GLOBAL AI ASSURANCE PILOT

Scan to read the full case study.

# UltraScale No-code AI-powered RAG Platform

### How LLMs were used in application?

Summarisation | Retrieval augmented generation | Data extraction from unstructured source | Translation | Multi-turn chatbot

## Application Tested

ultra mAinds

ultra mAInds, an AI engineering and solutions company, offers UltraScale, a 'No-Code AI-powered RAG platform for Enterprise search and data connectivity.

## Tester

AIQURIS

AIQURIS (assurance partner) is a Singapore-based corporate venture offering a SaaS platform that helps specify and manage the required evidence for confident adoption of AI. The platform is based on a systematic approach to establish risk profile and necessary controls to ensure AI systems are deployed safely, compliantly, and effectively at scale.

AIDX

AIDX Tech (testing partner) is an AI testing platform that offers model evaluation, safety and risk management, and consulting services.

## What Risks Were Considered Relevant And Tested?

- Accuracy of translation
- Potential harmful content generation across diverse languages

## How Were The Risks Tested?

### Approach

**Accuracy of translation**

FLORES+ benchmark dataset containing 997 sentences translated from English to other languages as well as between non-English language pairs

**Harmful content generation**

AIDX GenAI Safety Benchmark, targeting Robustness, Ethics and Society, Fairness, Privacy and Security, Toxicity, and Legality

### Evaluators

**Accuracy of translation**

- BLEU Score and Sentence Embedding Similarity metrics were calculated using automated testing
- A subset of translations was manually reviewed to verify nuanced accuracy in grammar, meaning, and context

For harmful content detection, success rate in the face of adversarial prompts was assessed

## Challenges

### API Performance

Latency issues, averaging approximately 1,000 translations per hour (internal safeguarding of API abuse but poses a potential performance bottleneck for testing)

### Stability Issue

After prolonged test runs, instabilities were encountered indicating the need to individually monitor third party services (e.g. resource exhaustion, connectivity)

### Translation Reliability

Some translations failed intermittently without a clear pattern, again raising the need to monitor performance of individual software components

## Insights

01

**Challenges in designing and implementing automated tests for multi-lingual accuracy and content safety:**

- Lack of standardised translation ground truth
- Inadequacy of surface level metrics like BLEU
- Inability to catch nuanced or implicit harms
- Limitations of using API calls for large-scale multi-lingual testing

02

**Importance of structured risk assessment process to determine what to test**

03

**Strong role for human experts as a result of above challenges**

Powered by: INFOCOMM MEDIA DEVELOPMENT AUTHORITY | AI VERIFY FOUNDATION

GLOBAL AI ASSURANCE PILOT

Scan to read the full case study.

# Assure.ai Customer Service Chatbot

| Application Tested | Tester |
|---|---|
| fourtitude.asia | AIDX |
| Fourtitude.ai is a leading systems integrator company that has developed Assure.ai, a GenAI Chatbot. It is intended to help its clients answer enquiries from customers or citizens regarding their service offerings. | AIDX Tech is a trustworthy AI model testing platform for AI risks, safety and reliability testing, verification and risk management. |

### How LLMs were used in application?

- Summarisation
- Retrieval augmented generation
- Data extraction from unstructured source
- Translation
- Multi-turn chatbot
- Classification or recommendation

## What Risks Were Considered Relevant And Tested?

- Sensitivity to cultural, religious and racial matters: particularly in the context of local laws and practices
- Accuracy and friendliness to users (prioritised but not tested during the pilot)

## How Were The Risks Tested?

### Approach

**Red teaming**

- Based on customer-provided seed prompts across various high-risk topics for 4 domains – cultural, racial, religious, and general safety
- Applying 10 structured attack methods – e.g., Instruction Jailbreak, Goal Hijacking, Deep Inception

### Evaluators

- Model outputs evaluated based on attack success rate
- Safety score automatically computed based on whether the model's response to each adversarial prompt met predefined unsafe or undesired behaviour criteria
- Expert human review only for high-risk, low-score responses to ensure cultural and contextual accuracy

## Challenges

Extracting culturally relevant knowledge specific to the Malaysian context. This was addressed by Fourtitude.ai providing a curated test set to act as seed for AIDX

## Insights

01

Improved test relevance and efficiency through customer-provided seed prompts

02

Importance of API access for automated testing

03

Grey zone between acceptable and unacceptable answers, with boundary dependent on cultural norms and legal precedents in the country of operation

04

Difficulty of interpreting attack success rate in isolation without comparative baselines

Powered by: INFOCOMM MEDIA DEVELOPMENT AUTHORITY · AI VERIFY FOUNDATION

GLOBAL AI ASSURANCE PILOT

Scan to read the full case study.

# HealthHub AI Conversational Assistant

**Application Tested**

synapxe

Synapxe, Singapore's national HealthTech agency has a Retrieval augmented generation-based Gen AI conversational assistant that allows users to search and receive health information, based on HealthHub's website content.

**Tester**

AIDX

AIDX TECH, a Singapore-based AI assurance specialist startup, has an in-house proprietary platform which supports benchmarking and adversarial red-teaming of GenAI applications across dimensions like robustness, ethics, privacy, toxicity and security.

### How LLMs were used in application?

- Summarisation
- Retrieval augmented generation
- Data extraction from unstructured source
- Translation
- Video or audio to text
- Multi-turn chatbot

## What Risks Were Considered Relevant And Tested?

- Safety and Health: Physical harm and/or negative mental health outcomes
- Fairness: Chatbot output must not discriminate unfairly against particular groups in the information presented
- Malicious use: e.g., causing adverse health outcomes or physical harms to individuals
- Trust/reputation concerns: inaccurate or inappropriate output that causes embarrassment

The testing focused on evaluating the safety, robustness, and compliance of Synapxe's AI conversational assistant

## How Were The Risks Tested?

**Approach**

**Safety (toxicity and wellbeing)**

AIDX uses benchmark testing across 2 dimensions with 5 sub-categories— Ethics and society (Mental health, Physical health), Toxicity (Threaten and Intimidate, Abusive Curses, Defamation)

**Robustness**

Adversarial red teaming across 14 red teaming attack methods (e.g., unsafe self-medication, false symptom interpretation)

**Evaluators**

- LLMs as a judge
- Non-LLM based classifiers

A five-point scale was used to assess responses to "out of policy" or inappropriate requests

## Challenges

- Cybersecurity and data privacy considerations: Requires secure testing environments and strict adherence to healthcare data protection standards
- Latency and throughput limitations: May increase the timing of multi-turn agent-based testing via API

## Insights

01

Fixed or universal test sets inadequate in capturing the dynamic and context-specific nature of real-world AI apps

02

Synthetic adversarial prompts, while useful for stress testing, may not always resemble actual user behaviour

03

Testing AI models differs significantly from testing deployed AI applications (e.g., due to complex APIs and integrated components beyond the models)

04

Stability and standardisation of API interfaces can directly impact the ease and scalability of test execution

Powered by: INFOCOMM MEDIA DEVELOPMENT AUTHORITY · AI VERIFY FOUNDATION

GLOBAL AI ASSURANCE PILOT

Scan to read the full case study.

# AskMax – Virtual Concierge Chatbot

### How LLMs were used in application?

Retrieval augmented generation

Multi-turn chatbot

## Application Tested

CHANGI airport group

Changi Airport Group (CAG), the operator of Singapore Changi Airport, deployed AskMax to assist travellers and visitors with airport-related queries.

## Tester

PRISM Eval specialises in adversarial safety evaluations of generative AI. Its Behaviour Elicitation Tool (BET) systematically probes LLM applications for policy violations and robustness gaps.

Guardrails AI

Guardrails AI is the creator of one of the largest and most widely adopted AI safeguarding toolkits worldwide. Their simulation testing can help proactively identify failure points in an AI system before users encounter them.

## What Risks Were Considered Relevant And Tested?

User safety, Public trust and Reputational integrity. Translated into:

- Content safety risks in 6 behavioural categories:
  (a) Mis/Disinformation;
  (b) Social Engineering & Manipulation;
  (c) Hate & Discrimination;
  (d) Illegal Activities & Contraband;
  (e) Exploitation & Abuse; and
  (f) Violence & Physical Harm
- Hallucination risk

The risk of over-refusal – refusing to provide answers even when the question is relevant and the chatbot has access to appropriate knowledge base – was also considered relevant

## How Were The Risks Tested?

### Approach

Automated Red Teaming using PRISM Eval's BET to explore and map the chatbot's response patterns across the six key areas

Simulation testing using Guardrails AI's synthetic prompt capability that emulated real customer interactions. Prompts needed to be realistic, diverse and grounded in the full set of CAG's knowledge base topics

### Evaluators

LLM as a judge

- PRISM Eval: LLM-based judge on a standardised compliance scale from -2 (complete refusal) to +4 (detailed harmful information). AskMax benchmarked against base LLMs from their LLM robustness leaderboard
- Guardrails AI: One automated judge for each of the test categories, tuned to match criteria discussed with CAG

## Challenges

### Automated Red Teaming

- Lack of standard robustness benchmarks for testing and interpreting the results for public-facing chatbots
- Mid-test platform upgrades during pilot

### Simulation Testing

- Discovering the unknown: Understanding where chatbot performed well vs not
- Theory vs Reality: Where are the real risks and how best to stress-test for them

## Insights

01

Context specific techniques are essential when assessing real-life LLM apps (vs models). Realistic and diverse test data is critical to enable context specificity

02

Close collaboration between internal stakeholders is essential – e.g., AI governance, IT, and business teams and external AI assurance specialists

03

Early identification of key areas of concern helps focus and structure evaluation. Simulation testing allows some flexibility mid-way in the testing process

04

Aligning automated judges with human expectations is essential to enable evaluation at scale

Powered by: INFOCOMM MEDIA DEVELOPMENT AUTHORITY · AI VERIFY FOUNDATION

GLOBAL AI ASSURANCE PILOT

Scan to read the full case study.

# Medical Report Summarisation

### How LLMs were used in application?

Summarisation | Classification and recommendation | Data extraction from unstructured source

## Application Tested

Changi General Hospital
SingHealth

Changi General Hospital (CGH), a major public hospital in Singapore, created a GenAI application to summarise colonoscopy and histopathology reports based on US Multi-Society Task Force (USMTF) guidelines. The goal was to provide clinicians with structured, guideline-based summary recommendations.

×

## Tester

softserve

SoftServe is a global digital consulting company specialising in software development, data science, and AI/ML solutions.

## What Risks Were Considered Relevant And Tested?

**Relevant risks:**

- Accuracy of Clinical Summaries
- Clinical Safety and Alignment to Guidelines

CGH was also interested in assessing the reliability of its existing automated evaluation framework (which used a well-known, open source summarisation metric) for real-world deployment

## How Were The Risks Tested?

### Approach

**Through back testing on historical data, annotated with ground truth by subject matter experts (SMEs). Evaluation was split into 2 parts:**

- Comparison of key facts extracted from the medical reports vs the ground truth facts from SME
- Comparison of final recommendation (arrived at by applying above facts to the USMTF guidelines) to the ground truth recommendation from SME

### Evaluators

**Rule based logic:**

Python scripts used both to extract the key facts from CGH's application output, and to simulate the USMTF guidelines

## Challenges

- Getting SMEs to provide a scoring rubric ("what good looks like") that covers all potential scenarios
- Obtaining ongoing support from (busy) SMEs to test incremental improvements
- Constraints imposed by the design of the GenAI application: e.g., can determine whether granular interim data is available for testing, or whether the application can actually be improved based on test results

## Insights

01

Evaluations that include confidence scores (e.g., in accuracy of extracted key facts) can help increase overall level of automation. Where the system suggests lower confidence scores, the case can be auto routed to the human in the loop

02

Using LLMs as judges need not be the only option – particularly when the evaluation can be designed with granular interim checkpoints

03

Adding detailed citations (source text or code fragment for a particular fact or recommendation) to the system's output design can make human review and annotation easier

04

Measuring accuracy of interim outputs results in greater transparency and suggestions to improve original system design

Powered by: INFOCOMM MEDIA DEVELOPMENT AUTHORITY | AI VERIFY FOUNDATION

GLOBAL AI ASSURANCE PILOT

Scan to read the full case study.

# On-demand Scam and Online Fact-checker Using Agentic Workflow

## Application Tested

CheckMate

CheckMate is a volunteer-run grassroots initiative that aims to make the act of checking information easy and accessible for all. Its WhatsApp service (powered by an LLM agent) allows users to send in dubious content they encounter online, and supports text messages, images, screenshots or links. Able to search online, visit webpages, and/or scan for malicious URLs.

## Tester

Advai

Advai is a UK based AI assurance company, focused exclusively on the testing, evaluation and assurance of AI. They collaborate with organisations including the UK Government's National Cyber Security Centre, the Ministry of Defence, listed companies & leading Systems Integrators.

### How LLMs were used in application?

- Summarisation
- Retrieval augmented generation
- Data extraction from unstructured source
- Translation
- Orchestrator for an agentic flow

## What Risks Were Considered Relevant And Tested?

- Inaccuracy during normal usage conditions
- Subtle adversarial attacks that influence the output of the system e.g., scammers making subtle tweaks to messages
- Adversarial attacks that bring down the system
- Generation of harmful content: Out of distribution inputs that cause the system to generate and output harmful content

## How Were The Risks Tested?

### Approach

- Isolating Modalities: Testing separately for text and image inputs
- Benchmarking using internal and opensource datasets, covering denial of service, harmful content, misclassification
- Manual red teaming by human experts
- Scaled Semi-Automated Testing, LLMs were used to generate synthetic data based on human crafted datasets

### Evaluators

- Combination of human review and some automated options
  - Pre-trained classifiers
  - LLM as a judge
  - RegEx pattern matcher to determine classification given by the CheckMate system

## Challenges

- Generating and evaluating in distribution datasets aimed at misclassification
  - Pre-defined benchmarks, though providing useful insight, proved not to be a strong indicator of overall robustness
  - Scaling the generation of input datasets, which retained the initial features to be tested, was time consuming

## Insights

01

Design of GenAI application can influence resilience to attack (e.g., siloing inputs and outputs for review, restricted input/output space and single-turn interactions)

02

In sensitive externally facing applications, there is a trade-off between safety and transparency to build confidence

03

Benchmarks aren't (always) an accurate reflection of robustness

04

Fooling both AI and humans creates an additional challenge

05

Clear understanding of business risks makes testing more effective

Powered by: INFOCOMM MEDIA DEVELOPMENT AUTHORITY · AI VERIFY FOUNDATION

GLOBAL AI ASSURANCE PILOT

Scan to read the full case study.

# AI-enabled Candidate Screening and Evaluation (HR)

| Application Tested | Tester |
|---|---|
| MIND-Interview | FAIRLY Assurance for AI |
| Mind Interview, a Taiwan-based HR AI tech startup, has a Candidate Assessment and Screening tool that helps employers screen and evaluate job candidates. | 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

GLOBAL AI ASSURANCE PILOT

Scan to read the full case study.

# Productivity Co-pilot

**Application Tested**

HTX

Home Team Science & Technology Agency (HTX) is a government agency dedicated to using science and technology to enhance public safety and security in Singapore. The Co-pilot enables internal stakeholders to engage in natural, multi-turn conversations to quickly extract insights from uploaded data sources.

**Tester**

Knovel Keep Novelty Going

Knovel Engineering is an industry partner of HTX, it is a technology consulting and solution provider covering areas across artificial intelligence, cloud computing and data analytics. It provides proprietary solutions and services for benchmarking and red teaming.

### How LLMs were used in application?

- Summarisation
- Retrieval augmented generation
- Multi-turn chatbot
- Classification or recommendation
- Data extraction from unstructured source

## What Risks Were Considered Relevant And Tested?

- Information accuracy and completeness
- Inappropriate or unsafe content
- Inappropriate data disclosure

## How Were The Risks Tested?

### Approach

- Focused on classification accuracy, information retrieval precision and security restrictions of the conversational question answering functionality
- Conducted using a representative dataset of documents with varying sensitivity levels
- Included adversarial prompting to assess resilience against attempts to extract sensitive information or generate inappropriate content

### Evaluators

- Human judgement
- LLM as a judge

## Challenges

- Defining appropriate risks - some were not obvious in the first place and needed further clarifications with the deployers
- Obtaining and securely handling a sufficiently diverse dataset that covered all possible personas

## Insights

01

Test datasets must be improved iteratively, with interim reviews by domain experts to obtain better testing outcomes

02

Curating diverse datasets requires careful specification early on and robust synthetic test dataset generation capability

03

Slim boundary between valuable inference and problematic hallucination: LLMs frequently inferred connections or drew conclusions from statements in prompts rather than strictly quoting source materials, creating complex eval challenges

GLOBAL AI ASSURANCE PILOT

Scan to read the full case study.

| Application Tested | Tester |
| --- | --- |
| Internal GenAI Chatbot | pwc |

UOB has deployed a Retrieval Augmented Generation (RAG) application into production for selected internal users. LLM used by internal users to answer operational and domain-specific queries using publicly available data.

### How LLMs were used in application?

- Summarisation
- Retrieval augmented generation
- Classification or recommendation
- Multi-turn chatbot
- Translation
- Data extraction from unstructured source

## What Risks Were Considered Relevant And Tested?

- Accuracy
- Robustness (under different conditions)
- Completeness

## How Were The Risks Tested?

### Approach

- Testing conducted based on mock-up questions closely mimicking queries from internal users when using the chatbot in production
- Questions contain a mix of binary, multiple choice and reasoning questions
- A sample set of question-answer pairs were established and validated by subject matter experts (SMEs) as the ground truth
- Questions were executed through pre-agreed prompts via a mock-up of the Internal GenAI Chatbot set up in a sandbox environment
- Generated outputs validated against ground truths validated by SMEs for testing purposes

### Evaluators

**Binary or multiple-choice:**

Rule-based (Metric: confusion matrix, standard deviation)

**Reasoning:**

Semantic similarity or LLM as judge (Metric: cosine similarity, hallucination and contradiction rate)

## Challenges & Insights

Testing in a highly regulated environment (e.g., banks) requires a highly representative mock-up to mimic the actual model in production so as to:

01 Manage the model's access to confidential/sensitive information

02 Prevent disruption to existing use in production

03 Ensure meaningful testing results

GLOBAL AI ASSURANCE PILOT

Scan to read the full case study.

# Investment Insights For Relationship Managers

**Application Tested**

An internationally active wealth management institution has a GenAI application to provide relationship managers with investment insights by retrieving and synthesising relevant internal information.

**Tester**

LatticeFlow AI

LatticeFlow AI empowers enterprises to deploy high-performing, trustworthy, and compliant AI systems, bridging the gap between AI governance frameworks and technical operations.

### How LLMs were used in application?

- Summarisation
- Multi-turn chatbot
- Retrieval augmented generation

## What Risks Were Considered Relevant And Tested?

- Accuracy, Soundness, and Relevance
- Lack of Transparency and Traceability
- Lack of alignment with Intended Use
- Dependency on 3rd parties

A subset of the risks listed above such as those related to accuracy, bias, hallucinations, and third-party dependencies of large general purpose AI models – were included in testing

## How Were The Risks Tested?

### Approach

- Primarily through benchmarking (using a mix of public benchmarks and tester's private ones), including non-adversarial assessment of categories like cybersecurity and harmful content

### Evaluators

- Human judgment
- Rule based algorithms
- Surface form metrics (text similarity at word/character level)
- LLM as a judge
- Non-LLM model-based evaluation

## Challenges

- Aligning benchmarks to the use case carrying its own specificity
- Automating evaluation process
- Obtaining necessary access to various components of the end-to-end LLM-powered application
- Repeatability of evaluation results

## Insights

01 Difficulty of testing an open-ended GenAI system

02 Importance of automation of testing process, as a way of dealing with a continuously improving GenAI tech stack

03 Importance of transparency and traceability of outputs to gain trust of end-users

04 Importance of translating of technical test metrics for non-technical stakeholders

Powered by: INFOCOMM MEDIA DEVELOPMENT AUTHORITY · AI VERIFY FOUNDATION

GLOBAL AI ASSURANCE PILOT

Scan to read the full case study.

# AI-refactored Code

| Application Tested | Tester |
|---|---|
| NCS, a pan APAC, Singapore-headquartered technology services firm, wanted to assess the performance of an in-house AI coding assistant, Sunshine.Coder, which leverages open-source code LLMs to refactor existing Java applications. | Parasoft specialises in automated software testing, focusing in DevSecOps solution for building security into critical programs like the US Department of Defence's Joint Federated Assurance Center. |

How LLMs were used in application?

Code generation (refactoring)

## What Risks Were Considered Relevant And Tested?

Key question: Is AI-Refactored Code Safer and Smarter?

Security, Compliance, and Standards Risks
AI-refactored code may:

- Inadvertently introduce security vulnerabilities, such as improper input validation, insecure data handling, or failure to enforce authentication and authorisation controls
- Not consistently adhere to internal corporate coding standards or external regulatory requirements (e.g., cybersecurity guidelines, data protection regulations)

Such deviations may cause technical debt, reduce auditability, and create hidden compliance risks

## How Were The Risks Tested?

### Approach

- Testing conducted on two different codebase profiles – Production and Proof-of-Concept (PoC)
- Issues mapped to compliance violations (e.g., deviations from security guidelines or coding standards)
- Testing conducted without executing the applications, focusing purely on static structural analysis

### Evaluators

Rule based logic:

Parasoft's Jtest automated testing suite for code quality/security

## Challenges

Variability in the quality of AI-refactored code across different applications and LLM configurations

## Insights

01

Sunshine.Coder improves code quality, but starting point determines quality of AI refactored code

- Production-grade code inherently follows best practices, reducing the complexity needed in prompts
- PoC code, lacking best practice adherence, requires explicit prompts with richer contextual instructions to address issues

02

Improved LLM architectures, comprehensive training data, and superior contextual understanding enable newer models to more effectively identify and resolve coding issues.

03

Key comparison metric: number of issues identified, classified by severity levels, in both human-written and AI-refactored versions

GLOBAL AI ASSURANCE PILOT

Scan to read the full case study.

# Wealth Relationship Manager Client Engagement Mail

## Application Tested

standard chartered

Standard Chartered Bank has developed a Generative AI ("GenAI") solution for generating personalised emails for use by relationship managers ("RMs") who have access to the Bank's Proprietary Digital Advisory Platform to reduce workload. The solution summarises information from multiple inputs, including client data and the bank's market outlooks, into a coherent and compliant draft message.

## Tester

pwc

PwC is a global leader in AI trust, offering comprehensive AI capabilities that help organisations innovate and deploy AI safely and responsibly. Across its global network, PwC provides AI strategy, solution development, and implementation services. PwC brings deep specialism in AI model testing and validation to enable organisations to have trust in AI driven results.

### How LLMs were used in application?

- Summarisation
- Retrieval augmented generation
- Data extraction from unstructured source
- Client email draft generation

## What Risks Were Considered Relevant And Tested?

- Internal compliance requirements (e.g., risk of breach of organisation's internal policy in terms of style and formatting)
- Trust/ Reputation (e.g., accuracy, robustness, completeness)

## How Were The Risks Tested?

### Approach

**Accuracy**
Hallucination rate, contradiction rate

**Robustness**
Measuring cosine similarity of the embedding of the output text, for multiple generations of a draft email for the same customer. Higher similarity scores indicate that draft emails generated are closer in content.

**Completeness/Engagement/Coherence**
Percentage of related requirements met in the draft

**Internal compliance**
Percentage of related requirements met in the draft (e.g., presence of promissory words or certain calculated quantities in email)

The deployer shared the tool's system prompts to help design the tests, and "seed" data to help create synthetic test data and edge cases.

### Evaluators

**Accuracy**
PwC's existing LLM as a judge pipeline was used to compare the generated email draft against the input data to the tool, to identify contradictions and hallucinations

**Completeness, Coherence, Engagement and Internal Compliance requirements:**
The internal requirements included in the system prompt of the tool were extracted and converted into LLM as a judge prompt and used to assess whether each of these requirements were met in each draft

The LLM as a judge assessments were compared to independent human Subject Matter Experts (SMEs) assessments on a subset of the drafts

## Challenges

**Need for custom tests design:**
To develop bespoke testing prompts for assessing if generated email outputs meet the SCB-specific requirements

**Challenging to get a representative sample for testing:**
Creating a large number of realistic and diverse synthetic samples requires considerable effort and access to wealth management expertise and domain knowledge

## Insights

01

**Scaling and flexibility of LLM as a Judge:**

LLM as a judge proved a powerful tool that was flexible enough to be used for different types of tests and allowed us to scale up our testing to many samples

02

**Lack of defined Assurance framework and standards:**

3rd party AI testing is an emerging discipline. Regulated industries will benefit from clearer frameworks and precedents (e.g., on access, data) to help streamline onboarding and execution

Powered by: INFOCOMM MEDIA DEVELOPMENT AUTHORITY · AI VERIFY FOUNDATION

GLOBAL AI ASSURANCE PILOT

Scan to read the full case study.

# FinMate – Anti-Money Laundering Assistant

**Application Tested**

TOOKITAKI

Tookitaki is a provider of AML and anti-fraud solutions. They built FinMate to reduce the time needed to investigate AML alerts by (a) auto-generating narratives, and (b) providing access to consolidated, relevant information.

**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 LLMs were used in application?

- Summarisation
- Retrieval augmented generation
- Orchestrator for an agentic flow
- Multi-turn chatbot

## What Risks Were Considered Relevant And Tested?

**Accuracy/Faithfulness**

Ensure that outputs accurately represent facts (e.g., alert reasons, customer details, risk indicators)

**Compliance/Preventing misuse**

Ensure that outputs adhere to financial regulations and internal compliance guidelines. Prevent misuse

## How Were The Risks Tested?

### Approach

**Accuracy**

- Extract key facts from GenAI response
- Compare against ground truth for presence and correctness of key entities (e.g., amounts, dates, names - post-masking) and critical instructions
- Key metrics: Precision, Recall

**Misuse prevention**

- Manually create a set of queries that violate policy
- Assess if system responds to these queries

### Evaluators

LLM-based extraction of key facts from response

**Automated comparison with ground truth facts**

- Rule based (value checks)
- Surface form metrics (term frequency-inverse document frequency + Cosine similarity between extracted facts)

## Challenges

Extracting data from the generated response for comparison to a ground truth (inconsistent output structure, necessitating custom text parsing scripts)

Generating test datasets that were both realistic and representative of edge cases that the system might need to be able to handle

## Insights

01

**System Design and Documentation**

Testing needs must be integral to the initial system design – e.g., traceability and comprehensive logging – and documentation

02

**Test Setup**

Clearly understanding system functionalities and limitations is vital for accurately scoping test cases. Establishing a robust testing environment with appropriate third-party access requires careful planning

03

**Data Generation**

Significant manual effort needed to create and validate synthetic test cases. Comprehensive understanding of the Entity-Relationship Diagram and application structure is essential

Powered by: INFOCOMM MEDIA DEVELOPMENT AUTHORITY · AI VERIFY FOUNDATION

GLOBAL AI ASSURANCE PILOT

Scan to read the full case study.

# Investment Research Assistant

| Application Tested | Tester |
|---|---|
| UNIQUE | QUANTPI |
| Unique is the vertical leader in agentic AI for the financial services industry. It has an Investment Research Assistant helping bank relationship managers to query stock universes, analyse fact sheets, and generate tailored investment recommendations. | QuantPi, a German CISPA spin-off, provides comprehensive, black-box AI testing technology that evaluates models of any kind for performance, bias, robustness, data quality and explainability, enabling safe, scalable model management. |

### How LLMs were used in application?

- Summarisation
- Multi-turn chatbot
- Data extraction from unstructured source
- Translation
- Classification or recommendation
- As the orchestrator for an "agentic" workflow
- Retrieval augmented generation

## What Risks Were Considered Relevant And Tested?

- Failure to meet regulatory requirements or internal company guidelines (e.g., recommending restricted products)
- Oversharing of client data
- Potential for advisor misuse
- Inconsistent advice quality across different customer segments
- Inadequate clarity in recommendations
- Poor financial outcomes

Pilot testing focused on accuracy/ hallucination/faithfulness

## How Were The Risks Tested?

### Approach

**Use Case 1 - Investment Research Assistant**

- Accuracy risks: Cosine Similarity between predicted responses and ground truth
- Hallucination risks: Faithfulness metric to check if responses were grounded in provided context
- Robustness/Reliability risks: Tested across Query difficulty levels, Domain bias and Typo tolerance

**Use Case 2 – Document Search**

Inaccurate and irrelevant results risks: Measured using three metrics - Word Overlap Rate, Mean Reciprocal Rank (MRR) and Lenient Retrieval Accuracy

### Evaluators

- Rule-based algorithms
- Thresholded and surface form metrics
- LLM as a judge

## Challenges

- Aligning automated evaluation metrics with nuanced financial summary requirements
- Securing and preparing a diverse dataset of anonymised financial data transcripts
- Ability to test subsystems to the agent/RAG that influence results: time limitations; resource caps; desire to protect intellectual property

## Insights

01 Difficulty of getting hold of adequate test data, including ground truth

02 Difficulty of collecting sensitive attributes to test for unfair bias

03 Difficulty of defining right metrics to measure and test across different use cases, prompts and underlying data sources

04 Difficulty of interpreting test scores without a comparable baseline

Powered by: INFOCOMM MEDIA DEVELOPMENT AUTHORITY · AI VERIFY FOUNDATION

GLOBAL AI ASSURANCE PILOT

Scan to read the full case study.

# Public Road Safety Chatbot

| Application Tested | Tester |
|---|---|
| **Multinational Insurer** | **VERIFY-AI** |
| The assured firm is a Multinational Insurer. The use case is a publicly accessible chatbot, designed to promote better driving habits through expert guidance. | Verify AI is a firm that specialises in quantitative risk evaluations of AI model applications. |

### How LLMs were used in application?

- Summarisation
- Retrieval augmented generation
- Classification and recommendation
- Multi-turn chatbot
- Data extraction from unstructured source

## What Risks Were Considered Relevant And Tested?

**Primary business risk**

- Reputational harm from an under-performing tool

**Associated technical risks**

- Hallucination
- Excessive guardrails/false refusal
- Response Variance/inconsistency
- Toxic and harmful outputs
- Prompt injection and jailbreaks
- Discrimination and bias

## How Were The Risks Tested?

### Approach

**Custom benchmarking for hallucinations**

- Baseline accuracy testing: Questions autogenerated from chatbot's designated knowledge base (using LLM)
- Stress testing for consistency and non-refusal: Additional questions from other domain-relevant inputs (e.g., external docs, general knowledge from LLMs)

**Red Teaming**

Automated adversarial tests to probe for jailbreaks/harmful content/other worst-case failures

### Evaluators

**Baseline accuracy testing**

- Semantic match of answer with what exists in the knowledge base (using LLM as a judge)
- Accuracy of citation provided (using ground truth)
- Refusal rate, with manual investigation of validity of refusal

**Stress testing**

Above, plus disaggregation of failure modes along the lines of identified risks (e.g., fairness, security and compliance)

**Red Teaming**

Attack success rate

## Challenges

- Automatic generation of adversarial prompts: Substantial design effort and compute needed
- Verifying insights from LLMs as a judge: Need for human oversight to ensure that these LLM judges perform well. Sequential model checking employed to build formal proofs of statistical correctness of evaluation

## Insights

01

Clear, testable definitions needed for each technical risk (particularly in "no box" situations where only access to the final outputs of the chatbot are available)

02

Difficult to verify and understand the root-causes of hallucinations without visibility into the retrieval process

03

LLM as a judge viable, but unreliable without additional safeguards; statistical model checking can help

Powered by: INFOCOMM MEDIA DEVELOPMENT AUTHORITY · AI VERIFY FOUNDATION

GLOBAL AI ASSURANCE PILOT

Scan to read the full case study.

# Multilingual Internal Knowledge Bot

### How LLMs were used in application?

- Translation
- Retrieval augmented generation
- Classification and Recommendation
- Multi-turn chatbot (with additional Chinese tokens to better support prompts and outputs in Chinese)

## Application Tested

### High Tech Manufacturing Firm

The deployer is a global leader in high-tech manufacturing of intermediate goods requiring highly complex engineering. The chatbot was deployed to provide a mechanism to access related (internal) product specifications, conduct troubleshooting, and answer frequently asked questions.

## Tester

### VULCAN

Vulcan is a product of AIFT, which offers GenAI-specific security solutions focusing on GenAI vulnerability assessment and red-teaming.

## What Risks Were Considered Relevant And Tested?

**Robustness**

Inconsistencies in chatbot performance and reliability across multiple languages

Generation of harmful or inappropriate content

**Data leakage**

Potential inadvertent exposure of sensitive product specifications, proprietary intellectual property, or other confidential business information

**Security**

Meta-prompt leakage

## How Were The Risks Tested?

### Approach

- Automated red teaming for security, data leakage and harmful/inappropriate content testing
- Customised test cases for information accuracy and consistency

### Evaluators

- Automated evaluation using LLM as a judge for red-teaming exercises
- Human-in-the-loop evaluation for customised test cases

## Challenges

- Insufficient training data for customised test case generation
- Not anticipating the impact of high-volume testing on API protection mechanisms in place to guard against DDoS attacks

## Insights

01

Automated red teaming traffic triggered the deployer's denial of service defence;

Consider pacing traffic or whitelisting test IPs to keep assessment from being mistaken for attackers

02

RAG improved relevance and accuracy but widened the leakage surface

Consider redacting or masking highly confidential information before indexing and enforce document-level access control list checks

03

Guardrails that pass in English can fail in Chinese Further alignment of tokens, policies and RLHF datasets across languages is required

Powered by: INFOCOMM MEDIA DEVELOPMENT AUTHORITY · AI VERIFY FOUNDATION