

Application Tested

Tester

Internal GenAI Chatbot




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?

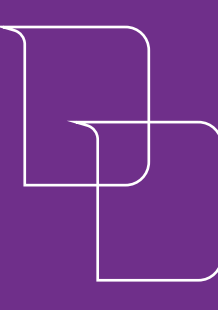
Approach

-  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

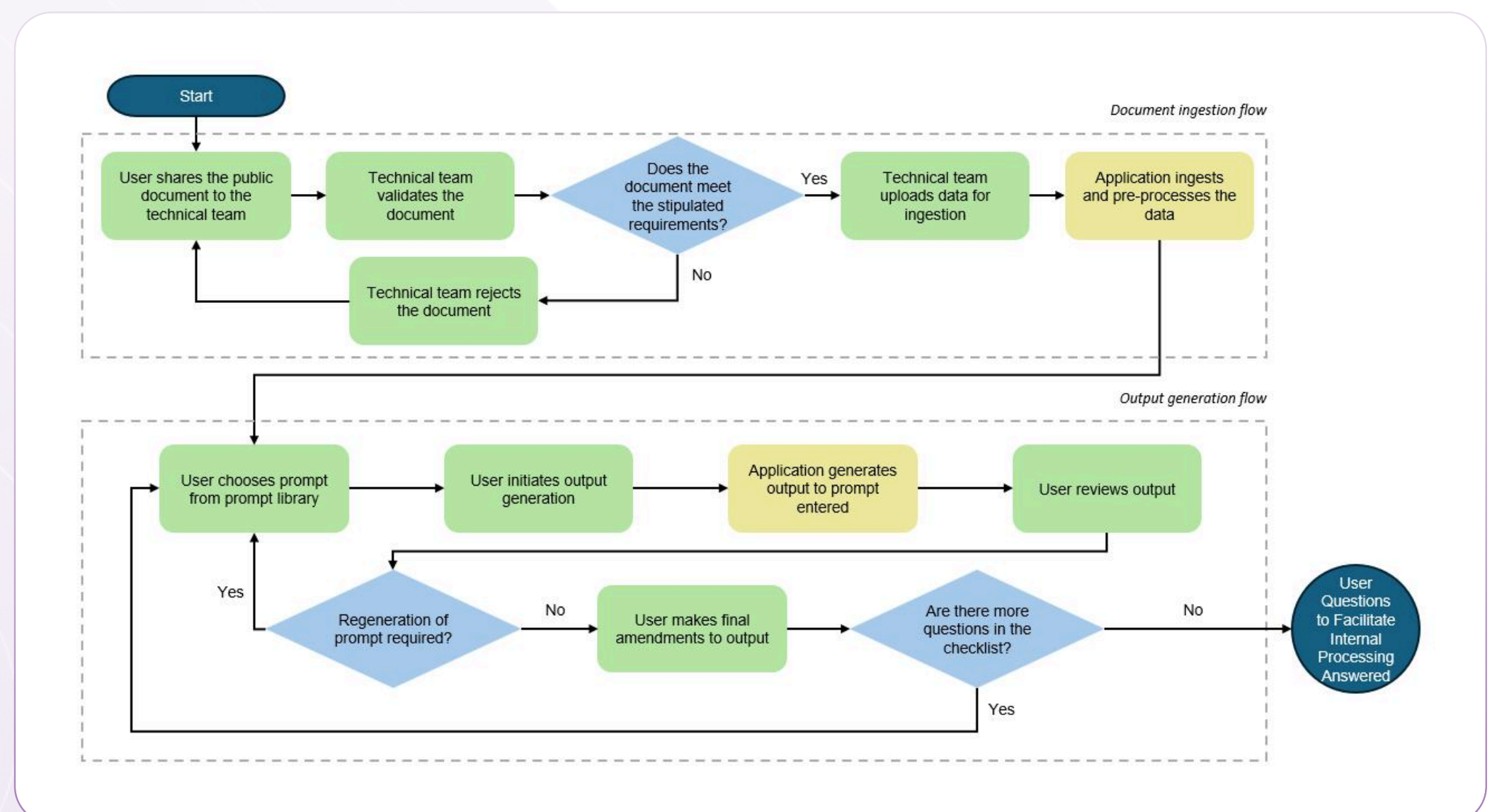
Internal GenAI Chatbot

UOB has deployed an Internal GenAI Chatbot Retrieval Augmented Generation (RAG) application into production.

The Internal GenAI Chatbot application utilises a Large Language Model (LLM) to retrieve, summarise and analyse information facilitating internal business processing by answering operational and domain-specific queries using publicly available data.

Outputs generated from the application are reviewed and amended by users (human-in-the-loop process) before use for internal processing. The application is only open to selected users in the bank.

The below sections illustrate the typical business process flow involving the Internal GenAI Chatbot and the high-level architecture of the RAG application.



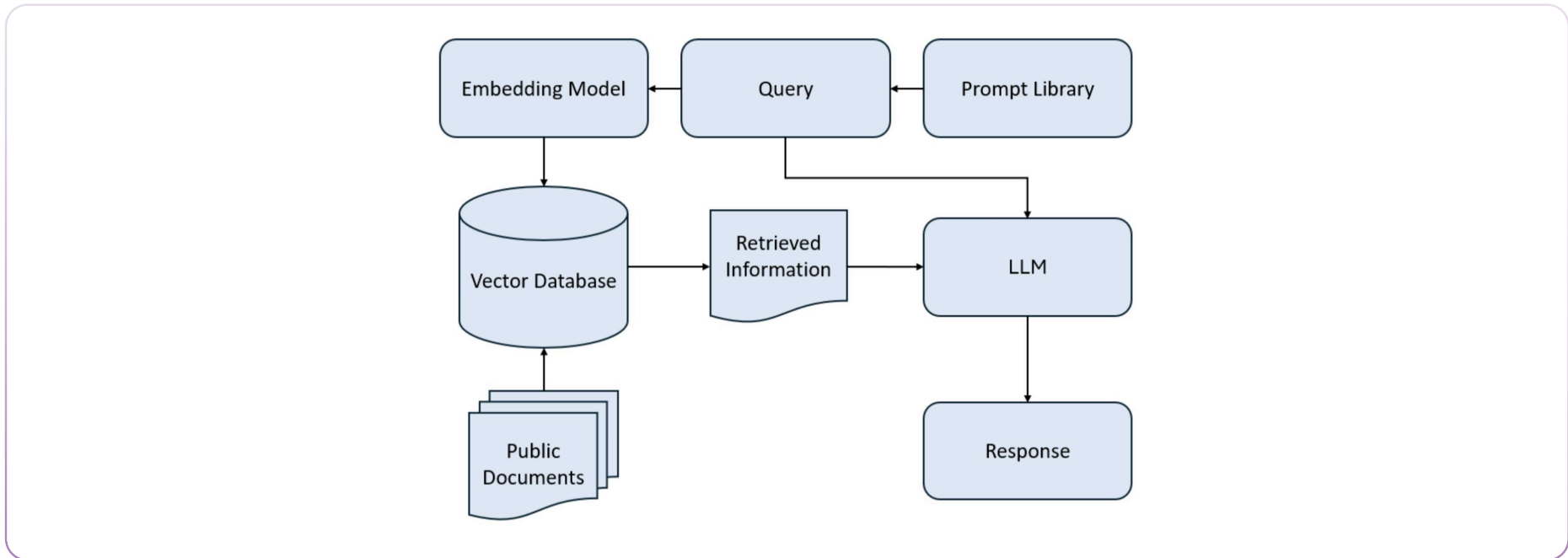
The user journey typically consists of the following steps:

- 01 The user submits the relevant company public disclosures (e.g., public sustainability reports, financial statements) to a technical team as the data source(s) the LLM refers to for answering user questions to facilitate internal processing
- 02 The technical team then validates and uploads the documents for ingestion
- 03 As part of pre-processing of ingested data, machine learning models are used for the purposes of guardrails and reranking
- 04 The application will determine if the document is in the appropriate format or contains sensitive information, upon which the document will be rejected
- 05 Upon successful ingestion of the document, users can select and input the relevant prompt from the prompt library to generate outputs as required
- 06 Thereafter, users will review the generated outputs and make the necessary amendments to facilitate internal processing

High-level architecture

The following diagram illustrates the high-level architecture of the Internal GenAI Chatbot being tested:

Data sources consist of publicly available documents. The underlying foundation model in the Internal GenAI Chatbot is Meta Llama 3.1.



Testing Partner and Testing Approach



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. Their extensive expertise and hands-on experience enable seamless integration of validation frameworks to deliver Gen AI model testing and monitoring with exceptional proficiency and insight.



Risk assessment approach

In risk assessment, the Tester will typically assess risks across 6 industry-standard dimensions: model risks, data risks, ethical risks, tech and security risks, deployment risks and legal risks. For the purposes of the pilot, the risk assessment focused on model risks, as well as other testing areas suggested for the AI Verify Foundation Global AI Assurance Pilot.



Technical testing approach

The Tester utilised a combination of proprietary technology ('LLM as a judge'), computational NLP techniques and Subject Matter Expert (SME) evaluation to perform the technical testing for this use case.

LLM as a Judge was used as an evaluator in this use case by leveraging on another LLM to assess the outputs generated by the Internal GenAI Chatbot, through custom instructions given to the LLM.

- Tests focused on accuracy, robustness and completeness through the following main approaches:
- ✓ Validation of final outputs against generated ground truth datasets for common questions with categorical outputs
 - ✓ Validation of final outputs against the public documents (from which the ground truth datasets were generated from), and ground truth reasoning provided by SMEs, to detect hallucinations and contradictions of the chatbot for the questions that require reasoning
 - ✓ Repeated testing using the same prompts to check for consistency in generated outputs

The risk assessment focused on model risks and the relevant testing areas suggested for the AI Verify Global AI Assurance Pilot, which in turn informed the testing scope for the Internal GenAI Chatbot.

As the Internal GenAI Chatbot is an internal-facing application, utilising publicly available data, and is primarily intended to be a productivity tool, the **Trust** testing area was prioritised to assess the accuracy, robustness and completeness of the responses generated by the tool, which were risks assessed to have a **High** impact.

Risks tested	Rationale for assessed impact level as 'High'
Accuracy of generated output(s)	Inaccurate outputs may impact internal processing subsequent downstream usage of the generated outputs.
Robustness of generated output(s)	Outputs which lack robustness may indicate the Internal GenAI Chatbot's inability to produce conclusive outputs expected by users under different conditions, impacting the utility of the tool and eroding user confidence.
Completeness of generated output(s)	Incomplete outputs may lead to material information being omitted in the generated outputs, affecting internal processing and downstream outputs.

Additionally, the following risks were considered but were not tested due to the assessed impact of the risks to be **Low** to **Medium**.

Risks considered	Rationale for assessed impact level as 'Low' to 'Medium'
Transparency of generated output(s)	Considering that the Internal GenAI Chatbot generates justification for any classification-related questions and that users have ready access to the original documents used as data sources, users can fact-check generated outputs independently. This may result in more tedious checking for users but may not necessarily translate into a 'High' risk level.
Bias present in generated output(s)	Considering that the usage of the Internal GenAI Chatbot still involves a 'human-in-the-loop' as a control to review the outputs before they are used in any downstream decisions/actions, the risk of bias may be mitigated through human review currently.

Scope of Testing

Due to confidentiality reasons, the testing would be conducted based on a set of common questions closely mimicking those used in the Internal GenAI Chatbot. The questions will contain a mix of binary, multiple choice and reasoning questions. All questions will be executed using the Internal GenAI Chatbot set up in a sandbox environment across 10 publicly available documents. Generated outputs will then be validated against ground truths, which have been validated by SMEs for testing purposes.

The following tests have been designed for the Internal GenAI Chatbot:

Risks	Expected Question Type	Evaluator Type	Metric(s)	High-Level Approach
Accuracy	Binary/ Multiple-Choice	Rule-based algorithm	Confusion matrix	Comparison with ground truths
	Reasoning	LLM as a judge, human judgement	Hallucination rate, contradiction rate	Checking for hallucinations or contradictions – with the original public documents as reference (LLM as a judge), or with ground truth reasoning from SMEs
Robustness	Binary/ Multiple-Choice	Rule-based algorithm	Stability metric	Comparison between different responses, when the same or similar questions are run multiple times
	Reasoning	Semantic metrics	Cosine similarity of responses across multiple instances of response generation (-1 to 1)	Comparison between different responses, when the same or similar questions are run multiple times
Completeness	Reasoning	LLM as a judge	Percentage of component of questions that the response attempts to answer	Checking for completeness in terms of the response addressing all the different sub-questions in the synthetic prompts

Table 4.1 Metrics used for Assessment of Accuracy, Completeness and Robustness



The test execution process comprised several key activities:

Generation of ground truths:

As ground truths validated by domain SMEs in a production setting were not available due to confidentiality reasons, a set of ground truths based on 10 arbitrarily chosen companies were manually generated by the Tester, with review from the Deployer.

Generation of outputs from Internal GenAI Chatbot:

The generated outputs were manually obtained through the Internal GenAI Chatbot in the Deployer environment and comprised of the following key steps:

- Upload relevant public document and ensure it has been ingested
- Select and copy prompt to be tested into the Internal GenAI Chatbot application's input field and generate output
- Record the generated output to be exported and run through the Tester's testing pipeline

Calculation of metrics

(excluding tests involving LLM as a Judge). The metrics used to test the Internal GenAI Chatbot are calculated as follows:

Accuracy for binary/multiple-choice questions:

The number of correct and wrong answers for binary/multiple-choice question types will be recorded.

The confusion matrix is then generated across i) all questions and ii) for each question to yield insights into the performance of the Internal GenAI Chatbot in terms of accuracy, precision, recall and F1 score.

Accuracy for reasoning questions:

- ✓ The accuracy of responses for reasoning questions will be measured via the detection of hallucinations and contradictions with the contents of the public document or the ground truth reasoning, if available.
- ✓ The approach for detecting hallucinations and contradictions begins with breaking down the output reasoning response into a set of clauses, using an LLM, and proceeds with checking whether each of the clauses constitutes a hallucination or contradiction by trying to retrieve relevant chunks from the original report. The LLM as a Judge will determine if chunks contradict the clause and categorise it as a contradiction, and if no relevant chunks to the clause can be retrieved, then the clause will be classified as a hallucination.

Robustness for binary/multiple-choice questions:

A set of responses will be generated for binary/multiple-choice questions. The same prompts will be run multiple times to generate several iterations of the outputs. The different answers will then be recorded and used to compute a Stability Metric, which measures the proportion of responses that form the majority prediction, or that remain unchanged when the question is run multiple times. An alternative version of this metric is to measure the proportion of responses which remain correct.

Robustness for reasoning:

The same prompts for reasoning questions will be run multiple times to generate several iterations of the outputs. Each of these responses will be embedded using a Sentence Transformer model and the average and minimum similarity across the embeddings will be computed and reported.

Completeness for reasoning questions:

Generated responses to each of the reasoning questions are given to an LLM as a Judge, which is tasked with identifying whether there are any sub-questions or requirements in the question that have not been addressed in the answer. The LLM as a Judge returns a list of missing information from the question. If the answer addresses all sub-questions in the question, then the list will be empty.

Resourcing and cost involved with testing

Overall, investments to facilitate the testing were largely on the resourcing required. The cost for LLM usage were immaterial. Below is a summary of resources invested for this testing:

- Deployer spent approximately 80 man-hours across 3 personnel
- Tester spent approximately 520 man-hours across 10 personnel

Challenges encountered in testing implementation

The following challenges were encountered while testing the Internal GenAI Chatbot:

- 01 Access to system prompts:

Due to confidentiality reasons, verbatim prompts were not used in the testing, which meant that the performance measured as part of the pilot can only be considered as a proxy of the performance of the deployed tool.
- 02 Lack of 'ready-made' ground truths for validation:

Due to confidentiality reasons, the original questions and corresponding ground truths validated by domain SMEs in a production setting were not available to the Tester in the pilot.
- 03 Usage of manual testing and sandbox environment:

Due to concerns around the performance of other use cases concurrently sharing the infrastructure supporting the Internal GenAI Chatbot, outputs could only be generated manually and in a sandbox environment, which meant that the performance of the Internal GenAI Chatbot in a production environment could not be validated during this pilot.

06 Insights/Lessons Learned

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:

- Manage the model's access to confidential/sensitive information
- Prevent disruption to existing use in production
- Ensure meaningful testing results