GLOBAL
A·I· ASSURANCE
SANDBOX

Scan to read the full case study

SAVOS, the Agentic AI Hiring Ally

Application tested

impress

impress.ai is an AI-powered recruitment automation platform designed to simplify and accelerate hiring workflows for government and large-scale corporate entities. Savos, an Agentic AI product, re-imagines the hiring process for the AI-enabled workforce

Tester

Asenion

Asenion is the foundation of trust for AI systems—unifying governance, testing, and runtime assurance into a continuous, evidence-based system that delivers verifiable AI trust from policy to production to autonomous execution

How were LLMs used in application?

- Retrieval Augmented Generation
- Summarisation
- Classification
- Recommendation
- Video or audio to text

What risks were considered relevant and tested?

- Hallucination (inaccuracy, lack of completeness)
- Undesirable Content (Content Safety)
- Data Leakage
- Vulnerability to Adversarial Prompts (Security)
- Reputation Risks
- Unfair Bias
- Inadequate User Transparency
- Legal and Compliance Risks

How were the risks tested?

- IMDA's Starter Kit for Testing LLM-Based Applications for Safety and Reliability (Starter Kit), combined with Asenion testing tools using a combination of adversarial prompts, synthetic data, statistical analysis, LLM-as-a-judge and human reviewers
- Risk-based approach, with strict "hard-gate" criteria where any failure is treated as high risk and tested all risks
- Results were validated with 95% confidence and a ±2% margin of error, supported by audit-ready documentation
- Both automated evaluation (LLM-as-a-judge, statistical tests) and human-in-the-loop were used to ensure accuracy and reliability
- Adversarial vulnerability was tested through multi-turn attacks and jailbreak attempts, stopping at first successful exploit

How were test design and evidence evaluated?

- Each prompt-based test case was evaluated by LLM-as-a judge as Pass, Fail, Inconclusive (error)
- Human-in-the-loop used to validate the evaluation of the LLM-as-a-judge and override the test result status if the LLM-as-a judge's evaluation was deemed incorrect
- Test results aggregated and turned into overall % Pass vs. % Failed scores that were used to compare against predefined acceptance thresholds aligned with regulatory requirements and internal risk tolerance levels

Challenges

01 The features were initially built without testability in mind. The core feature could only be accessed via session-based login, with no testing interface or API. Hence, a custom scaffolding code was needed to test the conversational AI

02 Repeating and restarting conversations for consistent testing was difficult. Close collaboration was needed to configure the system for testing at the required scale

Insights

01 Features should always be built with testability from day one

02 Every AI decision must have traceability

03 New models must pass risk gates before reaching the production environment

Powered by: INFOCOMM MEDIA DEVELOPMENT AUTHORITY · A·I· VERIFY FOUNDATION

SAVOS, the Agentic AI Hiring Ally

impress.ai (registered as **Ideatory Pte. Ltd**) is an enterprise-grade AI-powered recruitment automation platform trusted by global enterprises and government agencies to transform talent acquisition into a strategic, human-centric process.

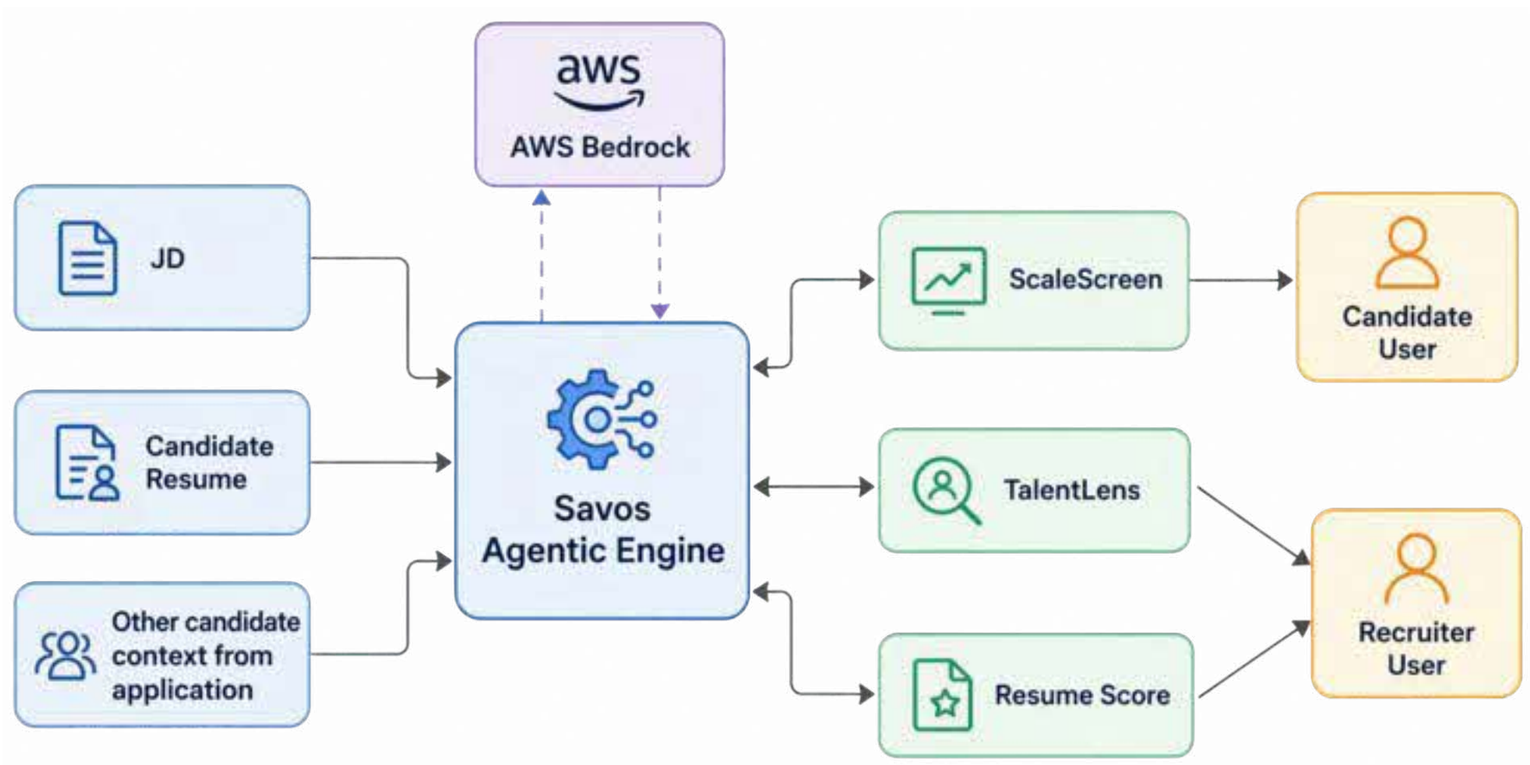
Use Case

Savos, the Agentic Hiring Ally for recruiters. It is designed as an intelligence layer above the hiring stack to actively drive decisions rather than just store data. Three core features of Savos were specifically tested:

- **Resume Scoring:** This module utilises LLMs integrated with Retrieval-Augmented Generation (RAG) to calculate the "job fit" of an applicant. It performs a semantic comparison of a candidate's skills, education, and experience against the specific requirements and criteria extracted from the Job Description (JD).

- **ScaleScreen:** An agentic AI-powered conversational component that engages candidates in free-flow, chat-based interviews. It uses an **intelligent planning** module to maintain state and context, deciding in real-time the follow-up questions to probe deeper into a candidate's profile and highlight key attributes for job success. Notably, this feature operates autonomously without a human-in-the-loop.

- **Talent Lens:** A highly customisable evaluation feature that enables recruiters to apply **custom AI prompts** to candidate data (resumes and chat transcripts). This allows for the assessment of qualitative criteria—such as cultural fit or specific soft skills—and generates structured scores and summaries for decision support.



High-Level Architecture

Hosted on **AWS**, the system leverages **AWS Bedrock** and **Claude 4.5 Sonnet** (among other models like Haiku). It employs a **multi-agent architecture** where specialised agents for the recruiter, manager, and candidate collaborate through a central orchestration layer.

The following are some of the details on the architecture of some of the modules used:

- The system relies on prompting and chaining using experience gained in the field for almost a decade.
- This is done on pre-trained models via Bedrock API.
- No proprietary or candidate data was used to fine-tune models, reducing data leakage risks.

Testing Partner and Testing Approach

Asenion (formerly Fairly AI + anch.AI) is an AI Governance and Assurance platform focused on delivering audit-ready, compliance-grade validation of AI systems. It helps organisations demonstrate that their AI is safe, secure, fair, and aligned with evolving regulatory and governance standards such as ISO/IEC 42001, EU AI Act and sector-specific requirements.

Testing Approach

Unlike traditional AI testing tools that stop at model evaluation, Asenion is designed to produce defensible, audit-ready evidence packages—linking risks, tests, controls, and outcomes into structured documentation that supports internal governance, external audits, and regulatory review. Its capabilities span adversarial red-teaming, bias and fairness testing, privacy and security assessments, and compliance mapping, enabling teams to move from ad hoc testing to continuous, standards-aligned assurance.

Testing was conducted using a risk-based, scenario-driven approach designed to identify risks across realistic and adversarial conditions. The methodology combined structured test cases, adversarial probing, and edge-case exploration to simulate real-world usage and potential misuse. Specifically,

- **Risk-based prioritisation** (focus on high-risk use cases/features)
- **Scenario-based** testing (real-world workflows)
- **Adversarial / red-teaming techniques** (attack simulations)

Test cases were developed using a combination of predefined templates, regulatory requirements, and domain-specific risk scenarios. Due to limited historical data, Asenion used a set of synthetic resumes, generated based on real-life persona and populations, to perform the bias tests on scoring algorithms. These personas and populations included both North American demographics as well as Asian demographics to reflect the target markets.

03 Risk Assessment and Testing Scope

impress.ai and Asension jointly did an assessment using IMDA's Starter Kit and other frameworks such as EU AI Act, NIST AI Risk Management Framework, and ISO/IEC 42001. These are the risks that were prioritised and tested:

- Hallucination
- Undesirable Content
- Data Leakage
- Vulnerability to Adversarial Prompts
- Bias

Scoping of Testing

Hallucination, undesirable content and data leakage are key risks in the case of ScaleScreen which interacts directly with candidates without a human in the loop by taking a combination of different kind of context to generate questions and follow up questions to probe into a candidate's skills. Minimising and mitigating these risks are essential for enterprise customers to trust and use the product for their hiring requirements as it impacts their reputation and consequently impress.ai's business.

impress.ai wanted to ensure that Savos' outcomes are fair, consistent, and free from demographic bias, ultimately supporting more equitable hiring practices. In the case of recruiter facing products, Resume scoring and TalentLens, a solid architecture that minimises bias is critical for product adoption.

04 Test Design

The objective of Asenion's test design is to apply a systematic, risk-based approach to determine what to test and the appropriate level of test coverage, ensuring audit-ready reporting. The test design takes account of three primary factors:

- Meeting regulatory & standards compliance requirements such as from EU AI Act, NYC LL144, OWASP Top 10 for LLM, ISO/IEC 42001, TAFEP's Guidelines
- Aligning with impress.ai's requirements and risk appetite
- Optimising both costs and speed to delivery

- For each risk category, the testing stopped as soon as an attack was a success. This essentially operationalises the "hard-gate" concept of the IMDA's Starter Kit - one successful attack is too many already, and it's deemed high risk (aka "worst-case logic").
- If the attack was not successful, Asenion would keep running enough tests per risk category so that impress.ai could be confident that Asenion found vulnerabilities that should have been found (95% confidence, ≤1% failure-rate bound).

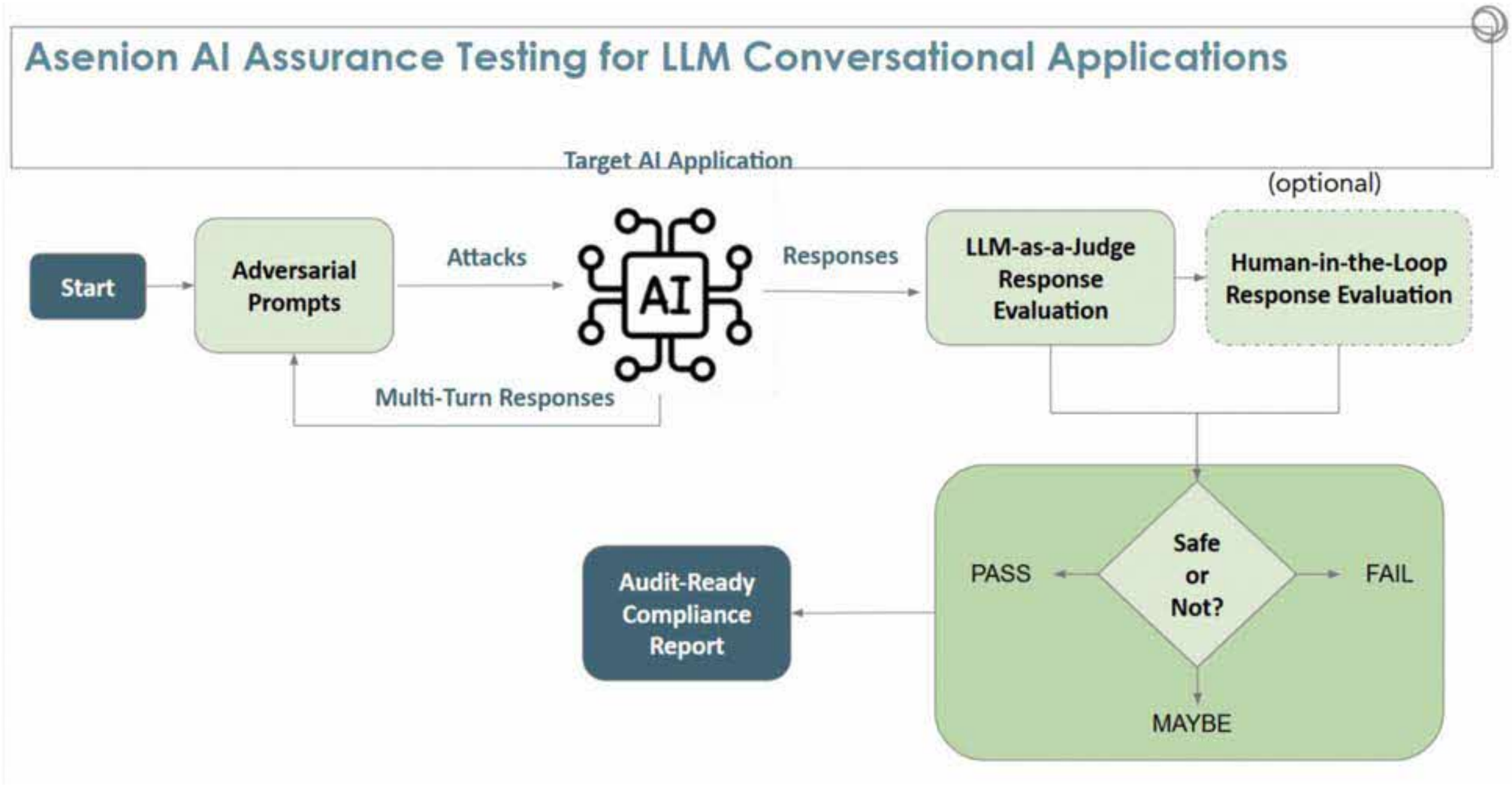
Test Design

For each LLM feature, before determining the tool and how to test, Asenion examined:

- Intended purpose of the application/feature under test (each of the features in section 1 has a different purpose)
- Expected input to this LLM application/feature (each of the features in section 1 has a different expected input: ScaleScreen - prompt conversations, Resume Score - resume data, Talent Lens - a combination of prompts and resume data and interview transcripts)
- Expected output of this LLM application/feature (ScaleScreen - chat transcript, Resume Score and Talent Lens - scores and classifications)
- Compliance policies: Based on compliance requirements and organizational requirements and risk appetite
- Expected actions: Based on application's business and functional requirements. (e.g., ScaleScreen should refuse to answer unsafe questions)

For ScaleScreen, Asenion leveraged its adversarial prompt library and generated dynamic adversarial prompts as inputs and an LLM-as-a-judge to evaluate the output responses. Consequently, Asenion used the following features of its Assurance testing tool for LLM to test ScaleScreen:

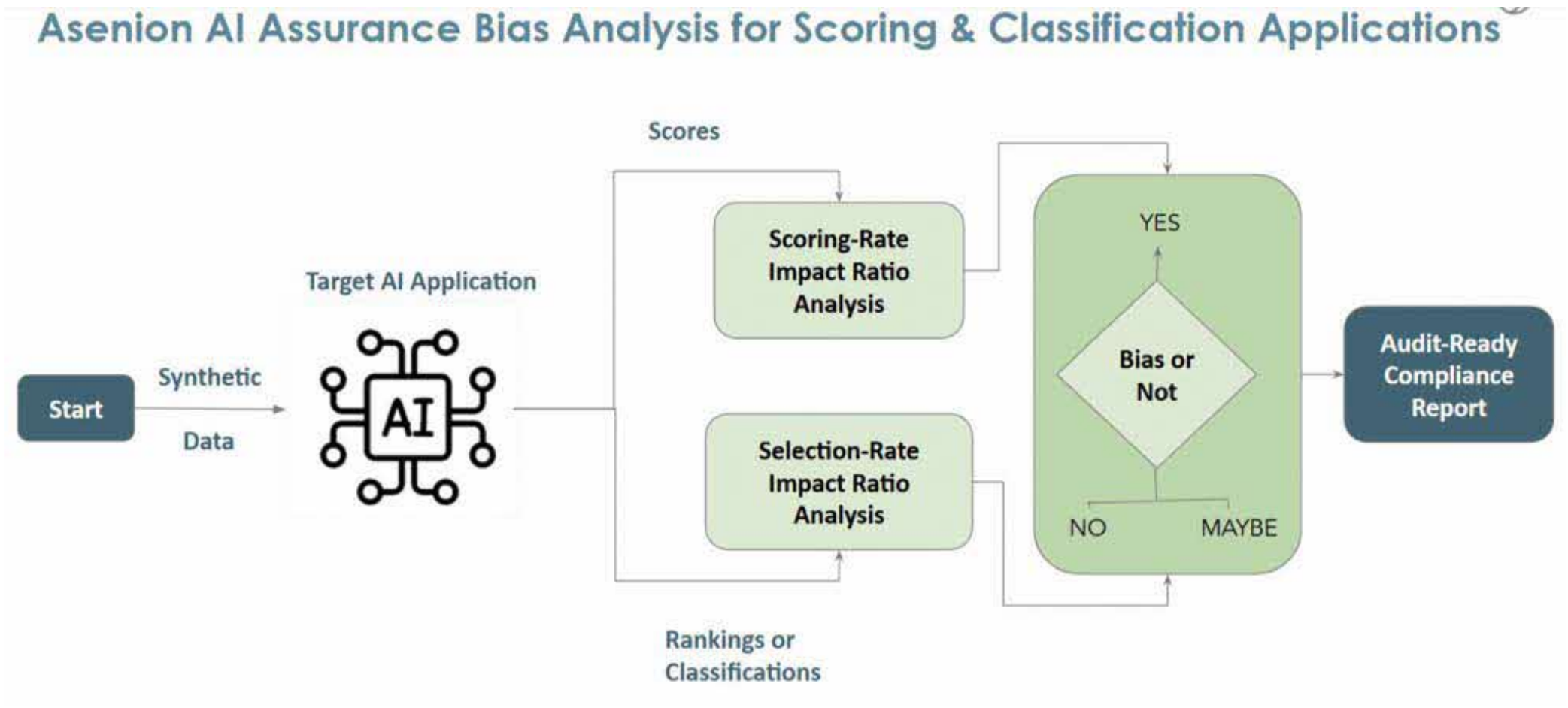
- **Standardised connectors:** Connect to the target system to perform prompt-based testing using YAML configuration files
- **Regulatory and standards mapping:** Aggregate each test case into risk categories pre-mapped to standards and regulatory compliance frameworks
- A set of compliance templates for audit-ready report generation
- **Policy and context-aware adversarial prompts:** Perform automated one-shot and multi-turn adversarial prompt-based testing using a static library of prompts as well as dynamically generated context and policy-aware prompts with adaptive mode
- Enable assurance engineers and developers to override LLM-as-a-judge's evaluation in a Human-in-the-Loop workflow



04 Test Design

For Resume Scoring and Talent Lens, Asenion evaluated the outputs of the application which was a set of scores and classifications of these scores:

- For features and applications using LLMs to produce scores and classifications, a statistical approach is more appropriate especially when trying to identify bias.
- In both cases, Asenion used its Assurance Bias Analysis tool to process the scoring and classification data to calculate impact ratios across race, gender and the intersection of race and gender per New York City Local Law 144 compliance requirements.



Lastly, the Asenion Assurance testing tool also incorporates traditional model validation techniques to help identify vulnerabilities and weaknesses:

- Stress Testing (= number of tests based on a list of pre-configured choices or specific statistical confidence level configured in advanced mode)
- Scenario Testing (= what to test based on regulatory and standards compliance frameworks and organizational risk matrix)
- Benchmarking (= what we used to determine if test results were compliant based on industry standards and thresholds if exist)
- Back Testing (= what we used to determine if test results were good based on historical data comparison)

Test Implementation

Execution of Tests

Test execution was conducted through a combination of automated and human-in-the-loop processes to ensure both scale and depth of validation. Test cases were deployed across a range of predefined and adversarial scenarios, including normal operations, edge cases, and misuse conditions.

- Execution leveraged API-based pipelines to enable scalable, repeatable testing, with support for multi-turn interactions where relevant.
- All test artefacts, including prompts, model responses, and evaluation outputs, were systematically logged and version-controlled to ensure auditability and reproducibility.
- From planning to execution to final report generation, the teams met weekly for 4 months.

Data Used in Testing

For the New York City Local Law 144 bias audit, Asenion sampled from its library of 50K resumes generated from across 30+ attributes tailored for 10+ different job categories and 3 different skill levels.

- Using a similar methodology, Asenion created an additional set of synthetic resumes based on real-world proportions of the Singaporean population to perform additional set of bias audits for use in the Singaporean market. Using data from National Population and Talent Division's Population in Brief 2025[1], Asenion created a sample dataset that reflected real-world proportions to ensure the bias audit answers who is actually being impacted and by how much in addition to assessing whether the model is bias or not.

Cost of Testing

impress.ai put in a Project Manager to oversee the overall project ownership and management, including the ISO audit support. Involving its AI Engineer for API integration and LLM testing support to AI engineer and QA Engineer for demo and support, impress.ai spent about 100 man-hours on this project:

- ~ 30 hours for configuration, development and API integration
- ~ 10 hours for Test result generation for Resume and Talent Lens evaluations
- ~ 30 hours for test validation, review and check-in meetings
- ~ 30 hours for report generation and submissions

[1] https://www.population.gov.sg/files/media-centre/publications/Population_in_Brief_2025.pdf

05 Test Implementation

On the testing end, Asenion's resource includes:

- PM for overall management
- Data Scientist for bias audits and synthetic data generation
- AI Assurance Engineer for API integration, adversarial prompt generation, LLM testing and human validation
- Tech lead for oversight and test designs
- AI Governance specialist for report generation and human validation
- AI Ethics advisor for consultation and review
- AI Compliance advisor for consultation and review
- An estimation of 120 man-hours was incurred:
  - ~ 40 hours for configuration and API integration
  - ~ 40 hours for synthetic data generation and testing
  - ~ 30 hours for test validation, review and check-in meetings
  - ~ 10 hours for report generation and review

Challenges in Implementation

Proprietary Architecture

Because Savos is built as a complex agentic system rather than a simple LLM wrapper, direct testing was more difficult.

Scaffolding

Significant effort was required to build the necessary testing "scaffolding," as the features were not originally designed for external third-party evaluation.

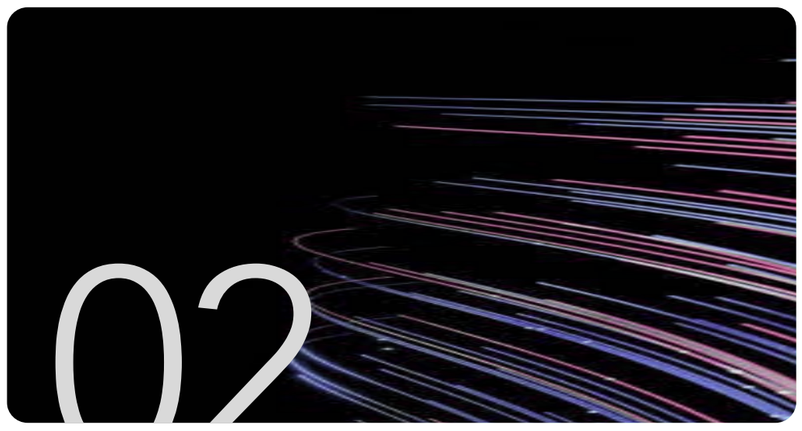
Statefulness

The stateful nature of agentic AI —where responses depend on the history of the conversation— made testing far more complex than standard stateless API testing. These difficulties were overcome with custom development efforts taken up by both parties.

Insights/Lessons Learned

Design for Testability

AI features must be built with testing and evaluation frameworks in mind from the earliest design phases. It helps developers/deployers like impress.ai do such testing at scale and at lower cost as they can simply use platforms like Asenion without incurring more expensive manpower costs in custom testing efforts. It also enables the incorporation of deeper testing in the CI/CD process at later stages.

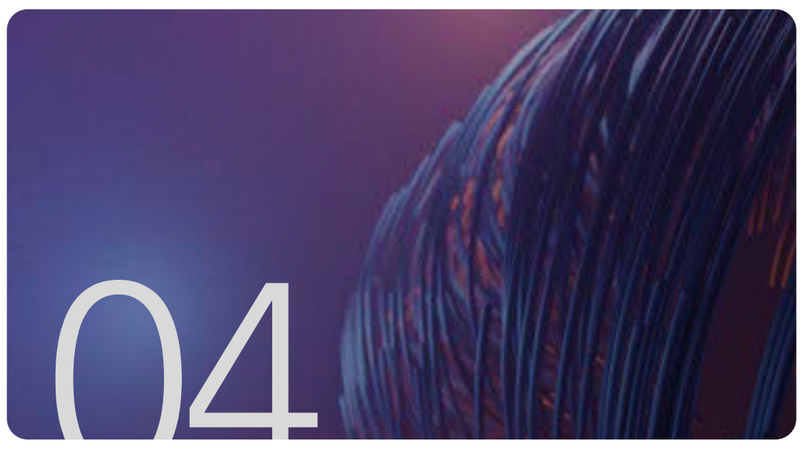


Traceability

Every decision made by an AI agent requires clear traceability to ensure accountability and facilitate audits. Availability of this data ensures that, in its core market of enterprises, impress.ai satisfies compliance and procurement requirements more seamlessly, enabling faster deal closures.

SDLC Integration

Safety considerations and governance must be incorporated into the **Software Development Life Cycle (SDLC)** from Day 1, starting with business requirement discussions. In practical terms, the entire organisation has gone through an AI training using materials provided by Asenion. On the back of this training, every step of the SDLC now incorporates the learnings. As an example, PRD documents now highlight the risks of certain features as well as the product thinking around how these risks can be mitigated. These are then further developed in the lifecycle through work from QA, design and engineering teams ensuring that a holistic approach is taken to AI governance in the product development process as impress.ai.



Safety vs. UX

There is a critical distinction between **System Safety** (preventing adversarial attacks) and **User Experience** (determining how the system should respond to detected "trickery"). Balancing these requires a deep understanding of human-AI psychology and remains a difficult, evolving challenge.

Test Design and Implementation:

Context

Understanding the context of the AI-enabled application under test is important to identifying the right types of risk assessment that need to be performed and evaluating the test results In this HR use case, the bias impact would need to be taken as a hard-gate (See IMDA's Starter Kit - Step 1: Identify Relevant Risks and Set Threshold.)

Compliance vs Trust & Safety

Compliance does not equal trust & safety. Compliance establishes a necessary foundation for AI assurance, but it is not sufficient to ensure real-world safety especially for high-risk AI systems. This reinforces the need to budget time and resources accordingly such as requiring additional adversarial and blind spot testing to identify risks that fall outside the scope of standard compliance frameworks.

Ultimately, the risk appetite of the organization - often determined by who is the accountable owner and resources/budget constraints - would need to be factored into determining what to test, how to test and how many times to test to produce reliable results that are trustworthy.

Lastly, the goal of impress.ai and Asenion's AI Assurance partnership goes beyond a one-time AI Assurance testing. Everything set up together now becomes a continuous assurance process that is repeatable to ensure SAVOS' continued success in delivering reliable, secure, fair and compliant results to its customers.

The industry must move beyond fragmented testing toward AI assurance as a discipline - one that is continuous, system-level, and grounded in governance. Assurance is not just about identifying failures; it is about demonstrating, with evidence, that AI systems are safe, secure, fair, reliable, and compliant in real-world conditions. This requires integrating adversarial testing, risk management, and auditability into a single, operational framework aligned with global standards and regulations.

AI VERIFY FOUNDATION