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

Relationship Manager Client Engagement Mail

Standard Chartered Bank is a global bank that connects corporate, institutional and affluent clients to a network that offers unique access to sustainable growth opportunities across Asia, Africa and the Middle East.

Use Case

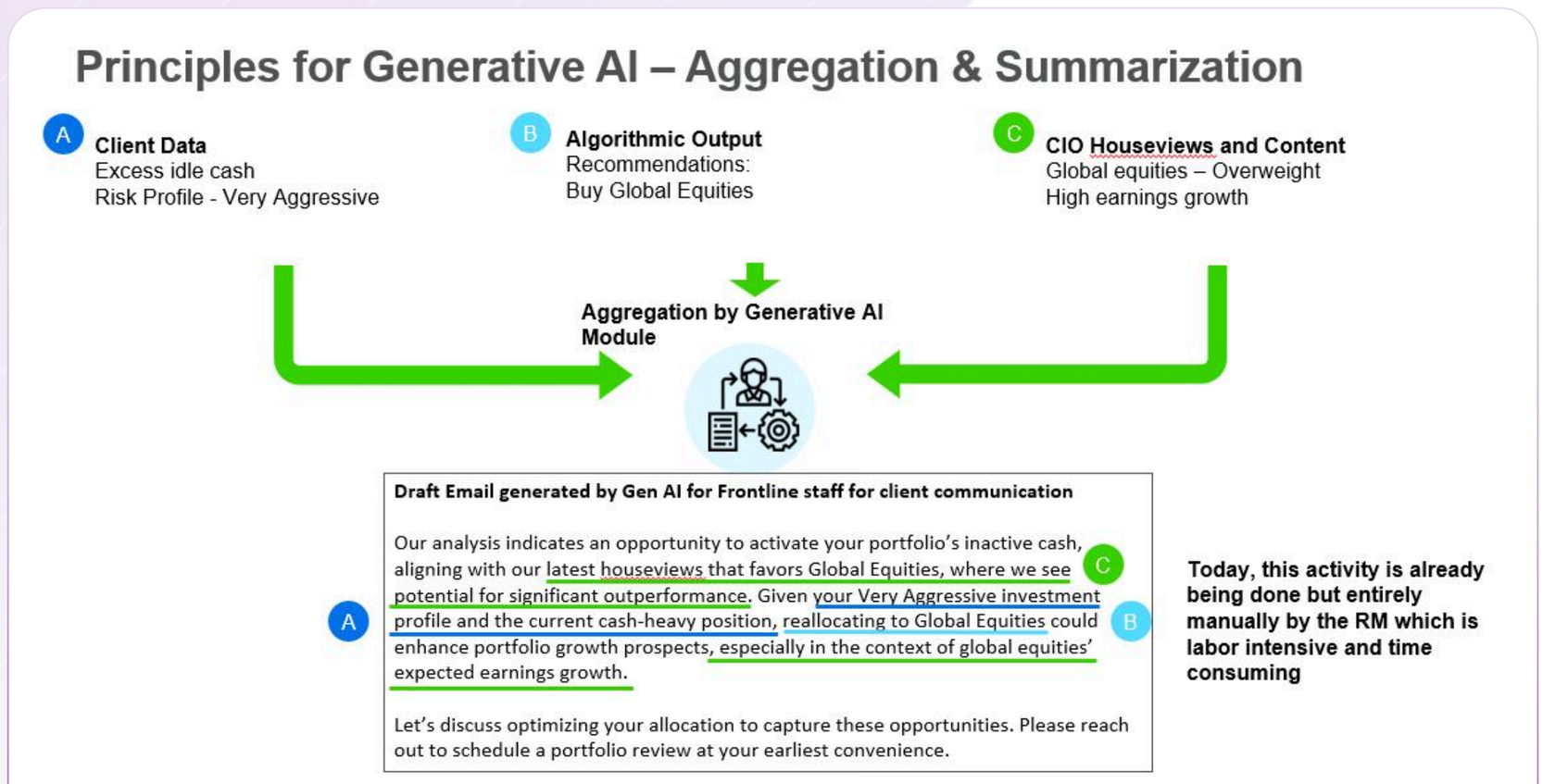
The pilot focused on a Generative AI (GenAI) solution designed to generate personalised external client emails for use by relationship manager (RM) who have access to the Bank's Proprietary Digital Advisory Platform. This solution is aiming to reduce RM workload and enable timely customer outreach. The solution summarises information from multiple inputs – such as client data, investment preferences, algorithmic recommendations, and market views of the Chief Investment Office – into a single, coherent, and compliant draft message. The solution uses a combination of PII masking, PII detector, and a Large Language Model (LLM) for aggregation and summarisation. omponents include:

High-level Architecture

The solution takes input from multiple sources in a JSON format, aggregates the information, and generates a nudge draft email for review. The architecture of the integration is based on a modular design to ensure scalability and flexibility. The core components include:

- **Generative AI Module:** This is the central engine which handles natural language understanding and generation
- **API Gateway:** Acts an intermediary between the Generative AI Module, and internal platform to manage data flow, and ensure secure communication
- **Data Integration Layer:** Facilitates the integration of various data sources, including client information, bank's view of market commentary and client's investment preferences and risk profile
- **User Interface:** Provides the frontend through which users interact with the Data Sources

Figure 1: Diagram of myWealth tool to be tested



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.

Testing Approach & Tools

In this context, "testing" refers to a suite of quantitative and qualitative assessment techniques used to evaluate the performance, reliability, and risk of LLM-powered applications. Unlike traditional software or statistical models, LLMs are non-deterministic, language-based, and highly context-sensitive – making testing a uniquely complex challenge. The inherent variability and complexity of LLM language outputs necessitate innovative testing approaches to ensure reliability and trustworthiness.

PwC's approach addresses this by focusing on the risks and limitations inherent in GenAI use cases: such as hallucinations, biases, factual consistency, and context misinterpretation. A three-pronged testing methodology is used, including:

- **Computational NLP techniques**

  (e.g., BLEU, ROUGE, BERTScore) to benchmark linguistic similarity and output consistency.

- **Human subject matter expert (SME) reviews,**

  drawing on expert judgment to assess factual accuracy, appropriateness, and clarity of outputs; and

- **LLM as a Judge,**

  an emerging approach where one LLM is used to evaluate the outputs of another, offering scalable, prompt-based evaluations that reflect the nuances of language use.

Tests focused on accuracy, robustness, completeness, coherence, engagement and internal compliance requirements through the following main approaches:

- **Output validation using synthetic user profiles:**

  Final outputs are validated against generated ground truth datasets. These are constructed using synthetic user profiles, algorithmic recommendations, and market views of the Chief Investment Office extracts as inputs to generate realistic draft emails for review.

- **Non-adversarial edge-case testing:**

  The model is further tested using synthetically generated "edge-case" data to identify limitations and assess the system's handling of less typical but plausible scenarios.

03

Risk Assessment and Testing Scope

The table below provides a mapping between the AI Verify Pilot's defined risk areas and the organisation's existing risk taxonomy. For the purposes of this pilot, two areas were identified as most relevant to the selected use case: **(Non-AI) Internal Compliance Requirements** and **Trust/Reputation**. These form the primary scope of testing for this exercise and are highlighted accordingly in the mapping below.

Category /Subject Area	Definition /Description	Example Issues	Risk Type (Risk Management Framework)
Safety and health	Potential for the AI application to cause physical or mental harm to individuals	Physical harm, direct negative physical or mental health outcomes	People Risk / Physical Safety and Security
(Non-AI) Industry-specific regulatory requirements	Potential breaches of regulatory requirements specific to the industry that do not involve AI	Displaying adult content advertisements next to children's content, collusion between pricing agents, mis-selling of financial products	Compliance / Regulatory
(Non-AI) Internal compliance requirements	Breach of the organisation's internal policies rather than external regulations	Negative commentary on named competitors, discriminatory language in AI-generated content. In addition, the generated content fails to meet the formatting and style standards set by the organisation	Data Risk / Responsible AI / People Risk / Culture
Unfair treatment of employees/ customers/users	Discrimination or adverse outcomes for specific groups or individuals due to inaccuracies or biases	Sexist language in AI content, chatbot failing to meet needs of under-represented groups, unfair classification of a student's essay as AI-generated	Data Risk / Responsible AI / Conduct / D&I
Transparency and recourse	Lack of understanding or ability to explain AI outputs and providing avenues for redress	Inability to explain AI decisions, lack of user understanding of AI outputs, difficulty in seeking redress.	Privacy / Responsible AI
Data disclosure	Unintentional, inappropriate disclosure of personal or confidential data by the AI application	Leakage of sensitive information	Information & Cyber Security
Malicious use	Potential for the AI application to be misused for harmful activities	Spreading misinformation, encouraging hatred or violence, committing fraud	Conduct / Ethics
Trust/Reputation	Risk of damaging the organisation's reputation or customer trust	Producing embarrassing content	Reputational / Stakeholder Perception
Financial loss	Inaccurate outputs potentially leading to financial losses	Failed automation, incorrect decisions, under-pricing products/services	Financial

Table 3.1: AI Verify Pilot risks mapping to the myWealth tool.

Testing Scope

Based on the two priority areas identified, the following model risks have been selected as the focus of testing for this Pilot.

Risks tested	Rationale for assessed impact level as 'High'
Accuracy of generated output(s)	Inaccurate outputs may lead to incorrect information being included in client communications. Misalignment between consolidated data (e.g., market trends, client data) and the generated email could result in factual errors.
Robustness of generated output(s)	Outputs which lack robustness may indicate the Email Drafting Tool'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 omit critical information, undermining the usefulness of the email. Failing to follow prompt instructions could result in incoherent or misleading client communication.
Coherence of generated output(s)	Coherent messaging is essential to avoid misunderstandings. Poorly structured or unclear emails may erode client trust and require additional follow-up clarification.
Engagement of generated output(s)	Emails should be engaging to ensure they capture the recipient's attention and foster positive client perception. Low engagement may reduce response rates and client satisfaction.
Internal compliance of generated output(s)	Outputs must align with internal policy. The inclusion of promissory language, restricted asset references, or inappropriate quantitative statements may violate compliance rules and erode business confidence.

Table 3.2: Scope of testing for the Pilot.

The Deployer has provided the Tester with the following:

- Excel files containing the data inputs detailing mock customers' financial situations, and corresponding outputs of the Deployer's algorithmic model for each client
- Task background and format containing the instructions as to how the email summary should be generated, e.g., what information to include in the email, how to structure it, what not to include
- A library of email templates across multiple scenarios, used for the purpose of prompt engineering
- Relevant extracts from the market views of the Chief Investment Office document detailing information specific to Global Equities, released by the Chief Investment Office

The Tester is using the files provided to generate further synthetic samples. Generated outputs will then be validated using a combination of LLM as a Judge and NLP techniques, as well as human reviews.

Test Design

The following tests have been designed for the Email Drafting Tool.

Risks Tested	Evaluator Type	Metric(s)	High-Level Approach
Accuracy	LLM as a Judge, human judgement	Hallucination rate, contradiction rate	Checking for hallucinations or contradictions – with the input data (client profile, algorithmic outputs and market views of the Chief Investment Office) as reference.
Robustness	Pre-trained models	Cosine similarity of generated drafts with the same inputs (-1 to 1)	Comparison between different responses, when multiple email drafts are generated with the same inputs.
Completeness	LLM as a Judge	Percentage of completeness-related requirements met in the draft	Extracted a list of several completeness-related requirements for generated email drafts. An LLM as a Judge will be used to determine how many of those are satisfied.
Engagement	LLM as a Judge	Percentage of engagement-related requirements met in the draft	Extracted a list of engagement-related requirements for generated email drafts. An LLM as a Judge will be used to determine how many of those are satisfied.
Coherence	LLM as a Judge	Percentage of coherence-related requirements met in the draft	Extracted a list of coherence-related requirements for generated email drafts. An LLM as a Judge will be used to determine how many of those are satisfied.
Internal compliance	LLM as a Judge	Percentage of internal compliance-related requirements met in the draft	Extracted a list of requirements for internal compliance (e.g., promissory words, presence of certain calculated quantities in email). An LLM as a Judge will be used to determine how many of those are satisfied.

Table 4.1 Metrics used for Assessment of the various metrics

Test Implementation

The test execution process comprises of several key activities.

Overview of Testing Process

The generated outputs were manually generated through the Email Drafting Tool in the Deployer environment.

Stage	Responsible Party	Key Activities
Generation of synthetic data for testing	Deployer, Tester	The Deployer provided an initial dataset comprising 20 synthetic customer profiles, associated algorithmic outputs, and fixed market views of the Chief Investment Office, along with two draft emails per profile generated by the tool. The Deployer and Tester then collaborated to generate additional samples, including edge cases to assess the model's performance under atypical conditions. A file containing the tool's system prompts was also shared with the Tester. These were used to derive a structured set of requirements for evaluating the completeness, coherence, engagement, and internal compliance risks.
Sharing of test results	Deployer	The Deployer shared all generated test outputs and related data files with the Tester.
Analysis of test results	Tester	The Tester evaluated all risks outlined in Table 4.1 via the following three approaches : › Computational NLP metrics and pre-trained models › Human SME reviews › LLM as a Judge
Report sharing	Tester	The Tester compiled the findings, insights, and any recommendations into a formal Testing Report, summarising performance against the defined risk dimensions.

Table 5.1: Test Execution Process

Testing Implementation

The tests described in Table 4.1 were run in batch using the Tester's proprietary LLM Testing library, which provides a suite of modular pipelines for evaluating language model outputs through both LLM as a Judge and NLP-based techniques.

01 For Accuracy and Robustness existing LLM as a Judge and similarity-based pipelines were utilised.

02 For Completeness, Coherence, Engagement and Internal Compliance requirements, custom LLM as a Judge instances were created to parse and evaluate the outputs against a structured checklist derived from system prompts, and then determine whether each predefined requirement under each metric was met within the generated draft emails.

03 In addition to automated testing, a subset of outputs was reviewed by Human SMEs. These human assessments were scored using the same metric framework as the automated tests (see Table 4.1). This allowed for validation of alignment between the human and automated evaluations, and acted as a sanity check to ensure the reliability and calibration of the automated testing approach.

Resourcing and Cost of Testing

01 The Deployer is spending approximately 100 human-hours across a team of 5 personnel in the testing of the Email Drafting Tool. This team is primarily focussed on generating synthetic data, configuring test scenarios, and ensuring inputs are made available to the Tester.

02 The Tester is spending approximately 520 human-hours across a team of 5 personnel in the testing of the Email Drafting Tool. This team is primarily focussed on test planning and metric definition, development and execution of automated test pipelines, custom prompt engineering for LLM as a Judge evaluations, human SME review and analysis of results, and reporting and documentation of findings.

03 LLM usage costs incurred during the cost of the testing are estimated to be in the order of a few $100s. The exact amount depends on the final number of samples evaluated and the volume of prompt-based queries executed during the test cycle.

Challenges in Testing Implementation

The following challenges were encountered in the course of testing the Email Drafting Tool.

01 **Limited access to SMEs**

Interpreting and validating LLM-generated email content requires deep contextual and domain understanding. The nuanced language used in financial communications means SMEs are critical to determining whether an output is not just linguistically correct but also appropriate, accurate, and compliant.

02 **Access to Email Drafting Tool Versioning**

As the Email Drafting Tool is still in a pilot phase, some intended features are not yet fully implemented or available for testing. Only those features that are operational within the internal pilot can be actively tested. This means that the testing scope is limited to what is currently available to these users, and other features or potential functionalities cannot yet be evaluated

03 **Access to Large Number of Testing Data Samples**

While synthetic data was used to simulate real-world usage scenarios, crafting realistic and diverse synthetic profiles, especially for edge cases, requires significant domain expertise. The process was time-consuming, making it challenging to generate a large volume of representative and high-quality samples for broad coverage.

04 **Custom Tests**

A substantial technical effort was required to convert the Deployer's specific requirements, related to completeness, coherence, engagement, and compliance, into operational LLM as a Judge test prompts. As these requirements were highly contextual to the tool and its intended use case, this translation involved careful parsing of system prompts and iterative prompt engineering to ensure accurate and measurable evaluation criteria.

Insights/Lessons Learned

Given the highly regulated nature of both the Deployer and Tester organisations, and the fact that AI testing is still an emerging discipline, a considerable amount of time was required to navigate internal risk reviews and contracting processes before testing could formally commence. This highlights the need for clearer frameworks and precedents in future AI assurance engagements to streamline onboarding and execution.

The comprehensive LLM testing that was conducted has yielded valuable insights into the strengths and ambiguities of the Client Email Drafting tool. These findings, although confidential, provide a robust foundation for informed decision-making, guiding the tool's future development and enhancement.

The structured evaluation approach, combined with the utilisation of an LLM as a Judge framework, proved invaluable in this process. This methodology enabled scalable, consistent, and nuanced assessments of the generated email drafts, while significantly reducing the resources typically required for manual evaluation. By automating the evaluation process, we achieved a more efficient and repeatable analysis, facilitating rapid iterations and continuous improvement of the tool’s performance.