GLOBAL AI ASSURANCE PILOT

Scan to read the full case study.

AI-enabled Candidate Screening and Evaluation (HR)

Application Tested

MIND-Interview

Mind Interview, a Taiwan-based HR AI tech startup, has a Candidate Assessment and Screening tool that helps employers screen and evaluate job candidates.

Tester

FAIRLY Assurance for AI

Fairly AI provides software and services for governance, risk and compliance management of predictive, generative and agentic AI systems.

How LLMs were used in application?

- Summarisation
- Classification or recommendation
- Data extraction from unstructured source
- Video or audio to text (not part of pilot)

What Risks Were Considered Relevant And Tested?

Bias

Potential for disproportionate scoring outcomes based on sex, race, or sex+race combinations, scrutinised under NYC Local Law 144

Privacy

Risk of inadvertently soliciting or inferring age-related information, prohibited under US/UK/Canadian employment laws

Quality

Risk of misalignment between candidate answer quality and resulting score; risk of inconsistency vs industry benchmarks

How Were The Risks Tested?

Approach

Bias

Synthetic candidate profiles varying demographics (sex, race, sex+race) with fixed candidate answers

Privacy

Manual review of top 3 questions and auto PII scan of summaries and reasonings of scoring

Quality

- Stress testing (validate that good answers get good scores, and bad answers get bad scores): Defined sample good/bad answers; score comparison to mean
- Benchmarking: Evaluating scoring consistency vs. external model

Evaluators

- Human judgement
- Rule-based algorithms
- Non-LLM model-based evaluations

Challenges and Insights

- Generation of realistic synthetic candidate profiles without introducing confounding variables
- Defining objective good and bad answers across different types of questions
- Variability and drift in the LLM outputs makes it non-trivial to determine how many times we need to run the same test to achieve the confidence level we need

Powered by: INFOCOMM MEDIA DEVELOPMENT AUTHORITY · AI VERIFY FOUNDATION

AI-enabled Candidate Screening and Evaluation

MIND-Interview

Mind Interview is a Taiwan-based HR AI tech startup offering tools to assist employers in screening and evaluating job candidates. The assured firm operates in the HR technology sector, offering tools to assist employers in screening and evaluating job candidates.

Use Case

The specific AI application under assessment is designed to score candidate responses to the top three interview questions as part of an initial screening process. The scores generated are intended for internal use only by hiring managers, not disclosed directly to candidates, and serve to support more consistent and scalable early-stage candidate evaluations.

High-level Architecture

The application utilises a sequential pipeline architecture:

01 The application uses a standard LLM-based text evaluation architecture where an LLM acts as the primary evaluator, generating initial scoring assessments of candidate responses.

02 A human feedback loop is incorporated during the model development and monitoring stages to refine and validate scoring behaviour, ensuring that it aligns with intended evaluation criteria and business requirements.

No external foundation model (e.g., GPT-4o-mini) is used in the production environment for scoring candidates. The production LLM was internally calibrated using proprietary methods and datasets. No fine-tuning was performed - calibration and prompt engineering were used to optimise model behaviour instead.

Data Sources

- Input: Candidate responses to the employer's top three interview questions.
- Training/ Calibration: Internal HR and hiring rubric materials; External synthetic and third-party candidate data.

Tool Usage

- Internal model orchestration platform
- Human reviewers during calibration and monitoring phases
- GPT-4o-mini API used only for benchmarking during external assurance testing

02

Testing Partner and Testing Approach

FAIRLY
Assurance for AI

Fairly AI is an AI assurance solution provider offering software and services for governance, risk and compliance management of predictive, generative and agentic AI systems.

Fairly AI applies ISO 42001 Management System for AI framework as the core approach. It conducts preliminary AI Risk Assessment, identifying inherent risk under different risk categories, applying risk controls, testing against these controls to measure residual risk. Tools include the Fairly AI Management System integrated with its Asenion Test Agent framework, to cover the different types of tests needed.

03

Risk Assessment and Testing Scope

Risks tested	Description
Bias Risk	Potential for disproportionate scoring outcomes based on sex, race, or sex+race combinations, scrutinised under NYC Local Law 144.
Privacy Risk	Risk of inadvertently soliciting or inferring age-related information, prohibited under U.S., U.K., and Canadian employment laws.
Security Risk	Out of scope. The algorithm is used internally and is not externally exposed.
Quality Risk	Risk of misalignment between candidate answer quality and resulting score; risk of inconsistency compared to industry benchmarks.

04

Test Design

Technical tests were designed to specifically address the identified risks, combining automated and manual methods.

Test Category	Purpose	Methodology	Key Metrics / Criteria
Bias Testing	Detect scoring disparities across demographic groups	Synthetic profile generation; constant answer text	Impact Ratios by sex, race, and sex+race (threshold: 80% rule)
Privacy Testing	Ensure no prohibited age-related questions are present	Manual review of top 3 questions and auto PII scan of summaries and reasonings of scoring	Compliance checklist (U.S., U.K., Canada laws)
Stress Testing	Validate that good answers receive high scores; bad answers receive low scores	Defined sample good/bad answers; score comparison to mean	Good answers > mean score, Bad answers < mean score
Benchmarking	Evaluate scoring consistency vs. external model	GPT-4o-mini scoring and deviation analysis	Mean Absolute Deviation within acceptable range

05

Test Implementation

Methodology:

- **Bias Testing:** Created synthetic candidate profiles varying demographics (sex, race, sex+race) with fixed candidate answers. Compared output scores for impact ratio analysis.
- **Privacy Testing:** Conducted manual legal review of the three interview questions and auto scan for PII (age) of the summaries and reasonings of the scoring results for compliance with employment laws regarding age discrimination.
- **Stress Testing:** Designed structured good and bad answers for each question and evaluated resulting scores against mean scores.
- **Benchmarking:** Input candidate answers into GPT-4o-mini, scored responses, and compared deviations from the Vendor's system.

Software and Tools Used:

- **Fairly AI's proprietary Bias Testing Toolkit**
- **OpenAI GPT-4o-mini API (benchmarking only)**
- **Standard statistical analysis libraries**

Metrics and Thresholds:

- **Bias:** Impact Ratio threshold at or above 80% (Four-Fifths Rule).
- **Stress Testing:** Good responses expected to score above mean; bad responses expected to score below mean.
- **Benchmarking:** Acceptable mean absolute deviation set by risk tolerance thresholds.

Interpretation Criteria:

- **Alignment with New York City Local Law 144 guidance.**
- **Quality testing assessed against statistical means and benchmark deviation tolerances.**

06

Insights/Lessons Learned

Challenge	Mitigation Approach
Generation of realistic synthetic candidate profiles without introducing confounding variables	Randomised controlled profile generation based on NYC LL144 audit requirements
Defining objective good and bad answers across different types of questions	Used human review and mean score of all responses as threshold
Variability and drift in the LLM outputs makes it non-trivial to determine how many times we need to run the same test to achieve the confidence level we need	Conducted multiple scoring runs and used statistical concepts to identity the minimum runs required for statistical significance

While these approaches cannot eliminate risk completely, they can be used to establish baseline testing to minimise risk over time. In other words, they provide a standardised structure to gather metrics so one can compare how the system is doing relative to its past metrics as well as out-of-the-box models.

AI VERIFY FOUNDATION