INFOCOMM MEDIA DEVELOPMENT AUTHORITY

AI VERIFY FOUNDATION

Global AI Assurance Sandbox

Overview of the Sandbox

What

A testing ground for builders or deployers of GenAI applications – not the underlying foundation models - to get them tested by specialist technical testers

Why

Reduce testing-related barriers to GenAI adoption – through
- Practical guidance
- Access to specialist testing partners

Provide inputs into (eventual) technical testing standards for GenAI applications

Support the growth of a viable AI assurance market

Who

Organisations that
- Have built or deployed a GenAI application
- Provide specialist software and/or services for technical testing of GenAI applications

Who should join, and why?

Builder/Deployer of GenAI application

You are launching or scaling up a GenAI application, and are looking for:

- Guidance on the appropriate testing to build trust in the application *(what to test)*
- Guidance on *how* to conduct those tests
- Introduction to potential partners that have relevant experience in such testing
- Opportunity to showcase your effort
- (Limited funding to access specialist expertise)

Specialist Technical Testing Vendor

You are building a business (software/service) around technical testing of GenAI applications, and are seeking an opportunity to:

- Validate your testing product/methodology with a real-life use case
- Get introduced to potential customers
- Contribute to emerging standards in GenAI technical testing
- Showcase your capabilities

Notes

- Execution of tests will always be by the testing partner, not IMDA/AIVF
- Testing is expected to be conducted inside the environment of the builder/deployer or the tester *(the sandbox will not provide a dedicated software environment for such testing)*
- Testing results will remain confidential to the builder/deployer and tester

INFOCOMM MEDIA DEVELOPMENT AUTHORITY

What are the qualifying criteria?

For Application to be Tested

- Involves use of a large language or multi-modal model*
- Live in production or intended to be live *(not purely experimental)*
- Focus on technical testing *(not process governance)* of the application *(not the underlying foundation model)*
- Makes a net new contribution to the AIVF/IMDA "body of knowledge"

(As of July 2025: this might involve a GenAI application archetype like Agentic AI, risk types like inappropriate data disclosure, industries outside those covered in the pilot, level of automation/ scale of testing not attempted in pilot)

For Technical Testing Partner

- Offers AI testing as part of product and/or service
- Demonstrates technical expertise in designing and scaling AI testing (e.g., benchmarking, red-teaming, automated evaluators, automated test data generation, human calibration)
- Able to distinguish between testing of underlying foundation model and the GenAI application

** Exception: video/image/voice applications using pre-LLM/LMM technology*

INFOCOMM MEDIA DEVELOPMENT AUTHORITY

Risk dimensions to consider during testing (not exhaustive)

Based on IMDA's Starter Kit

1. Hallucination
2. Undesirable Content
3. Data Disclosure
4. Vulnerability to Adversarial Prompts

Other Use Case Specific Considerations

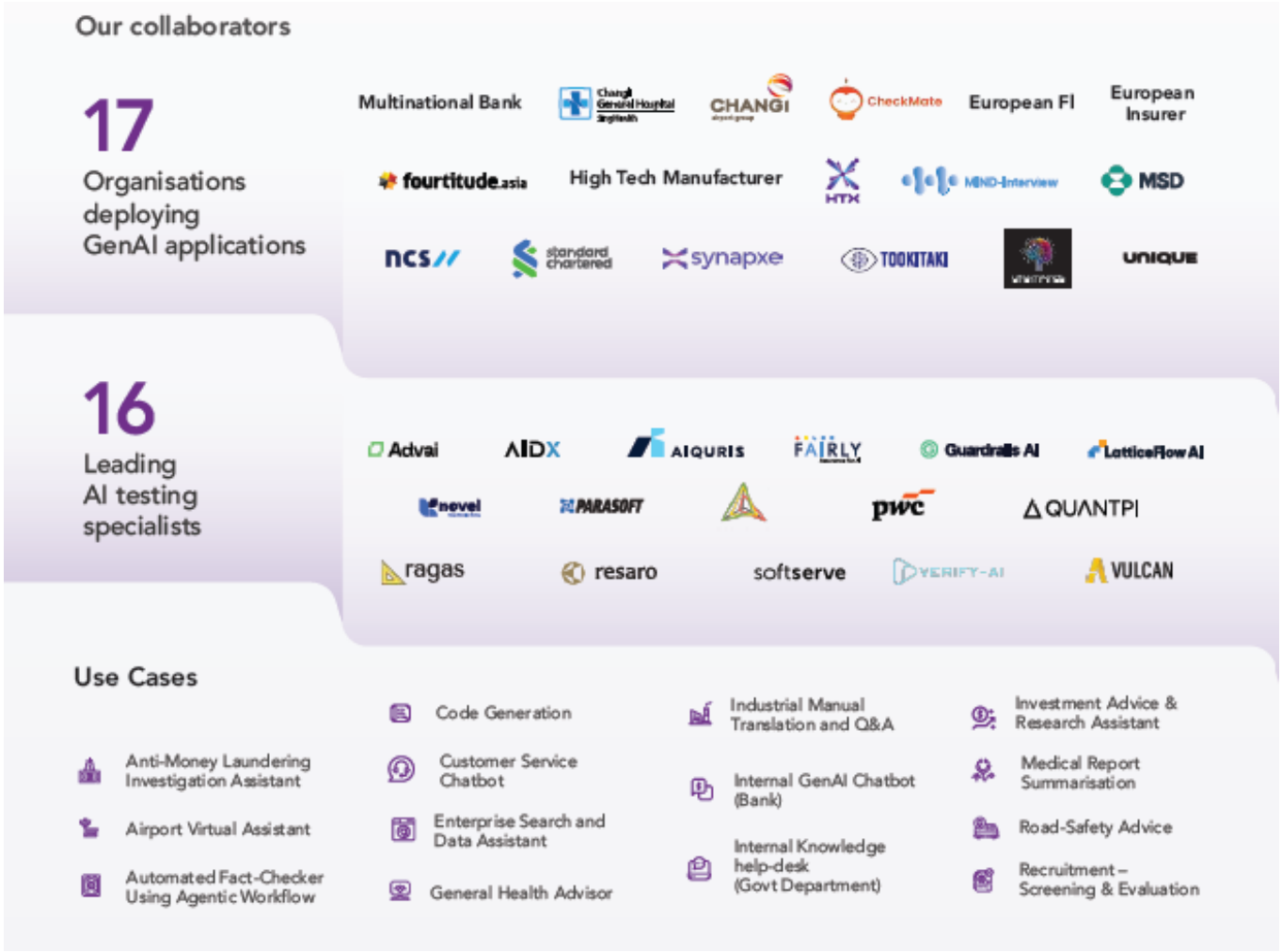
1. Impact on Safety & Health, Financial Concerns, Trust/Reputation Concerns, Unfair Treatment of Employees/ Customers/Users
2. Lack of appropriate level of human oversight/recourse
3. (Non-AI) Breach of industry-specific regulatory requirements
4. (Non-AI) Breach of internal compliance requirements

The Sandbox builds upon a successful pilot (1/2)

In February 2025, AIVF/IMDA launched the Global AI Assurance Pilot for testing of GenAI applications. The goal was to provide industry an opportunity to shape good practices in this rapidly evolving space.

By May 2025, the AI Verify Foundation and IMDA had paired 17 AI deployers with 16 specialist technical testers from around the world (refer to report and case studies for details)

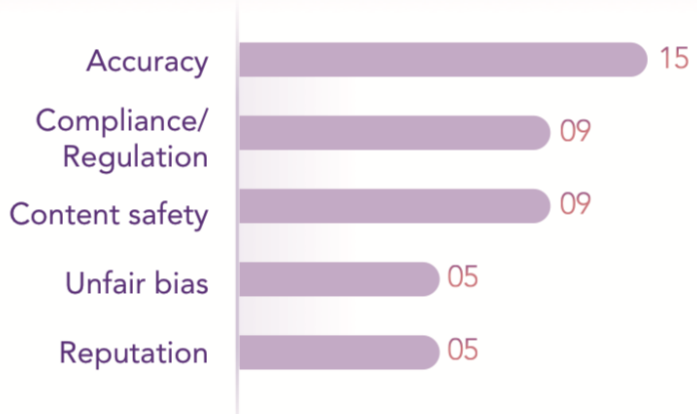
https://assurance.aiverifyfoundation.sg/pilot-overview/



INFOCOMM MEDIA DEVELOPMENT AUTHORITY

The Sandbox builds upon a successful pilot (2/2)

Top 5 Risks Tested



Testing Approaches

Techniques

- Use-case specific historical or synthetic test data
- Red-teaming (adversarial)
- Simulation testing (non-adversarial)
- Off-the-shelf benchmarks

Evaluators

- Human expert
- LLM as a judge
- Non-LLM model
- Rule-based logic
- Semantic or Similarity metrics

Insights

01 **Test what matters**

Your context will determine what risks you should (and shouldn't!) care about. Spend time upfront to design effective tests for those

02 **Don't expect test data to be fit for purpose**

No one has the "right" test dataset to hand. Human and AI effort is needed to generate realistic, adversarial and edge case test data

03 **Look under the hood**

Testing just the outputs may not be enough. Interim touchpoints in the application pipeline can help with debugging and increase confidence

04 **Use LLMs as judges, but with skill and caution**

Human-only evals don't scale. LLMs-as- judges are often necessary, but need careful design and human calibration. Cheaper, faster alternatives exist in some situations

The role of the human expert is still paramount for effective testing!

The IMDA starter kit will act as “baseline guidance” for conducting the testing

Set of **voluntary** guidelines to **coalesce rapidly emerging best practices and methodologies** for app testing

- **Practical step-by-step reference** for how to think about and conduct testing for common risks
- Provide consistency around app testing, **codifying soft standards** and **strengthening end-user trust**

Testing Guidance		Testing Tools & Resources
• Enables app developers to **validate overall safety of app** • Recommends tests and methods	+	• Complemented with **actual testing tools** (in Project Moonshot) to help conduct these tests

INFOCOMM MEDIA DEVELOPMENT AUTHORITY

Expected output from participants

Beyond the actual testing exercise, participants would also be expected to make limited details (not actual test results) public in a pre-defined format

1. **Industry and use case**
2. **Technical implementation:**
   a. High level architecture (e.g. RAG-based)
   b. Details of foundation model - embedding approach and vector database, whether fine tuning was used, data source
3. **Risk considerations**
4. **Test design:** Set of technical tests to support assessment of the risk considerations
5. **Test implementation**:
   a. **Methodology**
   b. Testing **tools**
   c. **Metrics** and/or evaluators within those tools
   d. **Interpretation/thresholds** - calibration used to determine “good/bad” results
6. **Practical challenges and mitigation**
   a. Difficulties faced (e.g., generating adequate test data, interpretation of results, accessibility, cost) and approaches to mitigate them
7. **Resourcing and effort from all parties**

Refer to the Pilot Case Studies published:
https://assurance.aiverifyfoundation.sg/case-studies/

INFOCOMM MEDIA DEVELOPMENT AUTHORITY

FAQ (1/3)

1. Will the AIVF/IMDA help me find a partner?

The AIVF will attempt to match interested testing specialists with firms that are deploying GenAI applications and are interested in exploring external assurance.

2. Will the testing results be shared with IMDA/AIVF or any other regulator?

No, unless explicitly requested by builder/deployer or tester, and agreed with IMDA/AIVF or other regulator.

3. Is it mandatory for the application to include a GenAI element?

Yes, since the focus is on technical testing of LLM applications. However, an exception to this rule may be considered if the application involves processing of image or video data using pre-LLM/LMM techniques.

FAQ (2/3)

4. Is it mandatory to conduct tests as per the IMDA starter kit?

The starter kit is meant to be just that - a starting point. However, as part of the testing exercise, IMDA/AIVF would like to get feedback (and rationale) on the aspects of the starter kit that were not considered relevant for the testing exercise.

5. Is it mandatory to use the open source tools provided by IMDA/AIVF? (E.g., Project Moonshot)

No. However, where participants are conducting testing in areas that are supported in the AIVF open source tools (e.g., selected external benchmarks), IMDA/AIVF would like feedback on the rationale for not using the AIVF-provided tools.

6. Will the testing be free of cost?

IMDA/AIVF will not charge for any guidance or introductions they may provide. However, the testing specialist may seek to use the sandbox to drive commercial outcomes, particularly if they have previously conducted a testing exercise for free in the pilot or sandbox.

FAQ (3/3)

7. Will IMDA/AIVF pay for such costs? If so, how much and in what circumstances?

IMDA/AIVF are working towards partial funding (provided to the builder/deployer) of the fees charged - if any - by their testing partner. Conducting the testing activity in Singapore will be one of the criteria in deciding funding eligibility.

8. Will any regulators be involved in the Sandbox?

In addition to IMDA/PDPC, horizontal or sector specific regulators in Singapore will be provided the opportunity to get involved in the Sandbox, at least as an observer. They may choose to do this to inform their own requirements around AI testing and get real-life feedback on the same.

However, a participant that does *not* want regulatory involvement will be able to avoid the same.

9. Will IMDA/AIVF or any other regulator provide any certification on the applications being tested?

At this stage, regulators are not expected to provide regulatory certainty even if they do get involved.

Next Steps: If you are interested

Our webpage

Email
assurance@aiverify.sg
on your interest

INFOCOMM MEDIA DEVELOPMENT AUTHORITY