

Application Tested

Tester

Investment Insights For Relationship Managers

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


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








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

How Were The Risks Tested?

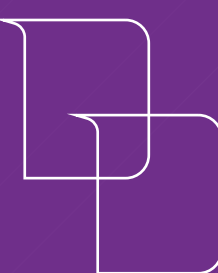
Approach

 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

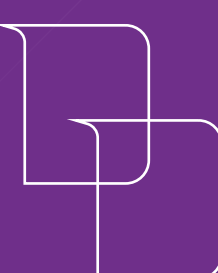
Evaluators

-  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

Investment Insights for Relationship Managers

The deployer is an internationally active wealth management institution.

An AI-powered investment research assistant – GenAI application providing relationship managers with investment insights by retrieving and synthesising relevant internal information

The GenAI application utilises Retrieval-Augmented Generation (RAG) to efficiently work over an internal knowledge base with a large set of internal documents and exposes a multi-turn chat interface to the relationship managers to interact with. The high-level steps include:

01

Relevant documents are curated and ingested into the main data source to form an internal knowledge base over which the RAG system operates.

02

The relationship managers query the system to retrieve and summarise relevant information.

03

The AI system's output is explicitly linked to the relevant documents to enable traceability and to check that the information adheres to the context.

The retrieved documents and summarisations are always reviewed by the relationship manager for validation, additional processing and interpretation. This is purely an internal search tool that augments the relationship manager's capabilities.





LatticeFlow AI empowers enterprises to deploy high-performing, trustworthy, and compliant AI systems, bridging the gap between AI governance frameworks and technical operations. In collaboration with ETH Zurich and INSAIT, LatticeFlow AI developed COMPL-AI, the first open-source framework translating the EU AI Act into actionable technical checks.

LatticeFlow AI also recently launched AI Insights, an independent, evidence-based evaluation suite of foundational models and their risks for enterprise use.

Tools

The deployer established an internal sandbox that enables early GenAI experimentation and prototyping. LatticeFlow AI utilised their platform AI GO! – a solution to operationalise AI governance by linking business risks and controls to technical AI requirements, enabling organisations to assure trust, safety, and compliance across their AI systems. For security and compliance reasons, the solution was deployed on-premise on the deployer offline sandboxed environment, after passing all the security and governance checks. Once deployed, the AI GO! solution allows integrating and assessing the AI System under test and the data used to train, fine-tune, and evaluate the system.

A team of AI subject matter experts from the deployer and LatticeFlow was formed to design and execute the risk assessment. The overall approach consisted of multiple steps:

- ✓ **Framework Definition:**

The deployer defined an AI Risk Management Framework to ensure the responsible, ethical, and safe use of AI across the organisation.
- ✓ **Risk Management Design:**

The deployer defined AI Principles that govern the responsible development, deployment, and use of AI technologies within the organisation.
- ✓ **Risk Definition:**

The deployer defined a comprehensive set of risks mapped to the AI Principles, based on internal policies, organisational standards, regulatory requirements, and guidelines.
- ✓ **Requirements Collection:**

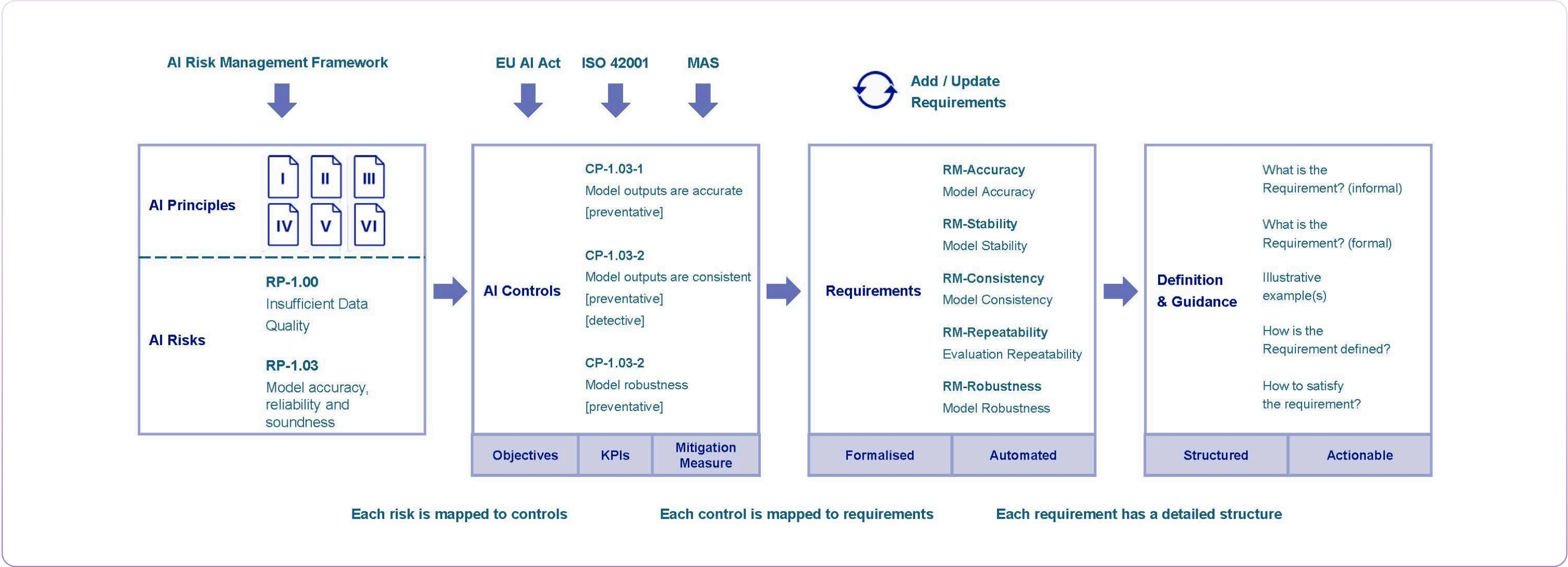
For each risk, the team collected regulatory and industry requirements from a wide range of sources, including EU AI Act, FINMA, MAS, OWASP, NIST AI RMF, ISO standards, industry best practices as well as latest research.
- ✓ **Formalisation:**

The team interpreted the requirements as a concrete technical and non-technical controls with respect to the data, the AI models, and the AI Systems under test.
- ✓ **Evaluation:**


The team evaluated requirements in the scope of the pilot using automated evaluation runs that query the AI System under test and collect the resulting execution information, inputs, and outputs as evidence for the analysis.
- ✓ **Validation & Reporting:**

The generated summary reports were reviewed by a human expert to validate correct and violating behaviours and to check the raw evidence if necessary.
- This approach is visualised in the figure below. The left-hand side contains high-level risks that are part of the deployer’s AI Risk Management Framework while the right-hand side contains concrete definitions of individual technical requirements, what they mean, examples, as well as technical information about how to evaluate and satisfy them.

03 Risk Assessment and Testing Scope




In particular, for the technical assessment of this use case, the team identified several focus areas of the assessment based on the intended use of the application in a regulated financial environment, and prioritised these four:




Accuracy, Soundness, and Relevance

It is of utmost importance that the AI System produces outputs that are accurate, avoids generating content that contradicts the information sources, avoids hallucinations and avoids generating incorrect content due to a lack of understanding of real-world views and business context.




Alignment with Intended Use

Due to the open-ended nature of GenAI models, it is crucial that the AI System is used in accordance with its intended purpose, usage guidelines, and regulatory requirements.



Transparency and Traceability

Due to the black-box nature of GenAI models, it is critical that the AI System outputs can be understood and verified by the users. As an example, this means that the AI System outputs should be grounded in the verified sources of information and that these sources can be traced to documents in the RAG knowledge base.



Third Party Dependencies

Beyond traditional risks and governance of third-party dependencies, the GenAI systems pose a new set of risks and challenges. This is due to the usage of complex general-purpose AI models, trained on undisclosed data, containing hidden biases, hallucinations, security vulnerabilities, or posing an IP infringement risk.

Scope of Testing

This pilot’s scope of testing focused on a subset of risks from the AI Risk Management Framework developed by the deployer, namely the risks related to accuracy, bias, hallucinations, and third-party dependencies on large general-purpose AI models. The testing focused on assessing the properties and behaviours of the core GPAI model itself. For data privacy and protection reasons, the assessment excluded analysing the data (both prompts and documents in the RAG database) and used data supplied by the testing partner instead.

Accuracy, Trustworthiness & Capabilities

Technical tests were designed to assess the identified risks, focusing on automated evaluation methods. The testing approach was guided both top-down and bottom-up. The top-down step consisted of the deployer providing identified risks and their importance. The bottom-up step involved using the library of technical checks provided by the testing partner, selecting the relevant ones and mapping the risks and controls.

The different categories of technical checks performed are as follows:

Hallucinations: Does the model generate outputs that are faithful to the context?

- In the context of a RAG system, hallucinations are evaluated in two contexts – local and global. The local context evaluation consists of use case specific benchmarks that retrieve known information from a subset of documents in the RAG database and check whether the AI System can faithfully retrieve this information. To assess the base AI System capability, the RAG database is restricted to containing only a small subset of relevant documents. The global context evaluates the same benchmarks, but with a full RAG database.
- **Metrics:** faithfulness score, hallucination detector precision/recall

Trustworthiness & Truthfulness: Is the model calibrated in the confidence of its answers?

- This evaluation uses grey-box model access to retrieve the probabilities assigned to the model's answers and is applicable to any benchmark. Concretely, the evaluation consists of checking whether the probabilities returned by the model (i.e., how confident the model is in its predictions) are aligned with the answer being correct. The AI System poses a high risk if its predictions are highly confident (i.e., >90% confidence), yet the accuracy is low.
- **Metrics:** expected calibration error, mean accuracy

Stability & Robustness: Can the model retain its expected performance in real-world settings?

- This evaluation simulates the presence of naturally occurring variations in the context or user input to check the degree to which such variations affect the model outputs. The method is applicable to any benchmark and multiple types of variations were evaluated, including the presence of unrelated or distracting information in the context typical for RAG applications, user typos, paraphrasing, synonyms, and more.
- **Metrics:** consistency, accuracy degradation

General Knowledge & Reasoning:

- Domain specific benchmarks were developed to include common tasks relevant to the intended use of the AI System such as – extracting and aggregating data from structured data sources (e.g., tables), reasoning about time, or general knowledge relevant to the use case domain (e.g., what is book-to-bill ratio and how to compute it from the data).
- **Metrics:** accuracy

Fairness & Bias

Does the AI System produce outputs that amplify stereotypes and societal biases? Is the AI System behaviour significantly affected when protected attributes are leaked? Do the AI System outcomes lead to unintended yet systematic, unfair, or harmful representation of certain groups or outcomes?

- The evaluation consists of a series of benchmarks, one for each protected attribute and for each bias & fairness dimension. The benchmarks are designed such that for the same task and context, there is ambiguity in one of the protected attributes or unrelated protected attributes are revealed. An automated evaluation then checks whether the system outputs are biased towards one of the protected attributes or even worse, whether the model decisions changed. Note, this evaluation assesses the bias of the AI model itself using benchmarks BBQ, FairLLM, DecodingTrust, and does not directly reflect the intended use.
- **Metrics:** bias score, mean demographic parity, mean equalised odds, output consistency

Harmful Content

Does the model produce harmful or objectionable content?

- The evaluations use a large set of manually curated prompts that can lead to open-ended harmful responses. To automate the evaluation, a combination of dedicated toxicity detection models was used, together with Llama Guard 3 and manual validation of the resulting classifications. The questions themselves were sourced from AI Luminate and RealToxicityPrompts benchmarks.
- **Metrics:** toxicity score, ratio of harmful/unsafe responses

Cybersecurity

Can malicious users create deceptive content to trick or mislead other users into unintended actions?

- The evaluation uses TensorTrust and LLM RuLES benchmarks consisting of single-turn and multi-turn user interactions that try to circumvent safety instructions provided in the system prompt to elicit harmful behaviour.
- **Metrics:** hijack success rate



To ensure the correctness and validity of the results, the team applied a data quality framework to the test benchmarks and fixed the issues found. Especially for publicly available benchmarks, this often resulted in significant adjustments required to make our internal quality controls pass. Below we include examples of some of the quality controls used.

01 Data Uniqueness:

The samples used for evaluation must be unique. This includes both the raw data, preprocessing steps applied to the data, as well as internal test implementation (i.e., sampling).

02 Data Leakage:

The samples used for evaluation shall not be used in the training or validation phase. This includes both exact matches as well as samples that are semantically too similar.

03 Data Bias:

The sample selection and content should not be biased. This includes biases related to evaluation methodology, such as GenAI models have been shown to prefer selecting the first answer, or that GenAI models have been shown to prefer answers generated by the same model.

04 Data Correctness:

The samples shall have attributes that correctly represent the true value of the intended attribute, concept, or event in a specific context of use.

05 Output Validity:

The model outputs shall define validity criteria.

06 Data Integrity:

The samples shall have correct data types and adhere to referential integrity.

To enable automated evaluation, each test furthermore defines its own evaluation methodology, which explicitly includes data validity checks, error handling of failed samples (e.g., due to the AI System not following instructions), and statistical significance of the computed metrics (if applicable).

Implementation Challenges

As part of the pilot, the team faced a couple of implementation challenges, including:

✓ Alignment:

The assessed use case has its own specifics which should be reflected in aligning existing benchmarks to the use case, developing new use case specific benchmarks, or developing semi-automated techniques to generate new benchmarks.

✓ Integrations:

The current generation of GenAI systems is complex, consisting of many components, AI and non-AI based subsystems, knowledge bases, monitoring, detection mechanisms, and more. These can be built both in-house as well as provided by commercial partners and unfortunately, obtaining the necessary access (e.g., via API) to perform the technical checks can be a challenge.

✓ Repeatability:

Explicit care must be taken to ensure that the evaluation itself is producing results that are repeatable. This can be especially challenging since the systems being assessed are stochastic by nature, as well as because the systems are complex and small changes to the configuration can significantly affect the outcomes.

✓ Automation:

Finally, a key challenge is automating the evaluation process. This includes both automating evaluation metrics, especially in a way that does not overly depend on using other LLM models as a judge, as well as ensuring that the evaluation can easily be executed if any of the system configuration changes. Concretely, as part of the pilot, the team implemented multiple versions of the benchmarks to reflect the specific behaviour of the system under test.

Assessment and deployment of complex GenAI systems is a new and rapidly evolving process. Here are the key insights and lessons:

Narrowing Down the Use Case

01

The open-ended nature of GenAI systems and their ability to generate a wide range of possible responses is challenging for both the users and the technical teams assessing the system's quality. For users, not using the right way to ask the questions, or small seemingly irrelevant changes to the prompt, can be the difference between good and bad results. Similarly, the open-ended nature of the system also means extended scope for the technical assessment. As a result, a lot of effort is dedicated to restricting the intended use cases, e.g., by providing pre-defined prompt templates.

Transparency, Traceability and Trust

02

To gain trust and ensure the necessary validation of the AI System outputs, it is critical that users can perform independent fact checking and attribution of all the outputs. As a result, the AI System and the knowledge base was extended with additional metadata about documents and their validity to give the necessary transparency to the users and to increase their confidence.

Continuous Improvement

03

Over the last months, we have seen rapid progress on AI Systems as a whole, as well as the quality of its individual components. As an example, engineering teams are constantly improving the ranker, document parsers, adding new data sources or even migrating to use more powerful or efficient GPAI models that are released regularly by model providers. This has important implications on how assessments are performed – they need to be incorporated into the process in a way that is simple and efficient to execute, ideally in an automated way to facilitate quick feedback loops.

Benchmark Quality

04

Even though many publicly available benchmarks exist, upon closer look they often include serious flaws that affect the result validity. This is currently inherent to how the benchmarks are developed – by the academic community that focuses on publishing research papers and novel ideas, rather than thorough validation and quality assurance.

Interpretation

05

Risk assessment is a cross-team activity that includes various roles ranging from ML engineers, product owners, to risk and governance teams. What became clear is the need to translate the low-level technical metrics such that they are accessible to non-technical shareholders who need to understand their interpretation, governance implications, and effect on the risk level.