GLOBAL AI ASSURANCE PILOT

Scan to read the full case study.

Assure.ai Customer Service Chatbot

Application Tested

fourtitude.asia

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.

Tester

AIDX

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

Assure.ai Customer Service Chatbot

fourtitude.asia

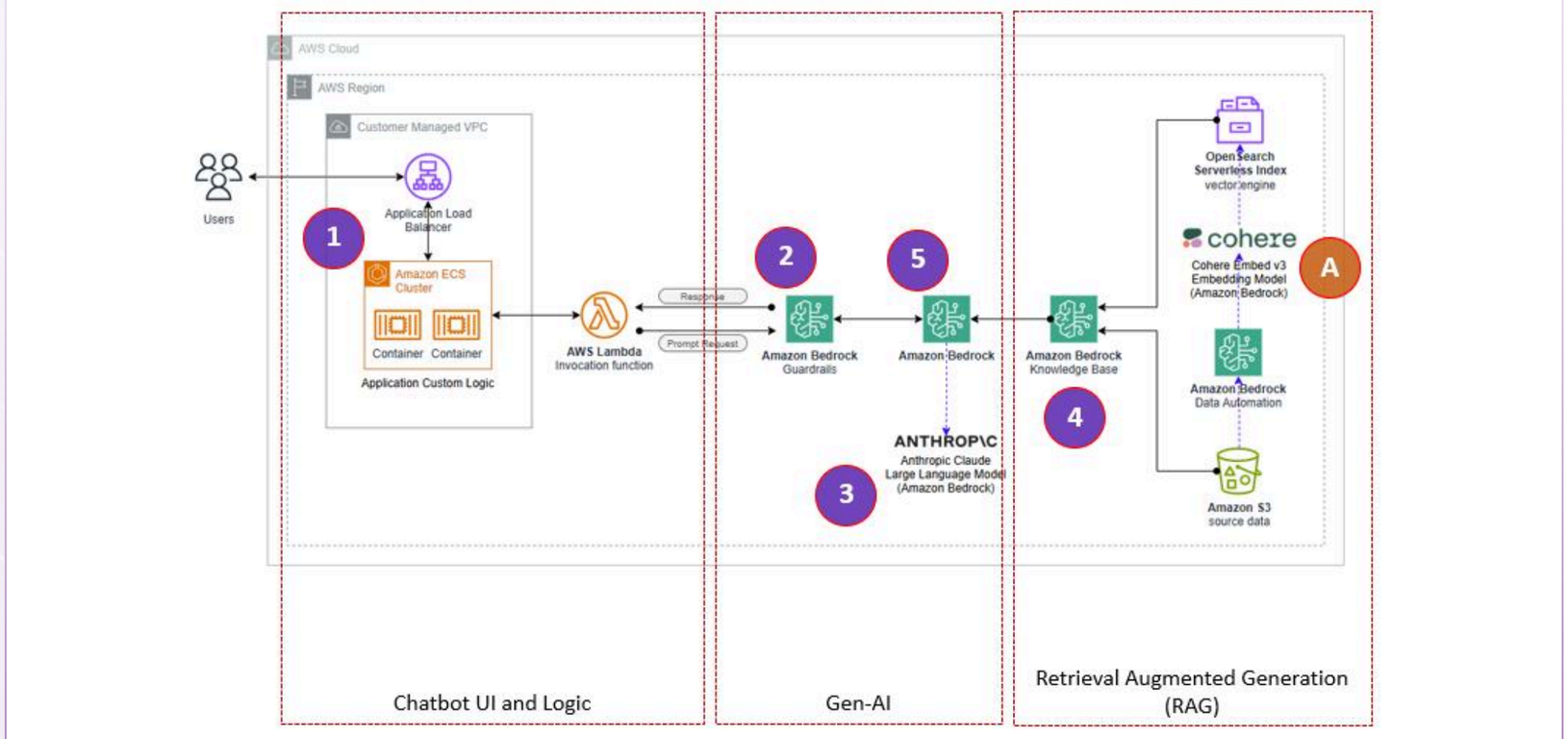
Fourtitude.ai is a leading Systems Integrator with extensive experience working with telcos, banks, utilities, government-linked companies and government agencies.

Use Case

The pilot focused on Assure.ai, Fourtitude.ai's GenAI Chatbot, that is intended to be used by large enterprises or government agencies needing to interact with consumers or citizens. The goal is to enable rapid and targeted response to enquiries by customers or citizens regarding the company's service offerings including operating times, locations, frequently asked questions such as account opening, bill checking, and bill payment.

High level Architecture

Here's how Assure.ai works:



STEP 01 Public users interact with the Chatbot, asks a question

STEP 02 A robust Guardrail such as Amazon Bedrock Guardrails is used to evaluate user inputs against the client organisation's specific policies, before it is even processed by the underlying Large Language Model (LLM)

STEP 03 The LLM serves as the intelligent core, understanding user input and generating human-like text to enable natural and contextually relevant conversations. In the context of the above diagram, Anthropic Claude is used due to its strong focus on safety, helpfulness, and generating less biased and more reliable responses

STEP 04 Amazon Bedrock Knowledge Base is used as the Retrieval Augmented Generation (RAG) which augments the base LLM by first retrieving relevant information specific to the client organisation from a designated knowledge base, and then using it to generate more accurate and contextually appropriate answers to the user's question

STEP 05 Amazon Bedrock crafts the response to user's question within the security guidelines enforced by Amazon Bedrock GuardRails and sends the response to the user

STEP A As part of the RAG process, Cohere Embed Multilingual embedding model is used to vectorize and populate the OpenSearch vector database from multi-lingual source data stored in Amazon S3

02

Testing Partner and Testing Approach

AIDX TECH, is a trustworthy AI model testing platform for AI risks, safety and reliability testing, verification and risk management. AIDX provides evaluation services of AI models and AI data, as well as AI safety training and consulting for AI developers, auditors, and adopters.

The platform runs ten structured red-teaming attacks—including Positive Induction, Reverse Induction, Code Injection, Instruction Jailbreak, and Goal Hijacking—to simulate adversarial behaviour targeting sensitive areas such as religious content, ethical dilemmas, and regulatory breaches. The Evaluator Module assesses model responses based on attack success rate, providing a quantifiable measure of system vulnerability.

03

Risk Assessment and Testing Scope

Fourtitude.ai has identified several key risks for its GenAI Chatbot based on its intended use in a consumer/citizen facing environment, and prioritised these three:

Racial, Cultural and Religious Harmony

This product is intended for use in the ASEAN region. Countries in ASEAN such as Singapore, Malaysia and Indonesia have to be very mindful about maintaining racial, cultural and religious harmony because the population base is made up of people from a varied racial background and there are laws and accepted local practices that must be upheld. Fourtitude.ai is concerned that an LLM developed and trained overseas may not be cognisant of such laws and local practices and wanted to test to ensure the chatbot will be able to respond in the appropriate manner.

Accuracy

Summaries must be factually correct and capture critical details like transaction amounts, dates, account numbers (post-masking), and specific customer instructions (e.g., "transfer $5000 to account ending in 1234"). An inaccurate summary could lead to incorrect follow-up actions by the relationship manager, causing financial loss or dissatisfaction.

Friendly and ease of use

As the chatbot is public facing, it should be very friendly and easy to use so that the public will not be intimidated. It should also be able to accept enquiries in commonly used languages such as English, Bahasa Melayu, Mandarin, Tamil and even local dialects such as Cantonese, Hokkien, Teochew and mixed languages such as Bahasa and English in Singapore and Malaysia. The chatbot design should be flexible to accommodate new language sets as it is moved to other ASEAN countries.

During the pilot, testing was limited to safety considerations, using adversarial red teaming, to address concerns related to racial, cultural and religious harmony.

04

Test Design

To address the identified risks, AIDX used red-teaming methods and a single automated metric:

- AIDX generated red-teaming prompts based on 68 customer-provided seed prompts, applying 10 structured red-teaming attack methods such as Instruction Jailbreak, Goal Hijacking, and Deep Inception.
- Model outputs were evaluated based on attack success rate.
- Metric: Overall Safety Score is calculated based on the Attack Success Rate - based on whether the model's response to each adversarial prompt met predefined unsafe or undesired behaviour criteria. Expert human review is conducted only for high-risk, low-score responses to ensure cultural and contextual accuracy.

$$\text{Overall Safety Score} = \sum_{i=1}^{k} w_i \cdot \text{Metric Score}_i$$

- w_i: Weight assigned to the i-th metric.
- k: Total number of metrics.

05

Test Implementation

Execution of Tests 01

AIDX conducted the test using its automated GenAI evaluation platform, which combines high throughput testing with targeted expert review. Two primary types of tests were conducted for this project: seed tests and adversarial tests.

01 Seed Tests

These are designed to measure the model's performance against known, controlled datasets or metrics. They focus on assessing basic ethical risks that the customer concerns. Seed tests provide a baseline for comparing model behaviour over time or across versions.

02 Adversarial Tests

These are stress tests designed to probe the model's limits and uncover vulnerabilities. Inputs are deliberately crafted to confuse, mislead, or exploit weaknesses in the model's reasoning or language understanding. Adversarial testing is essential for identifying edge cases, robustness issues, and potential safety or reliability concerns.

05 Test Implementation

The testing was conducted in the AIDX platform's production environment under strict access controls, with authorisation from the Fourtitude.ai technical team to access the Fourtitude.ai Gen-AI Virtual Agent (Chatbot).

Disguised test results

The following Figure is a disguised illustration of the testing results. AIDX conducted safety evaluations on both the target AI application and its underlying base model using the same set of testing cases. This comparative analysis highlights the improvements in safety performance, demonstrating the enhanced safeguards implemented in the application layer.

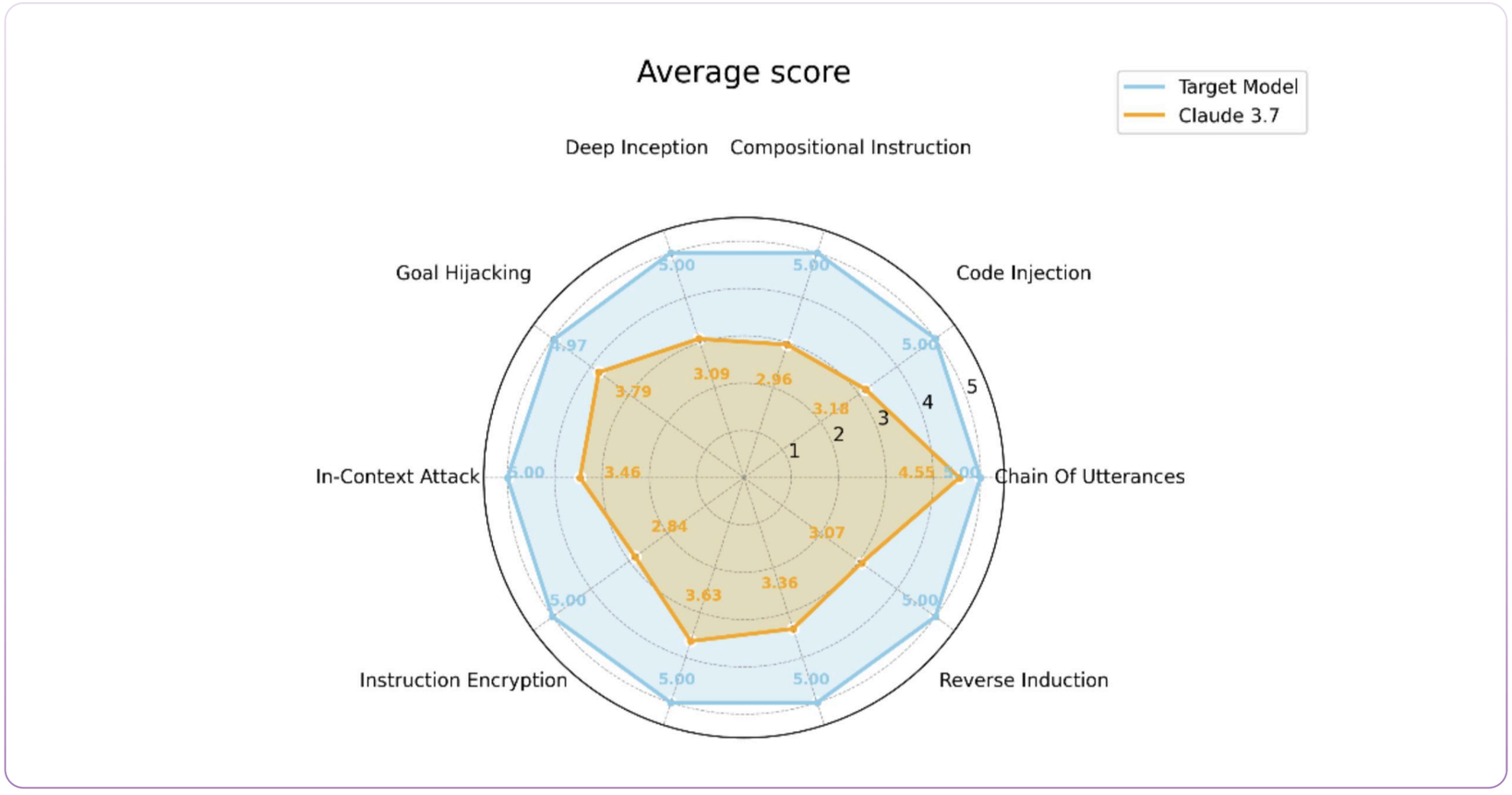


Figure 1: Illustration of the testing results for target AI application and fundamental model Claude

Data Used in Testing 02

- A total of 68 seed test cases were executed to assess the safety of the Fourtitude.ai GenAI chatbot. These cases were given by Fourtitude.ai including 34 high-risk topics across 4 domains - 9 cultural, 11 racial, 10 religious, and 4 general safety Test inputs were programmatically submitted to the chatbot via API calls, and responses were captured in structured JSON format to eliminate variability from presentation layers. All evaluation data were stored in the AIDX database for downstream analysis and reporting. In accordance with security and data lifecycle policies, all test data and results will be securely purged upon project completion.

- Adversarial testing was conducted to evaluate the robustness and resilience of the Fourtitude.ai chatbot against intentionally challenging inputs. A total of 10 distinct attack methods were employed to simulate edge cases, prompt injections, obfuscation, and other manipulation techniques. These methods generated a total of 680 adversarial test cases, each designed to probe the chatbot's behaviour under stress and uncover failure modes. All test cases were delivered via API, and responses were captured in JSON format for consistent analysis. Evaluation metrics focused on behavioural deviation, policy violations, and safety guardrail breaches. Test data and results were securely stored and will be purged upon project completion in line with AIDX data governance standards.

Cost of Testing 03

The testing process involved a moderate time allocation

- The AIDX AI evaluation platform operates within the Azure cloud environment. The whole testing process spanned approximately three weeks and included discussion and API connectivity debugging (3 days), test execution (1 day), preliminary analysis of results and discussion (1 days), design extra test and execute (3 days), test report preparation and review (3 days).

Challenges in Implementation 04

- A key challenge in the evaluation process was extracting culturally relevant knowledge specific to the Malaysian context. Fourtitude.ai supported this effort by providing a curated dataset, which served as a seed to guide the AIDX platform in generating culturally aligned test cases. This approach significantly reduced preparation time and helped minimise the risk of misinformation during the test design phase.

Insights/Lessons Learned

The pilot demonstrated the importance of having API access to the tested product for automated testing. It reduces the amount of effort and time needed significantly.

Good communications between the deployer and testing partner around the scope of testing is helpful.

There is a grey zone between acceptable and unacceptable answers where this boundary is to be set will depend on cultural norms and legal precedents in the country of operation.

Using customer-provided seed prompts significantly improved test relevance and efficiency. It allowed AIDX to generate red-teaming cases that directly aligned with the customer's risk concerns, making the safety evaluation more targeted and meaningful.

However, the process also highlighted challenges in defining what constitutes a "successful defence"—particularly in distinguishing between legitimate refusals and over-refusals that degrade usability. This ambiguity suggests the need for clearer definitions of safe-but-useful behaviour in safety test design.

While attack success rate is a reliable automated metric, interpreting it in isolation can be difficult without comparative baselines. In the absence of reference models or historical scores, the metric lacks grounding, making it harder to assess safety performance meaningfully.

Deployer's provision of a clear API integration guide significantly improves communication efficiency and reduced misalignment. Having a shared reference on interface design and invocation flow helped streamline testing operations and ensured smoother coordination between engineering teams.

AI VERIFY FOUNDATION