GLOBAL AI ASSURANCE PILOT

Scan to read the full case study.

Investment Research Assistant

Application Tested

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.

Tester

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

Investment Research Assistant

UNIQUE

Unique is the vertical leader in agentic AI for the financial services industry, providing cutting-edge solutions that empower financial institutions to thrive in a rapidly evolving landscape. It has over 40 customers, including LGT Private Banking, Pictet Group, Julius Baer, SIX, and other blue-chip finance firms. Unique has already certified their AI Management System according to ISO 42001 and is one of the first European companies to hold this certification leading the way in secure and safe GenAI applications for Financial Services.

Use Case

Unique's Investment Research Assistant is designed to streamline the process of querying stock universes and preparing detailed investment recommendations for banks and other financial services providers. The assistant can:

01 Extract stock data based on natural language queries that specify customer interests and criteria, such as sustainability ratings or buy signals.

02 Load fact sheets of identified stocks into a Large Language Model (LLM) to extract relevant rationales, providing detailed insights and justifications for each stock.

03 Generate follow-up emails to customers, including detailed investment stories and attaching relevant fact sheets.

Unique's technical architecture is designed to deliver enterprise-ready AI solutions, particularly for the financial services industry. The platform employs a Retrieval-Augmented Generation (RAG) approach, combining retrieval-based and generation-based methods to ensure accurate and contextually relevant responses. The platform also supports flexible deployment options, including cloud and on-premises setups, tailored to client needs.

Unique does not train its own models but works with pre-trained models from providers (LLMs) like Microsoft, Google while also supporting client-specific fine-tuning for proprietary data (on prem solution hosted by the client).

The system emphasises data security and compliance, hosting data in specific locations (e.g., Switzerland for Swiss Banks) and adhering to strict standards like SOC 2, ISO 27001 and relevant regulations. Unique's architecture includes features like data access control, pseudonymisation, and a human-in-the-loop mechanism to ensure privacy and accuracy.

Unique's AI governance framework incorporates continuous benchmarking, user feedback loops, and advanced prompt engineering to mitigate risks like hallucination and model drift. This recently has been certified with ISO 42001.

Finally, Unique's tools and services include connectors for internal knowledge bases, APIs for external data integration, and specialised modules for tasks like meeting transcript analysis. These capabilities are supported by a robust compliance layer and a focus on explainability and transparency in AI outputs.

02

Testing Partner and Testing Approach

QUANTPI

QuantPi GmbH is a Germany based spin-off of the CISPA Helmholtz Center for Information Security, the globally leading research center for Cybersecurity. QuantPi has translated over 8 years of research in mathematics and machine learning into a world-leading testing-technology.

PiCrystal is the AI testing engine at the core of QuantPi's platform. PiCrystal allows users –business experts, compliance experts, regulators, or other non-technical users – to understand the behaviour of any AI black-box by rigorously testing and assessing input and outputs. It offers standardised assessment for bias, fairness, robustness and more across any AI portfolio, and is model and data-type agnostic (Tabular, Text, Image for traditional and generative AI, Multi-modal). Additionally, it enables users to gain clarity through accessible visualisation of the testing logic used. It reduces computational costs through scalable testing automation and removal of redundant workloads.

PiCrystal is made of Embedders, Perturbers and Metrics.

01 **Embedders are data Categorisers/ Annotators and are use case specific.**

02 **An embedder is a function whose purpose is to extract properties from data. Embedders represent test scenarios.**

03 **Metrics are functions that take as input the model, a dataset, and possibly test-specific parameters and outputs values that represent model performance, robustness or fairness.**

04 **Perturbers are use case specific functions that add some augmentation/modification (noise) in the input data, enabling robustness testing.**

Using the PiCrystal Metric Compute python library, data scientists are able to generate and execute a suite of scenarios to evaluate black-box models on given data.

03

Risk Assessment and Testing Scope

As part of the project, the QuantPi and Unique teams agreed to focus on two primary use cases: Investment Research Assistant and Document Search. Various risks were identified with respect to the two use cases:

Non-Adherence to Consumer Protection and Duty of Care Regulations

- Impact: Risk to the firm's reputation and exposure to potential regulatory penalties

Non-Adherence to Internal Policies

- Impact: Potential inconsistencies in advice quality, or violation of internal guidelines like offering crypto currencies to a client who is not allowed to trade this

Poor Performance for Specific Groups

- Impact: Concerns about fairness and inclusivity in the advice provided

Inadequate Clarity in Explaining Rationale behind Recommendations

- Impact: Reduces advisor's ability to justify recommendations to clients, potentially eroding trust

Provision of Excessive Client-Specific Information

- Impact: Potential for data misuse or over-reliance on the application, compromising advisor's judgment

Potential for Advisor Misuse (to generate advice that benefits them financially but is not in the client's best interest)

- Impact: Ethical and regulatory risks, undermining the firm's duty of care

Poor Financial Outcomes

- Impact: Poor client satisfaction and firm's financial stability

Biases in underlying models (LLMs): e.g.,

- Market bias: overemphasising an irrelevant market
- Demographic bias: recommendations that are less appropriate for underrepresented groups
- Survivorship bias: over focusing on companies that survived, and not examining companies that failed
- Impact: Could lead to wrong investment decisions, biased investment decisions towards certain companies or countries

Test Design

Investment Research Assistant

In a pilot test of the Investment Research Assistant, a RAG system, QuantPi assessed **risks of inaccurate results, hallucination, and unfaithfulness.** This was quantified by calculating Cosine Similarity thresholded at 0.4 and 0.8 and Faithfulness across varied query lengths, question types, domains, complexities, and the introduction of typos in queries.

The metrics are defined as follows:

Cosine Similarity

Cosine similarity is a way to measure similarity between texts or documents based on their intrinsic meaning, rather than looking at the occurrence of the exact words. Mathematically, cosine similarity measures the similarity between the vector representations of the two texts.

In the context of this use case, the two texts are the ground truth and the predicted response generated by the agent. **Cosine Similarity was utilised and threshold it at 0.4 and 0.8**, to measure what proportion of the dataset crosses a specific threshold. A higher similarity score is considered better.

Faithfulness

Faithfulness metric assesses whether the response generated by the agent is grounded or not, i.e., how aligned is the agent's response with respect to the “Context” that was provided to it along with the query. Typically, a higher faithfulness is an indicator of lower hallucination. (A faithfulness score of 1.0 indicated that the response is generated purely based on the context, and nothing from outside the context).

The metric takes two inputs - the predicted response and the context (as a list of retrieved texts from the search layer). Both the inputs are decomposed into a list of claims, and we check how many of the claims in the generated response are also a part of the context (uses an LLM under the hood to make this similarity check).

Document Search

During the Document Search pilot, QuantPi focused on evaluating the search layer, identifying two potential failure modes: **inaccurate and irrelevant results**.

In general, the search system responds with a list of retrieved texts based on the limit parameter that is set. Test was conducted with the limit parameter set to 10, i.e., 10 most relevant chunks of text were retrieved. To quantify the quality of the search system, Lenient Search Accuracy, Mean Reciprocal Rank and Word Overlap Ratio were calculated across diverse input queries (varying lengths, question types and typos).

Word Overlap Rate:

Gives a measure of how many words in the ground truth is a part of the retrieved context from the search layer. This metric takes a bag-of-words approach, looking only at the occurrence of words and not the order of the occurrence. This has been done since often the retrieved context has a lot more information based on the chunking strategy. Ideally, the more words in ground truth that are a part of the retrieved text, the better it is, since it is an indicator that the question can be answered based on the retrieved context. Since retrieved context is a list of text chunks, we take the one with the highest overlap by default.

Mean Reciprocal Rank:

Mean Reciprocal Rank (MRR) is a critical evaluation metric for retriever systems, particularly in applications like search engines, recommendation models, or RAG systems. In the context of this use case, it measures how effectively the search system ranks the first relevant result (from a list of results) for a set of queries. Reciprocal Rank (RR) for a single query is 1 if the first relevant result is ranked 1st, 0.5 if ranked 2nd (1/2), 0.33 for 3rd (1/3), etc.

Lenient Retrieval Accuracy:

Lenient Retrieval Accuracy assesses the presence of relevant context within the set of retrieved documents, disregarding the order or rank. This metric yields a binary output (1 or 0) for each query, denoting whether the relevant context was identified or not, respectively.

Based on our post-assessment analysis, Word Overlap Ratio seems the most relevant metric for the given use case with the given dataset, since it looks at the occurrence of all the relevant words in the retrieved texts and does not look for an exact match of the word order between the ground truth and the retrieved texts. This is crucial since the test dataset is synthetically generated and could have quality problems, and the search mechanism under the hood is a combination of multiple different searches (e.g., elastic search, vector search).

05 Test Implementation

Set-up: organisational aspects

- NDA was quickly established using an existing template.
- Asynchronous communication via Slack worked well.
- Use case selection was agreed upon quickly, though test design was more challenging due to time constraints
- Resource requirements:

  Unique: 1 project manager, 1 data scientist (~10% workload per week)

  QuantPi: 1 project manager, 1 data scientist (~10% workload per week)

Test Execution

- Given a collection of annotators, a collection of metrics, and a collection of perturbers that can be used on the same set of inputs, outputs, and labels, PiCrystal generates assessment scenarios. These scenarios can then be used to assess the black box on the dataset. An advantage of this approach is that the same scenarios can be used on any other black box having the same signature (i.e., input and output format) and on any other dataset having the same format.
- Testing was conducted in a secure staging environment with strict access controls.

Data Used in Testing

- Automated testing: Used approximately 100 samples from the deployer, augmented by 100 modified samples from the tester. 200 samples in total.
- Data sources: Anonymised, but real, data from the deployer; Perturbed (i.e., synthetically modified) data from the deployer.

Cost of Testing

- Time allocation:

  12 hours: Deployer organisation's technical team for pipeline integration and data prep.

  4 hours: SMEs for manual review and gold standard creation.

  60 hours: Testing organisation's team for platform setup, execution, and analysis.
- Direct costs (model): $3.60

Challenges in Implementation

- Finding test data was the hardest challenge.
- Setting up access rights was complex due to high security standards on both sides.
- Setting up test accounts on both sides required more effort than expected.
- Test results showed low statistical significance due to small sample size.
- Interpreting and generalising results was difficult due to limited data.
- Efficiently integrating manual review workflows within the automated platform.
- Ability to test subsystems to the agent/RAG that influence final results
  - Time limitations
  - Resource caps
  - Desire to protect intellectual property

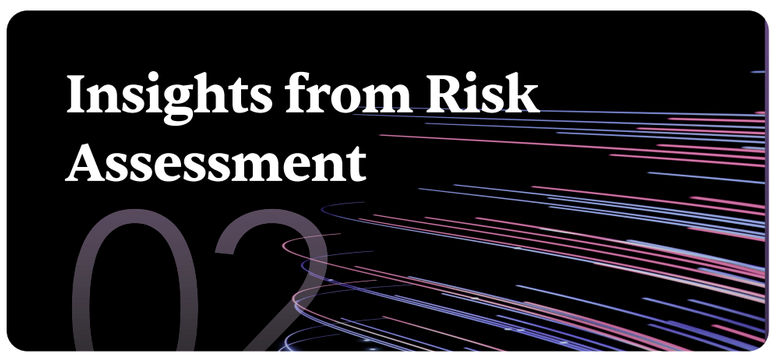
Areas for further research and collaboration

Unique and QuantPi identified the following areas to further test and the scope is outside of the AI Verify pilot:

Guardrail testing:	How does the introduction of a NIM Guardrail impact the output?
Location testing:	How does the introduction of location based rules impact the output?
Performance deviation across conversation turns:	Does the agent still perform as expected after 10 user messages?
Jailbreak scenarios:	How easy or hard is it to manipulate the assistant and circumnavigate guardrails?
GenEngine comparison:	How does switching out the LLM at the core of the agent change results? Azure GPT, Mistral, and Claude Sonet to be tested.

Insights/Lessons Learned

Deployers have resource constraints and are forced to focus on deploying new products. They really rely on testing partners and experts to evaluate solutions in the pre-beta phase.



Since GenAI is so new (and new models are emerging constantly), it is hard to define the most important risks. There are risks related to underlying data (also constantly changing), risks related to the LLM itself (and model providers), risks related to prompting and social engineering.

Test data is very hard to get and defining a “golden dataset” is challenging. Also, defining the right metrics to measure/test across different use cases, different prompts and changing underlying data sources remains a challenge. This makes the task challenging to decide which LLMs and prompts to use for which use cases.

Shareable data on sensitive attributes. Regulation prevents the collection of certain types of information (for example, ethnicity), which prevents testing whether or not the model is sensitive to that information.

Testing with a small dataset/sample does not allow for generalisation of results.

Interpretation of test results of one use case is difficult as the comparison is missing (e.g., is 0.4 a good or bad value can only be evaluated with a comparable use case).

Lack of best practices causes delays - we must all invent the wheel from scratch.