Singapore proposes framework to foster trusted Generative AI development

- *Framework furthers international consensus on governance of Generative AI*
- *Builds on previous AI Model Governance Framework*

Singapore, 16 January 2024 – The AI Verify Foundation (AIVF) and Infocomm Media Development Authority (IMDA) have developed a draft **Model AI Governance Framework for Generative AI.** This framework expands on the existing Model Governance Framework that covers Traditional AI, last updated in 2020[1].

2   Generative AI has significant transformative potential – above and beyond what Traditional AI has been able to achieve. This also comes with risks. While it remains a dynamically developing space, there is growing global consensus that consistent principles are needed to create a **trusted environment** – one that enables end-users to use Generative AI confidently and safely, while allowing space for cutting-edge innovation. The use and impact of AI is not limited to individual countries. Hence this proposed framework aims to facilitate international conversations among policymakers, industry, and the research community, to enable trusted development globally.

3   With Generative AI, there is a need to update the earlier model governance framework to holistically address new issues that have emerged. The proposed framework integrates ideas from our earlier discussion paper on Generative AI[2], which put forward a **conceptual foundation**. It also draws on earlier technical work to provide an initial catalogue and **guidance on suggested practices for safety evaluation** of Generative AI models [3]. On top of this, it draws on practical insights from ongoing evaluations tests, conducted within our **Generative AI Evaluation Sandbox**[4].

---

[1] "Model Artificial Intelligence Governance Framework Second Edition" 2020, https://go.gov.sg/AI-gov-MF-2

[2] Discussion Paper published by IMDA and Aicadium in June 2023: Generative AI: Implications for Trust and Governance), https://aiverifyfoundation.sg/downloads/Discussion_Paper.pdf

[3] "Cataloguing LLM Evaluations" published in October 2023 by IMDA and AI Verify Foundation. https://aiverifyfoundation.sg/downloads/Cataloguing_LLM_Evaluations.pdf

[4] Launched by IMDA and AI Verify Foundation in October 2023 to drive the development of evaluation benchmarks in Generative AI, https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/press-releases/2023/generative-ai-evaluation-sandbox

4   The framework looks at nine proposed dimensions to support a comprehensive and trusted AI ecosystem. The core elements are based on the principles that decisions made by AI should be explainable, transparent, and fair. Beyond principles, it offers practical suggestions that model developers and policymakers can apply as initial steps (see Annex for more information).

5   AI governance remains a nascent space. Building international consensus is key, as demonstrated by the successful mapping and interoperability of national AI governance frameworks between Singapore and the US, through the IMDA and the US National Institute of Science and Technology (NIST) crosswalk[5]. The proposed Model Governance Framework for Generative AI takes this one step further by covering the latest developments in Generative AI. In turn, this will inform Singapore's next steps, as we adopt a practical approach to maximise both trust and innovation. This framework will evolve as techniques and technologies develop.

6   For more details, please refer to the proposed draft framework https://aiverifyfoundation.sg/downloads/Proposed_MGF_Gen_AI_2024.pdf. We welcome views from the international community, which can be sent to info@aiverify.sg. This will support finalisation of the Model AI Governance Framework in mid-2024.

---

[5] IMDA-NIST Crosswalk harmonises Singapore's AI Verify and US' AI RMF for international alignment, https://aiverifyfoundation.sg/downloads/AI_RMF_and_AI_Verify_Crosswalk.pdf

About AI Verify Foundation

The AI Verify Foundation (AIVF) harnesses the collective power and contributions of the global open-source community to develop AI testing tools enabling responsible AI. The Foundation promotes best practices and standards for AI and seeks to build trust through ethical AI.

For more information, visit www.aiverifyfoundation.sg.

About Infocomm Media Development Authority

The Infocomm Media Development Authority (IMDA) leads Singapore's digital transformation by developing a vibrant digital economy and an inclusive digital society. As Architects of Singapore's Digital Future, we foster growth in Infocomm Technology and Media sectors in concert with progressive regulations, harnessing frontier technologies, and developing local talent and digital infrastructure ecosystems to establish Singapore as a digital metropolis.

For more news and information, visit www.imda.gov.sg or follow IMDA on Facebook (IMDAsg) and Twitter (@IMDAsg).

_____

For media clarifications, please contact:

Brenda Tan (Ms)

Manager, Communications and Marketing, IMDA

(65) 8180 0228

brenda_tan@imda.gov.sg

Annex: Nine dimensions of the proposed Model AI Governance Framework

This **Model AI Governance Framework for Generative AI** seeks to set forth a systematic and balanced approach to address generative AI concerns while facilitating innovation. It involves key stakeholders, including policymakers, industry and the research community to collectively do their part. Nine dimensions should be looked at in totality, to enable and foster a trusted ecosystem.

Fostering a Trusted AI Ecosystem

1. Accountability
Putting in place the right incentive structure for different players in the AI system development life cycle to be responsible to end-users

2. Data
Ensuring data quality and addressing potentially contentious training data in a pragmatic way, as data is core to model development

3. Trusted Development and Deployment
Enhancing transparency around baseline safety and hygiene measures based on industry best practices in development, evaluation and disclosure

4. Incident Reporting
Implementing an incident management system for timely notification, remediation and continuous improvements, as no AI system is foolproof

5. Testing and Assurance
Providing external validation and added trust through third-party testing, and developing common AI testing standards for consistency

6. Security
Addressing new threat vectors that arise through generative AI models

7. Content Provenance
Transparency about where content comes from as useful signals for end-users

8. Safety and Alignment R&D
Accelerating R&D through global cooperation among AI Safety Institutes to improve model alignment with human intention and values

9. AI for Public Good
Responsible AI includes harnessing AI to benefit the public by democratising access, improving public sector adoption, upskilling workers and developing AI systems sustainably