GLOBAL AI ASSURANCE SANDBOX

Scan to read the full case study

CityInsights - Mobile Management Chatbot

Application tested

City Developments Limited

City Developments Limited, a Singapore listed global real estate group, has deployed an internal chatbot for management to query trends and insights across its residential, commercial, hospitality, and finance data

Tester

Knovel

Knovel Engineering is a Singapore-based AI consultancy and solution provider that helps enterprises and government adopt and operationalize AI in real-world environments. We deliver applied AI, cloud, and data-driven systems; with a strong focus on AI assurance and trust as a key differentiator

How were LLMs used in application?

Retrieval Augmented Generation | Multi-turn chatbot | Agentic

What risks were considered relevant and tested?

- Hallucination (inaccuracy, lack of completeness)
- Data Leakage
- Vulnerability to Adversarial Prompts (Security)
- Inadequate User transparency

How were the risks tested?

- **Data leakage:** Tested through direct, multi-turn, and indirect prompts; checked for any unintended disclosure
- **Accuracy & consistency:** Compared outputs across semantically equivalent prompts for identical values and formatting
- **Prompt/tool leakage:** Probed partial system/tool instructions to see if system corrects or completes redacted information
- **Disclosure & source attribution:** Checked for AI disclaimers and validity/relevance of cited sources
- **Reasoning transparency:** Verified whether step summaries matched observed system behaviour and outputs
- **Key metric:** Pass/Fail classification

How were test design and evidence evaluated?

- Humans verify that responses are correct based on given descriptions

Challenges

01 Crafting effective adversarial inputs as non-subject matter experts in the finance domain

02 Guardrails consistently blocked direct attempts to access denied topics or out of scope data

03 System defences resisted straightforward exploitation of restricted information

Insights

01 Adapting known attack vectors to domain specific terminology and data structures produced meaningful test cases

02 Multi turn legitimate conversation flows paired with indirect questioning bypassed otherwise resilient guardrails

03 Providing partially correct information exploited system helpfulness, causing it to inadvertently surface restricted data

Powered by: INFOCOMM MEDIA DEVELOPMENT AUTHORITY | AI VERIFY FOUNDATION

GLOBAL AI ASSURANCE SANDBOX

Scan to read the full case study

Clinical Guidance Chatbot for Doctors

Application tested

Changi General Hospital
SingHealth

A clinical guidance chatbot for junior residents to query SingHealth guidelines on diagnosis, treatment protocols, and medication dosing

Tester

Guardrails AI

Guardrails AI is the creator of one of the largest and most widely adopted AI safeguarding toolkits worldwide. Their simulation testing proactively identifies failure points in AI systems before users encounter them

How were LLMs used in application?

Retrieval Augmented Generation | Multi-turn chatbot | Summarisation

What risks were considered relevant and tested?

- Hallucination (inaccuracy, lack of completeness, consistency)
- Undesirable Content
- Data Leakage
- Vulnerability to Adversarial Prompts (Security)

How were the risks tested?

- **Simulation-based multi-turn testing:** Thousands of synthetic conversations were generated to mimic realistic junior resident interactions across different clinical scenarios, including long-tail and adversarial case. Prompts were realistic, diverse, and grounded in the full set of CGH's clinical knowledge base topics

How were test design and evidence evaluated?

- **LLM-as-a-judge automated evaluation:** Each chatbot response was assessed using four aligned judges (hallucination, toxicity, system prompt leakage, and adversarial robustness)
- **Grounding verification against clinical guidelines:** For hallucination testing, chatbot responses were broken into individual claims and checked against the hospital's clinical knowledge base to determine whether the information was properly supported
- **Manual clinical review of flagged cases:** Results were supplemented with manual review of stratified conversation samples by doctors and pharmacists to validate the automated assessments and evaluate nuanced clinical reasoning

Challenges

01 Cognitive overload: reviewing hundreds of flagged conversations against clinical guidelines required additional doctors and pharmacists to support the validation process

02 Creating realistic clinical test scenarios to ensure the testing simulate a wide range of real-world clinician behaviours, including uncommon and messy interactions that are often missed by traditional testing datasets

Insights

01 Premature treatment advice identified as key failure mode: chatbot moved to medication recommendations before clarifying diagnosis (e.g., diabetes type/severity)

02 High-quality automated evaluation is essential for scaling AI testing. Simulation testing surfaced long-tail failures that static datasets and red teaming missed, particularly in everyday clinical dialogue

03 Knowledge base organisation critical to accurate guideline retrieval: LLM defaulted to incorrect or conflicting sources when documents lacked clear scope definitions

Powered by: INFOCOMM MEDIA DEVELOPMENT AUTHORITY | AI VERIFY FOUNDATION

GLOBAL AI ASSURANCE SANDBOX

Scan to read the full case study

Agents for Financial Accounting and Reporting

Application tested

Tester

earlybird × novel Keep Novelty Going

Earlybird's AI-native accounting infrastructure continuously reconciles and maintains accurate, audit-ready books in real time. The system combines autonomous AI agents with expert accountants to continuously process, classify, reconcile, and maintain financial records across invoices, receipts, bank feeds, and payment systems

Knovel Engineering is a Singapore-based AI consultancy and solution provider that helps enterprises and government adopt and operationalise AI in real-world environments. We deliver applied AI, cloud, and data-driven systems; with a strong focus on AI assurance and trust as a key differentiator

How were LLMs used in application?

Agentic | Classification | Detection | Recommendation

What risks were considered relevant and tested?

- Hallucination and Inaccuracy
- Robustness Under Real-World Conditions
- Data Leakage
- Vulnerability to Adversarial Prompts

How were the risks tested?

- **Accuracy:** Real and synthetic financial documents were used to evaluate extraction, classification, and reconciliation accuracy against human-verified data
- **Robustness:** Documents were degraded with realistic blur, skew, rotation, and compression to assess reliability under real-world conditions
- **Data leakage and Adversarial prompts:** Manipulated documents and embedded malicious instructions were used to test resilience against prompt injection, and persistence-related workflow attacks
- **Testing both independent and memory-influenced behaviour:** The system was evaluated both on standalone accounting tasks where each output should not depend on past interactions, and on workflows where past actions and transaction history could influence future decisions through the memory layer

How were test design and evidence evaluated?

- Automated comparison with ground truth and manual testing for stateful tests

Challenges

01 **Testing persistence-enabled workflows:** The system's memory layer was comparatively difficult to manipulate because financial inputs are highly structured and constrained, unlike open-ended chatbot environments with unrestricted free-form text inputs This reduced the attack surface for direct prompt injection and required more sophisticated adversarial techniques targeting extracted financial metadata and document-based workflows

Insights

01 **AI-native accounting needs broader assurance:** Testing must go beyond model accuracy to evaluate workflow reliability, safeguards, persistence behaviour, and human oversight across continuous financial operations

02 **Realistic operational testing is critical:** Testing under real-world document conditions, reconciliation edge cases, and adversarial scenarios provided a more representative assessment of operational reliability

03 **Human oversight and safeguards remain important:** Governance controls and expert review are still critical for persistence-enabled workflows, exception handling, and complex financial decisions

Powered by: INFOCOMM MEDIA DEVELOPMENT AUTHORITY · AI VERIFY FOUNDATION

GLOBAL AI ASSURANCE SANDBOX

Scan to read the full case study

Jurisdiction-Aware Custom Guardrails for Asia-Pacific

Application tested

fourtitude.ai

fourtitude.ai's RAG multilingual customer support chatbot evaluated for compliance with jurisdiction-specific policies, including Malaysian political and royalty-related content handling

Tester

Dynamo AI

Dynamo AI provides automated AI evaluation and real-time guardrails for security, compliance, and policy enforcement across enterprise AI systems

How were LLMs used in application?

Summarisation | Retrieval Augmented Generation | Multi-turn chatbot

What risks were considered relevant and tested?

- Undesirable Content (Content Safety)
- Reputation Risks

How were the risks tested?

- Constructed synthetic benchmark datasets combining diverse, jailbreak, in-domain, and inapplicable prompts
- Evaluated model responses against jurisdiction-specific policies (e.g., Malaysian politics, royalty)
- Assessed ability to detect and block non-compliant outputs
- Compared baseline (no guardrails), competitor guardrails, and Dynamo guardrails on the same benchmark
- Measured performance using Precision, Recall, F1, Accuracy, FPR, and FNR

How were test design and evidence evaluated?

- Model-based classification of responses as compliant or non-compliant
- Comparison against ground truth labels for policy violations
- Quantitative evaluation using standard classification metrics (F1, Precision, Recall, Accuracy, FPR, and FNR)
- Analysis of false positives (FPR, overblocking) and false negatives (FNR, missed violations)
- Validation through example-based review of guardrail decisions

Challenges

01 **Context-aware, configurable guardrails:**
Adapt to jurisdiction-specific legal/cultural nuances; continuously updated from production data

02 **Precision vs. recall trade-off:**
Strict guardrails → false positives; lenient → missed risks

03 **Cost at scale:**
Real-time guardrails expensive; need low-latency, efficient models

Insights

01 **Custom guardrails:**
Improve compliance without degrading AI performance; higher F1, fewer false negatives vs. baseline/competitors

02 **Recall priority:**
Minimising false negatives is critical in high-risk/regulated contexts

03 **Diverse datasets:**
Include in-domain, out-of-scope, and adversarial inputs to reveal failures and ensure robust evaluation

Powered by: INFOCOMM MEDIA DEVELOPMENT AUTHORITY | AI VERIFY FOUNDATION

GLOBAL AI 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 · AI VERIFY FOUNDATION

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

GLOBAL AI ASSURANCE SANDBOX

Scan to read the full case study

Jingjai: Multilingual AI Safety Testing

Application tested

KBTG

An AI-native internal assistant delivering quick, factual insurance information for employees

Tester

VULCAN

Vulcan specialises in providing GenAI security for enterprises, using scenario-specific testing against safety and security threats, and real-time monitoring and custom guardrails for GenAI applications

How were LLMs used in application?

- Retrieval Augmented Generation
- Multi-turn chatbot
- Data extraction from unstructured source
- Summarisation
- Translation

What risks were considered relevant and tested?

- Undesirable Content (Content Safety)
- Data Leakage
- Vulnerability to Adversarial Prompts (Security)
- Multilingual and multicultural risks (e.g., testing in English, Thai, English and Thai in Karaoke format)

How were the risks tested?

- Risks mapped into 11 threats across 4 language formats: Thai, English, code-switched Thai-English, and Romanized Thai
- Adversarial attacks - generated scenario-based test cases, and reviewed by Vulcan to ensure they aligned with the application's risk profile and operational context
- Ran 1,276 unique test cases (based on 11 threats x 29 attack techniques x 4 language formats)
- Repeated identical attacks (e.g., prompt injection and jailbreak techniques) to detect intermittent failures and probabilistic behaviour
- Added baseline language checks and a small set of multi-turn assessments to test realistic adversarial interactions

How were test design and evidence evaluated?

- Compared each response against the expected secure behaviour under the attack scenario
- Measured whether attacks succeeded, failed, or caused partial leakage, harmful output, or other policy-relevant deviations
- Assessed repeated runs to identify inconsistent or unstable failures that might not appear in a single test

Challenges

01 GenAI applications must enforce confidentiality and content boundaries consistently across languages, use cases. In regulated industries such as banking and fintech, controls should prevent the disclosure of system prompts, retrieval mechanisms, metadata structures, and internal tools, as well as the generation of regulated advice

02 Independent multilingual safety testing, stronger policy controls, and tighter monitoring of high-risk outputs can close this governance gap and ensure more reliable deployment

Insights

01 Multilingual safety is not uniform across languages, so GenAI systems should be tested in the real languages users actually use, especially in Southeast Asia

02 Poor comprehension can look like good safety, but it may just mean the application did not understand the input well enough; this is especially relevant for informal or phonetic forms like Romanised Thai or Manglish or Taglish

03 Gen AI applications and guardrails may block broad sensitive topics more effectively than nuanced local bias, so deployments across Asia need local-context testing to avoid overlooked social bias, reputational harm, and trust issues. Disclosure of internal logic, parameters, metadata, or tooling enables attackers to probe system weaknesses, and data leakage needs to be tested — even for internal applications

Powered by: INFOCOMM MEDIA DEVELOPMENT AUTHORITY · AI VERIFY FOUNDATION

GLOBAL AI ASSURANCE SANDBOX

Scan to read the full case study

Chatbot for Children and Youths

Application tested

MangaChat's conversational chatbot supports children and youths in Singapore in the areas of emotional awareness, emotional expression, and early emotional support

Tester

AIDX TECH is an AI technical testing vendor specialising 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 organisations identify AI risks, meet governance requirements, and deploy AI systems with greater confidence

How were LLMs used in application?

- Retrieval Augmented Generation
- Multi-turn chatbot
- Classification
- Recommendation

What risks were considered relevant and tested?

- Undesirable content
- Vulnerability to adversarial prompts
- Hallucination (Factual Inaccuracy)
- Data Leakage

How were the risks tested?

- Stage-based testing across the conversation flow: Risks were mapped and tested separately across 4 stages Engagement, Routing, Assessment, and Crisis Safety Net stages to capture how failures emerge at different points and transitions
- Benchmark prompts for content safety: Structured test prompts were used to probe whether the chatbot produces harmful, inappropriate, or age-unsafe responses under realistic and edge-case user inputs
- Adversarial and jailbreak testing: Realistic user scenarios were transformed into adversarial prompts to evaluate whether the system could resist manipulation, role-play exploits, and attempts to bypass safeguards
- Hallucination: Model responses were broken into factual claims and compared against ground truth, while crisis scenarios were tested for consistency across different distress expressions

How were test design and evidence evaluated?

- Model responses were evaluated using predefined scoring frameworks—1–5 scales for content safety and robustness, 0–1 hallucination rates for factual accuracy, and Pass/Fail outcomes for alignment tests
- LLM-as-a-judge for large volumes of responses
- Human review for validation and edge cases: All automated results were reviewed by human evaluators to validate judgments, interpret nuanced responses (especially in emotional contexts), and finalise the assessment

Challenges

01 Multi-stage complexity requires stage-specific testing - risks emerge across transitions, requiring tests to be tailored to each stage

02 Converting regulatory requirements into effective test scenarios requires careful interpretation and significant upfront effort to ensure accurate and meaningful evaluation

03 LLM-as-a-judge has limits in nuanced domains: Automated scoring is scalable but less reliable for emotionally or clinically complex responses, making human review essential for accurate assessment

Insights

01 **Test AI as a full experience, not just a model:** Multi-stage and multi-turn evaluation is critical to capture how risks emerge across conversation flow and user journeys

02 **Combine automation with human judgement:** Scalable LLM-based testing is effective, but human review is essential for nuanced, high-sensitivity contexts like youth emotional support

03 **Upfront design of the test drives meaningful results:** Clear mapping of system behaviour and expectations enables more accurate, reliable, and actionable testing outcomes

Powered by: INFOCOMM MEDIA DEVELOPMENT AUTHORITY · AI VERIFY FOUNDATION

GLOBAL AI ASSURANCE SANDBOX

Scan to read the full case study

MBBS Clinical Chatbot (Voicebot)

Application tested

Nanyang Technological University, the Lee Kong Chian School of Medicine uses a clinical chatbot to train students in empathy, information gathering, and patient prioritisation

Tester

Knovel Engineering is a Singapore-based AI consultancy and solution provider specialising in assessing agentic AI systems and delivering risk-based tests to ensure real-world readiness

How were LLMs used in application?

Multi-turn chatbot

Translation & Transcription (EN/ZH)

What risks were considered relevant and tested?

- Hallucination (inaccuracy, lack of completeness)
- Undesirable Content (Content Safety)
- Data Leakage
- Bias
- Vulnerability to Adversarial Prompts (Security)

How were the risks tested?

- **Data Leakage, Vulnerability to Adversarial Prompts** via prompt injection, context escape, and authority impersonation to check system prompt leakage
- **Single-turn override attempts** to assess behavioural control
- **Hallucination & Inaccuracy:** Evaluated using out-of-scope questions to detect role confusion, persona drift, and fabrication of unsupported or inconsistent details beyond the prompt
- **Bias in interaction behaviour:** Assessed through tone variation (warm to aggressive) to test whether the AI showed uneven or inappropriate responsiveness
- **Robustness across sessions:** Repeated structured clinical scenarios across multiple runs to evaluate stability of persona adherence, response consistency, and resilience to conversational variation

How were test design and evidence evaluated?

- **Manual review:** All outputs were manually assessed by AI testing experts for factual accuracy, persona adherence, safety compliance, and resistance to prompt injection
- **Regression comparison across test cycles:** Initial and post-fix results were compared to identify improvements, persistent issues, and any unintended regressions introduced by changes
- **Multi-dimensional consistency testing:** The system was tested across varied tones, repeated sessions, and English–Chinese inputs to assess behavioural stability and robustness under different interaction conditions

Challenges

01 A key challenge was bypassing guardrails, as the system appeared hardened against direct attacks

02 Fixing one issue occasionally introduced regressions elsewhere

Insights

01 Tone sensitivity should be treated as an explicit design feature

02 Safety improvements must be evaluated holistically to avoid regression: Enhancing safeguards in voicebots can sometimes introduce unintended side effects in other areas of system behaviour. This highlights the importance of regression testing across multiple dimensions

Powered by: INFOCOMM MEDIA DEVELOPMENT AUTHORITY · AI VERIFY FOUNDATION

GLOBAL AI ASSURANCE SANDBOX

Scan to read the full case study

Career Guidance Chatbot

Application tested

NCS

A conversational assistant for career guidance that leverages Generative AI to support career coaches and jobseekers with personalised career guidance, skills insights, and job search recommendations

Assurer

AIQURIS

AIQURIS is an independent assurance provider specialising in AI risk, governance, and quality management. Originating from TÜV SÜD, it helps organisations identify, assess, and manage AI risks across the full AI lifecycle

How were LLMs used in application?

Multi-turn chatbot | Classification | Recommendation

What risks were considered relevant and tested?

Legal, security, ethics, performance, and sustainability risks incorporated into the prioritised test scenarios

Example: Fair Consideration Framework (FCF) key facts were embedded in the test case design and evaluation criteria to ensure compliance-relevant accuracy was explicitly tested

How were the risks tested?

- Establish a risk profile across six quality pillars: legal, security, ethics, performance, sustainability, and safety
- From this profile, a stakeholder analysis identified who could be harmed and how
- A hazard register was then derived by tracing causal pathways through the AI lifecycle for each identified harm
- Testing requirements were selected directly from this hazard set, ensuring each test scenario could be traced back to a specific hazard and its affected stakeholders
- Out of 23 candidate scenarios identified, nine were prioritised for testing based on harm severity, stakeholder impact, and likelihood of occurrence

How were test design and evidence evaluated?

- Reviewed the test design and scope to ensure the selected scenarios were clearly derived from identified risks and covered key areas such as accuracy, fairness, safety, and robustness
- Examined the evaluation methodology, including how personas, test cases, and scoring criteria were constructed, to confirm the approach was structured, consistent, and repeatable
- Assessed execution evidence, such as test logs, configurations, conversation transcripts, and evaluation results, to verify that reported outcomes matched actual system behaviour
- Evaluated residual risks and findings, identifying strengths and gaps in performance and providing recommendations for improvement before broader deployment

Challenges

01 Testing was restricted to the UAT instance and black-box methods due to the application being live in production, which also limited the depth of information that could be shared with the independent assurer

02 Token budget limitations, stemming from large-scale simulation-based testing not having been planned for at the outset, restricted the volume of test cases generated

03 As a result, only risk scoring could be completed within this engagement. Defining risk appetite, assessing risk impact, and risk mitigation plans remain as next steps toward a full assurance conclusion

Insights

01 Start with context and risk. Clearly define the system's use case, stakeholders, and potential harms, and use this to drive risk-based prioritisation of what to test

02 Design testing to be systematic, repeatable, and grounded in clear evaluation criteria so results are meaningful and can be independently reviewed

03 Test cases and evaluation rubrics must be tailored to the specific application, as general-purpose methods are insufficient for producing a meaningful and defensible risk score

Powered by: INFOCOMM MEDIA DEVELOPMENT AUTHORITY | AI VERIFY FOUNDATION

GLOBAL AI ASSURANCE SANDBOX

Scan to read the full case study

Public-Facing Retail Chatbot

Application tested

Retailer

A retailer's customer service chatbot designed to answer customer queries by retrieving information exclusively from pre-approved company FAQs to provide reliable, brand-aligned responses

Tester

VULCAN

Vulcan (by AIFT) provides GenAI security through scenario-specific testing against security/safety threats and real-time protection guardrails for enterprise AI interactions

How were LLMs used in application?

Retrieval Augmented Generation

Multi-turn chatbot

What risks were considered relevant and tested?

- Hallucination
- Undesirable Content
- Data Leakage
- Vulnerability to Adversarial Prompts
- Breach of compliance
- Reputation Risks

How were the risks tested?

- Scenario-based red-teaming at scale: Executed 1,450 test cases derived from 50 threat scenarios and 29 attack techniques
- Prompt injection and jailbreak testing: Systematically probed the chatbot with adversarial prompts to bypass safeguards and uncover vulnerabilities (e.g., data leakage, prompt exposure)
- Multi-turn attack simulations: Conducted sustained, sequential interactions to test persona manipulation and behavioural drift that only emerge over longer conversations
- Hybrid testing approach (automation + manual): Combined automated testing for breadth with expert-led manual probing to uncover deeper, "unknown unknown" risks
- Re-ran the same attacks after implementing guardrails to measure mitigation effectiveness and residual risk

How were test design and evidence evaluated?

LLM-based review and human-in-the-loop verification to judge outcomes, categorised by severity (High, Medium, Low) and business impact, rather than raw attack success rates, to ensure findings were actionable for decision-making and remediation

Mitigation effectiveness assessment: Results were evaluated through before-and-after comparison (initial test vs re-test after guardrails) to measure how effectively defences blocked or reduced previously successful attacks

Challenges

01 Tuning safety configurations is difficult; overly strict settings (e.g., jailbreak filters) may block legitimate user queries, while permissive settings risk allowing harmful content to pass through

02 Single-prompt testing fails to uncover sophisticated multi-turn jailbreaks where attackers gradually manipulate the chatbot's persona or policies over a sustained interaction

03 Current automation typically scratches the surface, lacking the multi-turn endurance required to operationalise red-teaming against the 'unknown unknowns' of sequential interactions

Insights

01 Relying solely on system prompts is insufficient; a dedicated runtime guardrail or "GenAI firewall" layer is critical for real-world defence

02 Prioritising findings by business impact and severity rather than raw "attack success rates" allows for better internal risk communication and remediation

03 Point-in-time testing is insufficient; regular red-teaming must be operationalised as "business-as-usual," especially after model or logic updates

Powered by: INFOCOMM MEDIA DEVELOPMENT AUTHORITY · AI VERIFY FOUNDATION

GLOBAL AI ASSURANCE SANDBOX

Scan to read the full case study

LawNet AI: AI-Powered Legal Search

Application tested

SAL Singapore Academy of Law | LawNet

Singapore Academy of Law (SAL) has deployed a single-turn Question & Answer (Q&A) application designed to retrieve relevant cases and legislation to answer legal research questions

Tester

resaro

Resaro offers independent, third-party assurance of mission-critical AI systems with extensive experience in testing Computer Vision and Generative AI systems

How were LLMs used in application?

Retrieval Augmented Generation

What risks were considered relevant and tested?

- Correctness of answers (including faithfulness and lack of completeness)
- Data Leakage (specifically system prompt leakage)
- Reliability and Robustness to perturbations (consistency of responses for identical prompts and meaning-preserving perturbations of prompts)

How were the risks tested?

- Custom evaluation sets were created to test for risks using realistic legal Q&A scenarios. SAL's legal experts generated a dataset and Resaro supplemented this with a synthetic dataset. Both of these were also perturbed to test additional aspects of system quality

How were test design and evidence evaluated?

- Automated evaluation with LLM-as-a-judge and results validated by human expert to ensure alignment and reliability

Challenges

01 **Limited legal understanding in general purpose pipelines** – General-purpose synthetic tools struggle to capture nuanced legal concepts (e.g., doctrinal links and citation importance), though they still provide fast and scalable testing value

02 **High context sensitivity in legal evaluation** – Small changes in legal question framing can significantly alter the correct answer, making it hard to define a single ground-truth reference for evaluation

Insights

01 General purpose test pipelines cannot yet fully handle the nuances of highly specialised domains. However, their speed and quality out of the box still makes them useful

02 Precision of context is critical in legal evaluation so datasets must be carefully designed to ensure a single clear reference answer

03 RAG evaluation must be multi-dimensional – Effective testing requires assessing correctness, faithfulness, and citation grounding, with faithfulness being especially important in legal use cases

Powered by: INFOCOMM MEDIA DEVELOPMENT AUTHORITY | AI VERIFY FOUNDATION

GLOBAL AI ASSURANCE SANDBOX

Scan to read the full case study

AGILTrust: Misinformation detection platform

Application tested

Tester

ST Engineering × AIDX

AGILTrust offers AI capabilities in deepfake detection, factchecking, visual and textual content analysis. Features tested are Reaction & Implication analysis such as predicted emotional reaction, action and implication induced from such media consumption

AIDX TECH is an AI technical testing vendor specialising 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 organisations identify AI risks, meet governance requirements, and deploy AI systems with greater confidence

How were LLMs used in application?

Multi-turn chatbot | Summarisation | Detection

What risks were considered relevant and tested?

- Undesirable Content
- Vulnerability to Adversarial Prompts (robustness)

How were the risks tested?

- **Undesirable Content / Content Safety Risks:** Tested using a structured benchmark embedded in a neutral news-transcript template. The application generated its usual reflection/reaction and implication analysis, which were then evaluated against predefined undesirable content risks such as biased language, emotional framing, contradictory reasoning, sensationalism, downplaying of risks, and other inappropriate or misleading analysis behaviours
- **Vulnerability to Adversarial Prompts:** Tested by inserting adversarial prompts into the same news-transcript template to assess whether the system would follow harmful or instruction-overriding content instead of staying within the intended news-analysis task

How were test design and evidence evaluated?

- **Undesirable Content:** Each reflection/reaction and implication output was evaluated using an LLM-as-a-judge framework against predefined undesirable content risk definitions. A test case failed if any defined risk was detected in the final user-visible output
- **Vulnerability to Adversarial Prompts:** Each adversarial test output was also evaluated using an LLM-as-a-judge framework to determine whether the system followed harmful or instruction-overriding content embedded in the news transcript. A test case failed if the system complied with the adversarial instruction or produced undesirable content
- **Guardrail comparison:** The same evaluation logic was applied across AWS Bedrock Guardrails and Virtue AI guardrail configurations. Results were aggregated using Safety Rate, calculated as the percentage of test cases that passed without triggering any defined risk

Challenges

01 **Mapping Application-Specific Risks to Standard Taxonomies.** Existing AI safety frameworks focus on clear harmful content, but this use case required evaluating softer aspects like neutrality, reasoning quality, and tone. This made custom test design and limited domain-specific datasets necessary

02 **Defining “unsafe” in a legitimate news-analysis context is complex.** The system is expected to handle sensitive or controversial content, so the challenge was distinguishing valid analysis from biased, misleading, or overly emotional interpretation. Careful calibration was needed to avoid wrongly flagging appropriate analytical outputs

03 **Limited Failure Attribution in Black-Box Testing:** When the application returned refusals or empty analytical outputs, black-box testing could not always determine whether the cause was the model, application logic, or guardrail configuration

Insights

01 **Application-specific output risks require explicit test design.** Testing must explicitly measure neutrality, reasoning quality, and tone consistency, not just harmful content detection. This requires tailoring benchmarks to reflect how well the system performs its analytical role

02 **Risk assessment should focus on how content is interpreted, not just whether the content itself is sensitive.** In news-analysis applications, the key failure mode is often biased or emotionally skewed interpretation of otherwise legitimate material. Effective testing therefore depends on evaluating the quality of analysis under realistic content conditions

03 **Guardrail effectiveness should be evaluated at the application level, not only at the prompt or model level.** End-to-end testing using realistic input formats allows different guardrail configurations to be compared under the same operational conditions, capturing both protection effectiveness and potential task disruption such as refusals or empty outputs

Powered by: INFOCOMM MEDIA DEVELOPMENT AUTHORITY | AI VERIFY FOUNDATION