterfactual fairness to evaluate the AI system:

| | |
|---|---|
| **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** | Statistical techniques are used to measure the extent to which the counterfactual samples experience a different score from the original samples. |
| **Score** | The results are mapped to a score for that demographic group from 0 to 100 (more on our scoring system below). The group scores are averaged to create an overall score for the verification. |
| **Review** | Our team of AI auditors perform spot checks on the AI system and reviews the interpretation of the results. |

# Methodology

## Scoring system

For each verification performed, we summarize the results with the following outputs:

| | Description | Values |
|---|---|---|
| **Test samples** | The total number of inputs that were submitted to the system. | 1 to 10,000 |
| **Score** | The score quantifies the AI system's performance on a scale from 0 to 100%, reflecting the degree to which the evaluated system can be trusted on this dimension. | • From 0% to 100%<br>• • A score of 100% means there was complete consistency between the original and counterfactual samples.<br>• A score of 0% means there was no consistency between the original and counterfactual samples. |
| **Grade** | The high-level categorisation of the level of trustworthiness for this result. | The grade is determined based on the following score ranges:<br>• Excellent: 95-100%<br>• Good: 90-95%<br>• Moderate: 75-90%<br>• Poor: 50-75%<br>• Very poor: 0-50% |
| **Interpretation** | A simple statement that states the key take-away from the results in layman's terms. | Example: "The system is largely resilient to linguistic variations, ensuring reliable user experiences." |

# Methodology

## Data sets

Our evaluation framework uses proprietary datasets to conduct tests on AI applications. The datasets are designed to be representative, covering diverse scenarios and populations to ensure the thorough assessment of AI models.

### Representativeness

To deliver effective evaluations, we focus on ensuring that our datasets are representative of real-world conditions. This involves including a wide range of data types, scenarios, and demographic characteristics. By doing so, we aim to accurately reflect the diversity and complexity of the environments in which AI applications operate.

### Ethical sourcing and data protection

All datasets are ethically sourced, adhering to the highest standards of data collection practices. We are committed to maintaining confidentiality and protecting personal data. To this end, any collected data is anonymized, with personally identifiable information removed to uphold privacy.

Despite the anonymization, 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. This consent process is documented and adheres to legal and ethical guidelines, reflecting our commitment to ethical AI.

# Disclaimer

This AI Assurance Report has been prepared by Warden AI Ltd. to provide an independent evaluation of the AI system developed by the AI provider in question, based on our proprietary methodologies and datasets. The evaluations and conclusions presented in this report reflect our best professional 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 related to fairness, robustness, 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 evaluated 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 AI Ltd. 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 AI Ltd. harmless against any claims, damages, or liabilities that may arise from the use of the evaluated AI system.

Warden AI

# Popp AI Trust Assurance Report

March 6, 2024