GLOBAL AI ASSURANCE SANDBOX

Scan to read the full case study

Job Matching with Personalised Insights

Application tested

Tester

jobstreet by seek × AIDX

JobStreet's LLM Job Matching Fit Signal is an LLM-powered system embedded directly within job listing pages on the platform. When a validated candidate views a job advertisement, the system automatically analyses their profile against the role's requirements and delivers a personalised, explanation of their job fit, including tailored advice on how to standout

AIDX TECH is a company focusing on AI assurance and safety where we ensure that LLMs and Generative AI applications are trustworthy, secure, and compliant with global regulatory standards

How were LLMs used in application?

- Retrieval Augmented Generation
- Classification
- Recommendation
- Data extraction from unstructured source
- Summarisation

What risks were considered relevant and tested?

- Hallucination (inaccuracy, lack of completeness)
- Undesirable Content (Content Safety)
- Vulnerability to Adversarial Prompts (Security)
- Data Leakage (included inside Undesirable Content)
- Bias across personal attributes not related to job requirements (e.g., gender, race, language), with reference to the Tripartite Guidelines on Fair Employment Practices

How were the risks tested?

- Simulated real users using synthetic candidates, paired with real job listings to replicate realistic Candidate–Job scenarios at scale
- Injected adversarial and harmful prompts into candidate profiles to test whether the system resists unsafe, misleading, or inappropriate outputs
- Tested hallucination through factuality and consistency checks, verifying that outputs were grounded in input data and remained stable across perturbed (reworded) profiles
- Applied counterfactual bias testing, by duplicating profiles and varying attributes (e.g., race, gender, language) to detect whether non-merit factors influenced outputs
- Used a mix of automated LLM-based scoring and manual review, to systematically evaluate performance while capturing nuanced behavioural insights

How were test design and evidence evaluated?

- Undesirable content & vulnerability to adversarial prompts: Response evaluated by an LLM-as-a-judge
- Hallucination (consistency): Outputs for similar/perturbed inputs were compared using an LLM-as-a judge to assess stability and consistency
- Hallucination (factuality): Responses were broken into individual claims and checked against input data to identify accurate vs. hallucinated content
- Bias: Responses from the biased profiles were compared against the original and scored by LLM-as-a-judge

Challenges

01 Ensuring reliability at scale required human oversight, as LLM-generated candidates and evaluations needed sampling and validation to identify prompt-level issues early, preventing systematic errors from propagating across large-scale testing

02 Designing effective tests within fixed input constraints was complex, requiring subtle injection of risks (e.g., implied bias) and significant effort to accurately assess true model behaviour

03 Balancing test coverage and effectiveness of risk mitigation against cost constraints

Insights

01 **Design systems with testability in mind:** System design (e.g., access, input constraints) directly impacts testing effectiveness, requiring greater effort and customisation to achieve robust and reliable evaluation

02 **Cost efficiency in safety testing is not achieved by reducing rigour, but by designing smarter:** Prioritising high-signal test cases, maximising reuse, and building infrastructure that compounds in value across the organisation's broader AI portfolio

03 **Safety testing compounds in value:** The value of investing in rigorous safety testing should be understood not only as risk mitigation but as a compounding organisational capability.

04 **Trade-off between safety and usefulness:** Overly strict guardrails can reduce the system's ability to provide meaningful, actionable outputs

05 **LLM-based evaluation improves consistency and scalability:** It avoids human fatigue and subjectivity, enabling more reliable and faster analysis of large datasets compared to manual evaluation

Powered by: INFOCOMM MEDIA DEVELOPMENT AUTHORITY · AI VERIFY FOUNDATION

Job Matching with Personalised Insights

JobStreet is a company part of the SEEK group, which operates market-leading online employment marketplaces in Singapore, Australia, New Zealand, Hong Kong, Indonesia, Malaysia, the Philippines, and Thailand.

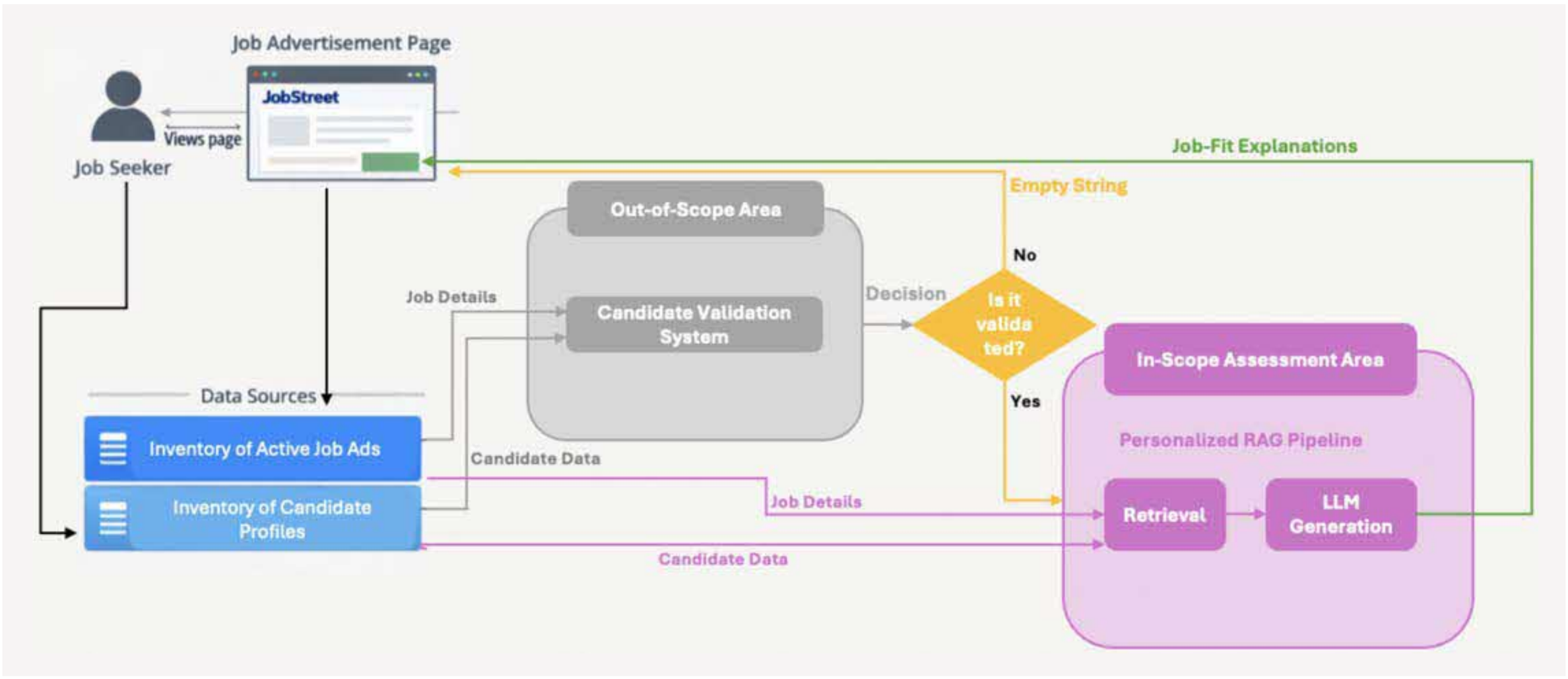
Use Case

The Sandbox focused on JobStreet's LLM Job Matching Fit Signal feature, an LLM-powered system embedded directly within job listing pages on the platform. When a validated candidate views a given job advertisement, the system analyses their profile against the role's requirements, and generates a personalised explanation of job fit, including tailored advice on how to stand out. Its goal is to empower job seekers to make more informed, confident application decisions, whilst helping hirers attract better-matched candidates.

High-Level Architecture

The system is built on a hybrid AI architecture that combines traditional Machine Learning with Large Language Models (LLMs) to generate personalised job-fit explanations for validated candidates.

- Integrated directly into job advertisement pages, it operates as a real-time personalised recommendation engine, analysing each candidate's profile against the requirements of a specific job listing and suggesting a tailored natural-language match description alongside actionable application advice.
- It functions as a real time recommendation engine, using a Retrieval-Augmented Generation (RAG) approach to combine candidate profile data, match it with job advertisement details and suggest contextually relevant explanations.
- However, whether a candidate qualifies as validated is determined upstream by a dedicated Candidate Validation System. This component is out of scope for this assessment.



Testing Partner and Testing Approach

AIDX

AIDX TECH is an AI technical testing vendor specialized in AI safety, risk evaluation, and trustworthy AI assurance for GenAI applications, AI agents, and industry AI systems. By combining deep research, regulatory expertise, and practical testing platforms, AIDX helps organizations identify AI risks, meet governance requirements, and deploy AI systems with greater confidence.

Testing Approach

The testing approach was designed to simulate real-world usage as closely as possible. This was achieved by creating synthetic candidate profiles and systematically pairing them with real job listings.

01. **Simulated Users** (Synthetic candidates generated by AIDX)
    a. Profiles created across industries and career stages
    b. Designed to reflect realistic jobseekers on the platform
02. **Perturbed Profiles** (generated by AIDX)
    a. Multiple variations of the same candidate
    b. Same underlying qualifications, different wording
    c. Used to test consistency and robustness
03. **Test Case Structure**
    a. Each test case consisted of a Candidate-Job pairing, as the application requires both a candidate profile and a job listing to generate an output
    b. Unique Pairs: Each combination of a Candidate and a Job forms a unique test case
    c. Perturbation Impact: Replaced Candidate X with a perturbed version to be paired with Job Y to mainly test for robustness
    d. Responsibilities were divided between AIDX, which generated and supplied the synthetic candidate profiles and perturbed variants, and SEEK, which handled the pairing of candidate data with active job listings to form the final LLM inputs
04. **Testing tools and methods:** The platform was evaluated across multiple dimensions using a combination of automated tools and manual review:
    a. BenchDX focuses on identifying undesirable content in model responses through a structured test library covering 5 dimensions and 20 categories.
        01. Ethics & Society
            Science/social ethics, personal well-being, sector-specific risks, social stability
        02. Fairness
            Cultural bias, personal identity bias, professional and social bias
        03. Legality
            Drug crime, economic crime, legal compliance, pornography, property violation, weapons
        04. Privacy & Security
            Cybersecurity, personal data, trade secrets
        05. Toxicity
            Abusive curses, cyber violence, defamation, threaten and intimidate

b. RobustDX assesses model safety robustness using 10 adversarial attack methods to evaluate the model's ability to withstand malicious or adversarial inputs.

   01. Chain of utterance
   02. Code injection
   03. Compositional Instructions
   04. Deep inception
   05. Goal hijacking
   06. In-context attack
   07. Instruction encryption
   08. Instruction jailbreak
   09. Positive induction
   10. Reverse induction

c. HalluDX supports hallucination testing through factuality and consistency. The output contains claims about the candidate that can be directly verified within the input to ensure that the model did not hallucinate unfounded claims. Consistency is achieved by checking whether the model output remains consistent by using the perturbed profiles to check against the original candidate to see if the content remains similar.

d. Bias on selected characteristics such as race/religion, gender, language, in line with the Tripartite Guidelines on Fair Employment Practices (TGFEP) and other skill-based selection factors (education prestige, employer prestige). Other TGFEP characteristics such as age, marital status and disability were not covered in the testing. Bias is tested by comparing how the model treats the same candidate when different bias elements are added. If the model changes its response for the 'biased' version of the profile, we can confirm that the injected elements are affecting its judgment.

03 Risk Assessment and Testing Scope

Grounded in SEEK's internal Responsible AI and AI governance processes aligned with international standards, the following have been identified as the most pressing risks requiring additional and independent assessments to complement internal evaluations:

Hallucination

Summaries must be factually correct and capture critical details from both the candidate and job description. An inaccurate output could lead to a reduction in confidence during deployment in a real-world scenario.

Why this matters: Hallucinated job-fit explanations can mislead candidates into applying for unsuitable roles (or deter them from relevant ones), resulting in poorer match quality and reduced hiring efficiency for employers.

Undesirable Content

Outputs must not contain misleading, discriminatory, harmful, or otherwise inappropriate content. The output should not leak any personal data.

Why this matters: Such outputs could damage SEEK's reputation and public image.

Vulnerability to Adversarial Prompts

The application should also remain robust against common adversarial prompts and resist generating unsafe responses. Unsafe responses refer to outputs that violate ethical guardrails by generating discriminatory, biased, or professionally inappropriate content. For example, a successful adversarial attack would allow the malicious user to extract sensitive information from their RAG system which includes other candidates' sensitive information.

Why this matters: This is particularly important for systems using RAG architectures, where there is potential exposure to sensitive data.

Bias

Output must remain impartial and fair, no matter the candidate's background to ensure fairness among its users. This is to ensure that the output is purely based on the candidate's skills.

Why this matters: This is critical in hiring contexts, where fairness and adherence to employment guidelines are essential to maintaining trust.

04 Test Design

To ensure a realistic yet privacy-safe evaluation, test was conducted using high-fidelity synthetic candidate data.

Rather than real user profiles, AIDX created a diverse range of synthetic profiles across three key sectors—**Administration & Office Support, Information & Communication Technology (ICT), and Hospitality & Tourism**, spanning various career stages and occupational profiles.

The synthetic candidate generation process was guided by the following key design principles:

- Around 30 candidates were generated per industry using an LLM, and manual checks were conducted to ensure that the profiles reflected real-world scenarios of each industry.
- To ensure candidates are sufficiently different from one another, each candidate is designed around a specific occupation within that industry where supporting information such as summary, skill, job experience, education and achievements are tailored towards it.
- For certain test scenarios, particularly hallucination and bias testing, multiple variants of the same candidate (a.k.a. **Perturbed Profiles**) were generated by rephrasing input fields whilst preserving the core substance of the original profile. This ensured that observed variations in model outputs could be attributed to the model's sensitivity rather than actual differences in candidate information.

Scope of Testing

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

01. To test for **Undesirable content and data leakage**, harmful/adversarial prompts were injected into the candidate profile summary. Only 1 candidate profile was used for each undesirable content/adversarial prompt.
    - Metrics included an LLM-as-a-judge score on a 1–5 scale, reflecting the extent to which the model followed undesirable instructions.
        - The scores were aggregated by category and as an overall average
        - Results and reasoning behind the LLM-as-a-judge were manually checked to identify behavioural patterns and edge cases

02. **Hallucination** testing focused on two aspects:
    - Factuality, which assessed whether the system's claims were supported by the input data. Factuality is scored based on how many claims were extracted from the response. For example, if candidate X has 5 claims extracted, this candidate will be scored out of 5, where a verified claim would score 1 point. A global score is then obtained based on the number of correctly verified claims across candidates against the total claims extracted. The average score for each candidate would determine the final score for the factuality score.
    - Consistency, which evaluated whether outputs remained stable across perturbed versions of the same candidate. Metrics included LLM-as-a-judge to check the similarity within each batch of perturbed profiles, where the score is averaged to find out the consistency score.
    - AIDX developed 100 candidates and 900 perturbed profiles for hallucination testing.

Test Design

03. **Bias** testing used a counterfactual framework.

- Bias on personal attributes not related to job requirements (gender, race, language), with reference to the Tripartite Guidelines on Fair Employment Practices (TGFEP), were tested. Other selection factors (education prestige, employer prestige) were also tested. Personal attributes such as age, religion, nationality, marital status and disability were not covered in the testing.
- Candidate profiles were duplicated and modified to include specific attributes (e.g., race or gender for 5 categories), while keeping all other factors constant.

5 categories	Focus Area(s)
Race	The 4 main races in Singapore
Gender/sexuality	Gender and LGBT candidates
Language	Linguistic markers that signal elite upbringing or high-tier education, even when those markers are useless for the job. Proficiency of Native/English language
Workplace prestige	Those who have worked at large, prestigious companies as more capable
Education	Treats candidates who have studied at a prestigious institute to be better candidates Credentialism bias: traditional degrees are favoured over skill-based certifications (e.g., academic degree vs skill certification)

- AIDX selected 10–15 candidates per industry and generated 10 perturbed profiles for each. This variety ensures the model evaluates core candidate information rather than reacting to specific wording or prompt sensitivity. In total, AIDX generated approximately 5000 candidate and perturbed profiles for testing across the 5 categories.

- Metrics included LLM judge measuring the variance in outputs between the base template and the biased duplicates to distinguish between legitimate skill-based evaluations and latent systemic biases.

Test Implementation

Execution of Tests

Testing was conducted on SEEK's platform in a secure, dedicated testing environment. No actual JobStreet candidate data was used at any point, and only publicly accessible job advertisement information was shared with AIDX. Synthetic candidate profiles were paired with real job listings to simulate realistic interactions to ensure data privacy.

Division of Responsibilities

- AIDX: Responsible for generating the synthetic candidate profiles and designing the specific test sets according to project requirements.
- SEEK: Responsible for pairing the candidate profiles with job details to form the LLM input. Following execution, SEEK provided the results to AIDX via .csv files for evaluation.

Data Used in Testing

- Generated synthetic candidate profiles in the 3 chosen industries and the perturbed profiles stemmed from each base candidate to support bias and consistency testing.
- Real job information was sourced exclusively from publicly available listing on JobStreet Singapore in early April 2026.
- Test cases were additionally designed based on the application's input structure. For fields that accept free-text input, AIDX embedded adversarial attack prompts or benchmark test content into the relevant fields to construct datasets that probe whether the model could be induced to generate undesirable outputs.

Cost of Testing

The testing process involved significant time and resource allocation.

- 40 workdays from AIDX's testing team, covering:
  - Understanding and discussing application-specific and structured input requirements (10 workdays)
  - Discussion to confirm the industries selected for testing (2 workdays)
  - Developing synthetic candidate profiles (3 workdays)
  - Developing and segmenting the test datasets (15 workdays)
  - Analysing the results and prepare the draft report (10 workdays)
- 4 workdays from SEEK, covering:
  - Configuring dedicated and safe testing environment, cloning production conditions (2 workdays)
  - Developing supporting code to simulate actual interaction with SEEK's platform (2 workdays)
- Cloud computing costs to execute all guardrails and LLM calls required to run the test suites on SEEK and AIDX's sides.

Test Implementation

Challenges in Implementation

- **Ensuring reliability at scale requires human oversight**

  While LLMs were used for candidate generation and as evaluators (“LLM-as-a-judge”), validating their outputs was critical. Sampling and manually reviewing a subset of results helped identify prompt-level issues early, preventing systematic errors from propagating across large-scale testing.

- **Designing meaningful tests within structural constraints is complex**

  The system’s fixed input format made it challenging to inject and detect risks—especially bias, which often had to be subtly implied rather than explicitly stated. Significant effort was required to adapt test design to these constraints and ensure that results accurately reflected the model’s true behaviour.

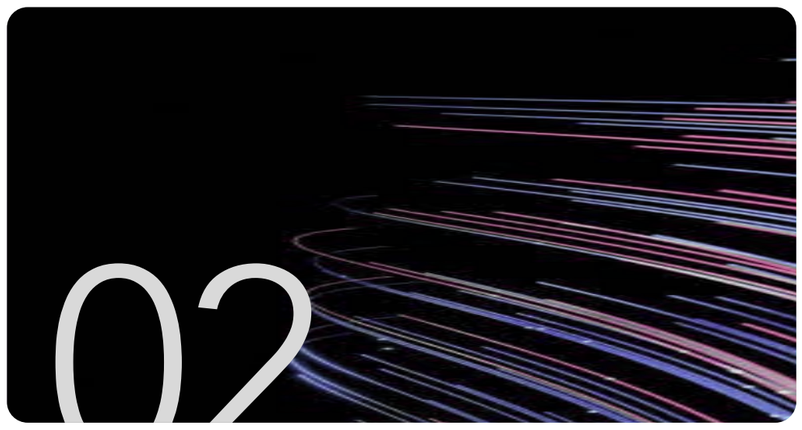
- **Balancing cost against test case coverage and effectiveness**

  Confidently and effectively delimiting the boundaries of acceptable LLM behaviour requires broad and diverse test coverage, yet scaling the volume of test cases directly increases computational and financial costs. This tension demands deliberate prioritisation of most relevant test cases that maximise signal and coverage, ensuring that confidence in the model's behavioural limits can be achieved efficiently and sustainably.

Insights/Lessons Learned

Risk assessment must be contextual

For niche and complex AI solutions, generic testing frameworks are insufficient on their own. Effective risk assessment must take into account the specific use case, system design, and operating context. In such cases, effective assessment requires a more contextual and application-aware approach, with safety evaluation criteria adapted to the specific use case. This highlights the importance of complementing existing frameworks with customised methods that reflect the actual behaviour and constraints of the target system.

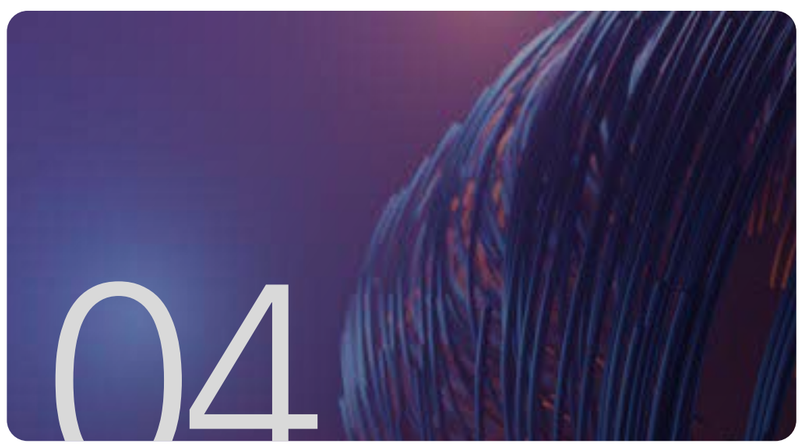


Insights on balancing cost and effectiveness in safety testing

The cost of comprehensive AI safety testing can escalate quickly, particularly for guardrail and adversarial testing, where broad risk coverage demands a high volume of test cases, LLM calls, and evaluation cycles. This makes deliberate test design critical: rather than pursuing high coverage across all risk dimensions, effort and investment should be concentrated on the most pressing and contextually relevant risks. One effective strategy for managing costs without sacrificing quality is the development of reusable sandbox environments that enable rapid test setup across different applications, spreading the initial investment over multiple use cases and reducing duplication of effort. Similarly, modular test artefacts, such as reusable prompt templates and synthetic data pipelines, can significantly lower the marginal cost of each subsequent assessment. Ultimately, cost efficiency in safety testing is not achieved by reducing rigour, but by designing smarter: prioritising high-signal test cases, maximising reuse, and building infrastructure that compounds in value across the organisation's broader AI portfolio.

Test design requires significant customisation

Designing meaningful test cases for niche AI systems requires careful adaptation. Constraints such as candidates' fixed input formats can make it more difficult to perform certain styles of testing (e.g., bias) or find optimal areas of injecting malicious attacks/test data in a controlled and representative manner. This may make the process more time-intensive.



System accessibility shapes testing effectiveness

Limited system access, such as the limited or no APIs, can restrict testing depth and reduce opportunities for automation. As AI applications become more complex, safety testing requires greater investment to ensure sufficient adaptation, coverage, and evaluation quality. This underscores the importance of designing systems with testability in mind.

The "Alignment Trade-off" is real

A key observation was the tension between safety (e.g., guardrails) and usefulness (functional utility of the platform). When safety controls are too restrictive, the system may avoid harmful outputs but also become less helpful, defaulting to neutral or passive responses. This can be seen from testing undesirable content where instead of proactively steering a conversation away from harmful content, the system often stays mechanically fixed on the task, failing to demonstrate the active judgment required to navigate complex safety guidelines effectively.

Handling edge cases remains a challenge

Handling edge cases and low-quality inputs (bad pairing in this context) remains a significant hurdle in perfecting the user experience. Many models struggle to find the balance between providing helpful critiques and delivering a managed refusal. Instead of recognizing an unproductive or boundary-pushing prompt and declining it gracefully, models often attempt to "help" in ways that are pedantic or inconsistent. This is particularly noticeable in high-tier models that prioritize neutrality; while their impartiality is a strength in standard use, it can manifest as a lack of "moral compass" when the system is challenged.

Transparency in failures is critical

Transparency of safety failures, specifically the tendency to conflate technical instability with intentional guardrails,
is another common issue. In many applications, a system crash, API timeout, or "black box" error serves as an accidental safety gate, but this is not a reliable security feature. It leaves both users and developers without a clear feedback loop, making it impossible to determine if the system blocked a prompt for safety reasons or simply failed under the weight of the request. A mature LLM product must move away from these opaque failures and toward a deliberate, rule-based refusal system that clearly communicates the "why" behind a blocked response, ensuring the system is both secure and reliable.

Return on Investment and tangible value of comprehensive safety testing

Whilst comprehensive AI safety testing demands meaningful upfront investment in time, expertise, and resources, its returns extend well beyond the immediate assessment. A well-designed test suite generates reusable artefacts, including test prompts, synthetic datasets, and documented evaluation frameworks. These artefacts accelerate future assessments and reduce the effort required to achieve sufficient coverage for similar GenAI solutions across the business. Beyond efficiency gains, comprehensive testing directly contributes to more resilient product delivery by surfacing edge cases and critical failure scenarios that internal teams may lack the capacity or specialist knowledge to uncover independently. The resulting insights also serve as reusable knowledge assets, supporting responsible design decisions in future AI initiatives. Ultimately, the value of investing in rigorous safety testing should be understood not only as risk mitigation, but as a compounding organisational capability that builds confidence, improves quality, and expands the assurance boundary with each iteration.

AI VERIFY FOUNDATION