GLOBAL AI ASSURANCE PILOT

Scan to read the full case study.

UltraScale No-code AI-powered RAG Platform

Application Tested

ultra mAInds, an AI engineering and solutions company, offers UltraScale, a 'No-Code AI-powered RAG platform for Enterprise search and data connectivity.

Tester

AIQURIS

AIQURIS (assurance partner) is a Singapore-based corporate venture offering a SaaS platform that helps specify and manage the required evidence for confident adoption of AI. The platform is based on a systematic approach to establish risk profile and necessary controls to ensure AI systems are deployed safely, compliantly, and effectively at scale.

AIDX

AIDX Tech (testing partner) is an AI testing platform that offers model evaluation, safety and risk management, and consulting services.

How LLMs were used in application?

- Summarisation
- Retrieval augmented generation
- Data extraction from unstructured source
- Translation
- Multi-turn chatbot

What Risks Were Considered Relevant And Tested?

- Accuracy of translation
- Potential harmful content generation across diverse languages

How Were The Risks Tested?

Approach

Accuracy of translation

FLORES+ benchmark dataset containing 997 sentences translated from English to other languages as well as between non-English language pairs

Harmful content generation

AIDX GenAI Safety Benchmark, targeting Robustness, Ethics and Society, Fairness, Privacy and Security, Toxicity, and Legality

Evaluators

Accuracy of translation

- BLEU Score and Sentence Embedding Similarity metrics were calculated using automated testing
- A subset of translations was manually reviewed to verify nuanced accuracy in grammar, meaning, and context

For harmful content detection, success rate in the face of adversarial prompts was assessed

Challenges

API Performance

Latency issues, averaging approximately 1,000 translations per hour (internal safeguarding of API abuse but poses a potential performance bottleneck for testing)

Stability Issue

After prolonged test runs, instabilities were encountered indicating the need to individually monitor third party services (e.g. resource exhaustion, connectivity)

Translation Reliability

Some translations failed intermittently without a clear pattern, again raising the need to monitor performance of individual software components

Insights

01

Challenges in designing and implementing automated tests for multi-lingual accuracy and content safety:

- Lack of standardised translation ground truth
- Inadequacy of surface level metrics like BLEU
- Inability to catch nuanced or implicit harms
- Limitations of using API calls for large-scale multi-lingual testing

02

Importance of structured risk assessment process to determine what to test

03

Strong role for human experts as a result of above challenges

Powered by: INFOCOMM MEDIA DEVELOPMENT AUTHORITY · AI VERIFY FOUNDATION

No-code AI-powered RAG platform

ultra mAInds is an AI engineering and solutions company, covering the entire AI transformation journey from architecture to implementation. ultra mAInds offers custom delivery models based on client needs, including capability services, AI technology solutions, and industry-specific AI-driven solutions.

UltraScale is one of ultra mAInds offerings and provides a 'No-Code AI-powered RAG (Retrieval-Augmented Generation) platform designed for enterprise search and data connectivity. By integrating over 550 SaaS interfaces, databases, and internal systems, UltraScale enables seamless access to dispersed data across an organisation.

Use Case

The pilot focused on a GenAI application designed to interact with an organisation's knowledge from a variety of unstructured sources. As part of this test pilot, two AI agents were selected to be reviewed.

- YourGPT

  An AI agent with a rich predefined prompt library that grants seamless and unified access to all LLMs within the UltraScale ecosystem. The prompt library is tailored for tasks such as creative exploration and advanced problem-solving.

- TranslationAgent

  The AI agent allows to translate any PDF document into multiple target languages while preserving document structure. It is intended to be used for international communication and global collaboration within and external of an organisation.

High level Architecture

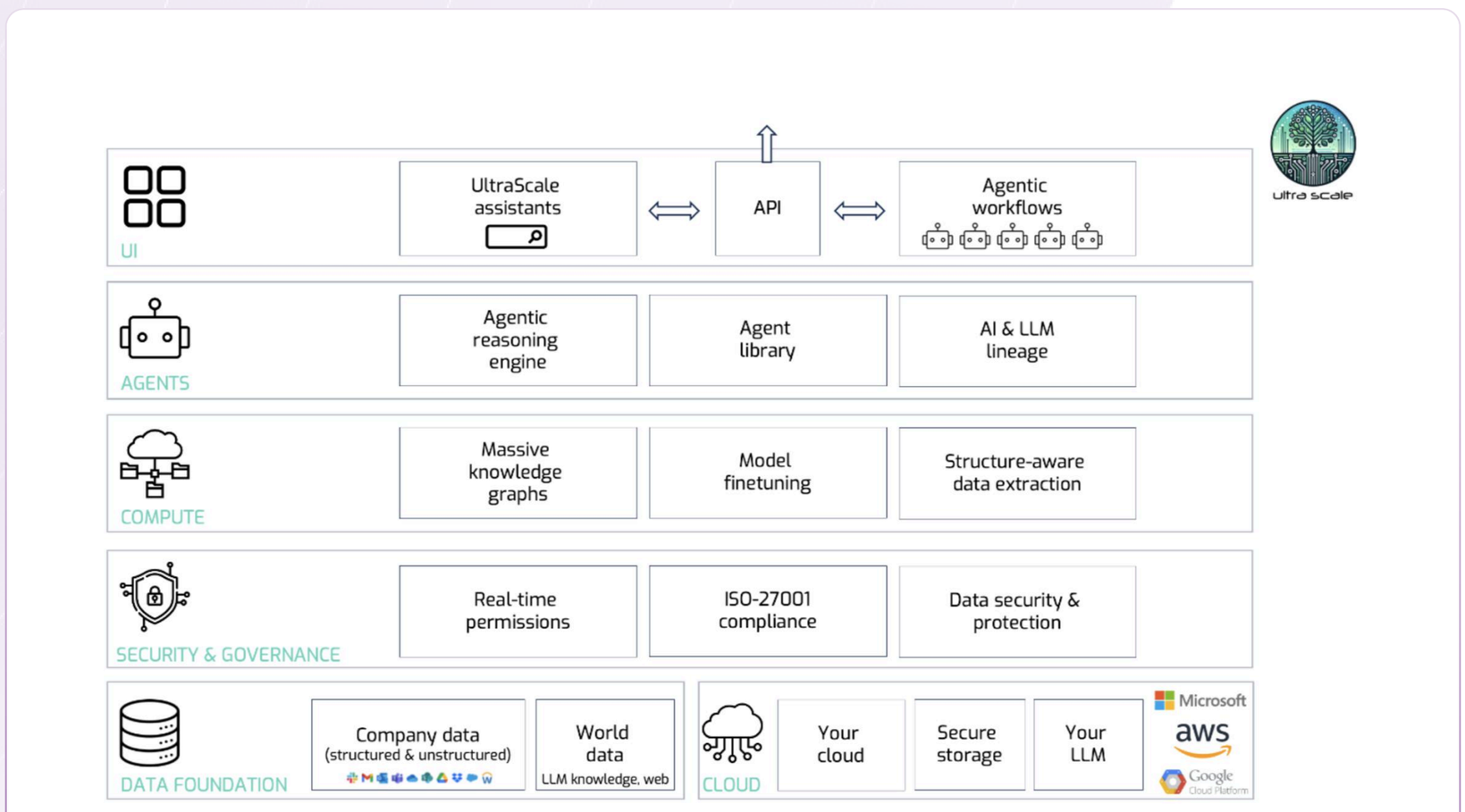


Figure 1: High level system architecture

01 Deployer and Application

The UltraScale system is an enterprise-grade AI co-pilot designed to empower users to interact with their unstructured and structured data through intelligent agents. It provides a secure, scalable, and explainable pipeline for document ingestion, understanding, and interaction. The end-to-end flow is indicated below:

Document Upload & Data Source Linking:

Users upload files or connect external data repositories.

AI Pipeline Activation:

The system transcribes, indexes, and semantically enriches the content.

Agent-Driven Processing:

Agents perform tasks such as summarisation, Q&A, extraction, and translation.

Output & Feedback Loop:

Users view results, interact with agents (e.g., chat), and can iteratively refine outputs.

User-in-the-Loop:

Throughout the process, users remain in control to validate and guide AI actions.

02 Assurance/ Testing partners and approach

Assurance partner:

AIQURIS Pte Ltd (AIQURIS) is a Singapore-based corporate venture offering a SaaS platform for AI quality and risk management. The platform enables organisations to assess, monitor, and mitigate risks associated with AI systems through automated profiling, compliance tracking, and advisory support. In this use case, AIQURIS provided a risk and quality profile for the AI system, from which risk mitigation requirements can be derived, including process and technical controls such as AI tests.

Testing Partner:

AIDX Tech Pte Ltd (AIDX) is a trustworthy AI model testing platform for AI risks, harmful content and reliability testing, verification and risk management. AIDX provides evaluation services of AI models and machine learning data, as well as AI training and consulting for AI developers, auditors, and adopters.

Testing Approach & Tools

For the use case in this pilot, AIDX provided an automated multilingual AI testing solution that evaluates model accuracy and potential of harmful content generation across diverse languages. The platform assesses linguistic fidelity, semantic alignment, and robustness against adversarial threats. The entire process, from test execution to report generation, is fully automated, while a human expert reviews a subset of outputs to ensure domain-specific accuracy and compliance. The result is a concise evaluation report that documents quantitative benchmarks and expert validation.

03 Risk Assessment and Testing Scope

Based on AIQURIS's comprehensive 6 pillar methodology, several high and critical risks were identified for the UltraScale AI system across the Legal, Security, Ethics, and Performance pillars. The pillars Safety and Sustainability have been assessed to not pose unacceptable risks.

- The risk profiling found that **Legal risks** are most pronounced, with critical concerns relating to **accountability**, **contractual compliance**, **safeguarding of confidential data, non-discrimination**, and the **potential for misuse or abuse of the system**, which in aggregation exposes the organisation to potential legal liabilities.
- **Security risks** are elevated, confirmed by ambiguity in roles, insufficient audit trails and the AI specific vulnerabilities that could lead to **undetected manipulation or manipulated system output.**
- **Ethical risks** were found to include **privacy** and the **potential for discrimination**. The AI system produces content such as marketing material and emails, which may entail gender, age, race or other forms of unwanted bias.
- **Performance risks** are high in suitability, efficiency, compatibility, robustness, and reliability, reflecting the need for the system to consistently deliver **accurate, timely, and contextually appropriate outputs** under complex and variable conditions.

To address these risks, both process and technical controls are required. Given the scope of the AI Verify Foundation Assurance Pilot, the remainder of this document covers only the tests conducted against (a subset of) the technical controls. Translation and protection against harmful content generation were selected to be tested in this pilot. The former was particularly relevant because the translation of technical documents that preserves layout is a key distinguishing feature of the UltraScale platform.

04

Test Design

Two primary types of tests were conducted for this project:

- Machine translation quality
- Content generation behaviour evaluation

AIDX's testing framework was designed to evaluate both the accuracy and potential of harmful content generation of multilingual LLM outputs using a combination of automated tools and expert human review.

Accuracy Tests (automated)

Accuracy was assessed using the FLORES+ benchmark dataset. Target languages included French, German, Italian, Russian, Japanese, Chinese, and Spanish. Evaluation metrics included:

- BLEU Score (automated) – measuring n-gram overlap with reference translations to evaluate the quality of text that has been machine-translated from one language to another
- Sentence Embedding Similarity (automated) – evaluating semantic consistency. A subset of translations was manually reviewed to verify nuanced accuracy in grammar, meaning, and context.

Harmful Content Tests (automated)

The AIDX GenAI Benchmark was used to gauge the potential of harmful content generation, which tests for predefined risk categories without red-teaming or prompt injections.

- Model responses were scored using automated detection tools, and the attack success rate (ARR) was calculated based on predefined criteria.
- A subset of high-risk responses was manually reviewed to ensure accurate interpretation of harmful or sensitive content.

05

Test Implementation

Execution of Tests

Tests were executed using the AIDX automated GenAI evaluation platform, which integrates high throughput automated testing with targeted expert review. The testing was conducted in a secure staging environment with strict access controls.

- For automated tests, the benchmark datasets were processed in batches via the UltraScale chat API, with structured prompts enforcing JSON-formatted responses. The platform automatically computed BLEU, sentence embedding similarity, and attack success rate metrics to evaluate translation quality and harmful content generation risks. Outputs below predefined confidence thresholds were flagged for further inspection.
- For manual tests, selected test cases (e.g., low-scoring results and responses in high-risk categories) were routed to expert evaluators via a secure AIDX review interface.

Data Used in Testing:

- Machine translation quality was evaluated using a benchmark dataset across eight languages (FLORES+): English, French, German, Italian, Russian, Spanish, Japanese, and Chinese. A total of 997 sentences from these languages were assessed. The evaluation process involved making structured API calls, with each call submitting one sentence and requesting translations into the remaining seven target languages.

  All translation results were stored in the AIDX database for subsequent analysis. Upon completion of the project, all test data and results are purged.

- The content generation behaviour of the UltraScale platform was evaluated using an AIDX benchmark dataset. The evaluation focused on six key dimensions: Robustness, Ethics and Society, Fairness, Privacy and Security, Toxicity, and Legality. Approximately 160 prompts per language were submitted via API calls to assess the model's behaviour across these dimensions. The responses were analysed using the AIDX benchmark evaluator, which assigned a quantitative score ranging from 1 to 5. Upon completion of the project, all test data and results are purged.

Cost of Testing

The AIDX AI evaluation platform operates in the Azure cloud environment. The testing process spanned approximately two weeks, covering API connectivity debugging, test process monitoring, and preliminary result analysis. The costing only reflects compute costs, as expert costs and project overhead were waived for this pilot. It is expected that automation will steadily increase and minimise the need for expert intervention.

	Description	Cost (SGD)
Azure Container App	For API call, result analysis and partially report generation, total 2 instances used for this project	$ 150.00
Azure firewall	To protect working environment	$ 100.00
Azure MySQL database	Persistent transactional data for result analysis	$ 20.00
Azure GPU (NC6) about 80 hours	Use language model for result evaluation, last about one week	$ 150.00
Third party managed AI services	The UltraScale platform integrates with third party services and monitors usage in credits.	$ 100.00

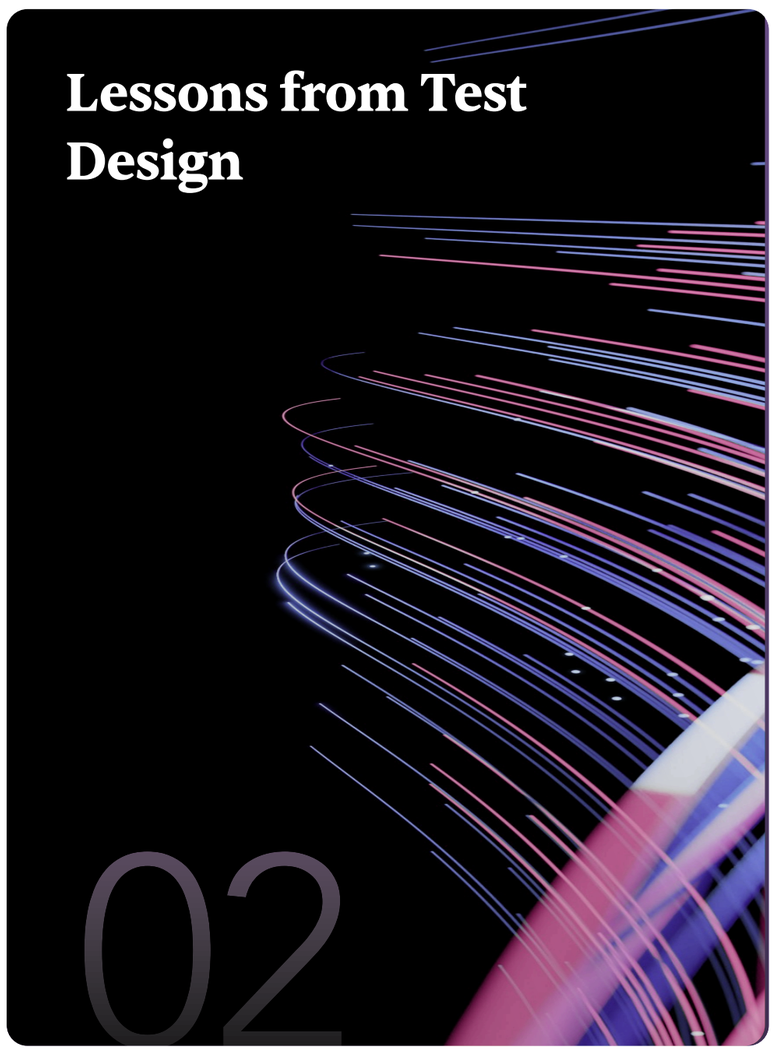
Challenges in Implementation

- **API Performance:** Under real-word conditions, the API responses exhibit latency issues, averaging approximately 1,000 translations per hour, which indicates safeguarding of API abuse but poses a potential performance bottleneck for testing.
- **Stability Issue:** After prolonged test runs, instabilities were encountered indicating the need to individually monitor third party services.
- **Translation Reliability:** Some translations failed intermittently without a clear pattern, again raising the need to monitor performance of individual software components.

Insights/Lessons Learned

The testing pilot reinforced that technical controls should be implemented based on a structured and comprehensive risk profile of the AI use case. Testing can then be used to confirm the qualitative estimation of risk quantitatively.

Structured risk profiling also helps with communication among the partners. For example, the term “Safety” commonly refers to physical injuries or equipment damage. However, in the AI community it is often ambiguously used to also include ethical concerns such as discrimination, insults or other harmful content.



Designing effective multilingual accuracy and content tests presented several key challenges:

- A major limitation was the lack of standardised translation ground truth (especially for less-resourced language pairs and domain-specific technical terms) which forced evaluators to assume the reference translations were accurate, despite potential inconsistencies.
- Additionally, testing for content risks across languages revealed that automated benchmarks alone could miss nuanced or implicit harms, reinforcing the need for human insight during test design.
- Classic metrics like BLEU were selected for their ease of implementation, but it became evident that they often failed to reflect semantic correctness in real-world scenarios. For instance, a translation might score low on BLEU yet be contextually and semantically valid. This highlighted the importance of complementing surface-level metrics with embedding-based similarity and selective human validation.

In practice, executing large-scale multilingual tests through API calls introduced operational challenges. The LLM API responses were occasionally too long or returned in malformed JSON, requiring error handling and reprocessing. Furthermore, non-English input handling varied, sometimes causing empty or inconsistent outputs. These issues complicated automation and underlined the fragility of relying solely on structured API assumptions.