LEGAL UPDATE

IMDA AND AI VERIFY FOUNDATION FINALISE GOVERNANCE FRAMEWORK FOR GENERATIVE AI AND LAUNCH PROJECT MOONSHOT

1. On 30 May 2024, the Infocomm Media Development Authority ("**IMDA**") and AI Verify Foundation ("**AVF**") finalised the Model AI Governance Framework for Generative AI in Singapore (the "**finalised Gen AI Framework**"). This followed public consultations on the draft Gen AI Framework which was issued on 16 January 2024 and "seeks to set forth a systematic and balanced approach to address generative AI concerns while continuing to facilitate innovation". The public consultation garnered over 70 responses from local and international companies, tech MNCs, audit firms, and Government agencies. A summary and commentary on the draft Gen AI Framework can be found in our earlier article (click here).

2. On 31 May 2024, IMDA and AVF launched Project Moonshot, an open-source Large Language Models ("**LLM**") evaluation toolkit. The toolkit allows businesses to assess their LLM applications against specific benchmarks and manipulation, ensuring that AI applications which are placed on the market can be safe and trusted.

3. This legal update outlines the key changes made to the finalised Gen AI Framework and provides an overview of Project Moonshot.

Revisions from the draft to the finalised Gen AI Framework

4. The finalised Gen AI Framework retains the same 9 dimensions first seen in the draft version: Accountability, Data, Trusted Development and Deployment, Incident Reporting, Testing and Assurance, Security, Content Provenance, Safety and Alignment R&D, and AI for Public Good. However, the finalised Gen AI Framework provides several examples and clarifications to key terms being added:

    a) Data

    As regards the use of copyright materials in training datasets and obtaining consent from copyright owner, the finalised Gen AI Framework clarifies that "*remuneration and licensing to facilitate such uses*" are pertinent issues. Further, debiasing and removing inappropriate content were added as examples of data cleaning under the "Facilitating Access to Quality Data" section.

    b) Trusted Development and Deployment

    It will be recalled that the draft Gen AI Framework discussed the need for safety best practices across the model AI development lifecycle, around development, disclosure and evaluation. The finalised Gen AI Framework adds that the

27 June 2024

For any queries relating to this article, please contact

Tan Tee Jim, *SC*
tanteejim@leenlee.com.sg

Basil Lee
basillee@leenlee.com.sg

Authors:

Tan Tee Jim, *SC*
Basil Lee
Chee Kai Hao
Poon Chong Ming

Lee & Lee
25 North Bridge Road
Level 7
Singapore 179104
Tel: +65 6220 0666

principles articulated in the 2020 version of the Model AI Governance Framework for traditional AI solutions remain relevant for generative AI.

The finalised Gen AI Framework also clarifies that the relevant areas of disclosure (i.e. Data Used, Training Infrastructure, Evaluation Results, Mitigations and Safety Measures, Risks and Limitations, Intended Use and User Data Protection) are only a standard baseline – developers of customised or advanced models should consider disclosing additional information.

Under the section on "Evaluations", models with very niche capabilities have also been stated to potentially require customised evaluations as opposed to using standardised methods.

c) Incident Reporting

In addition to setting up reporting channels for safety vulnerabilities in AI systems, the finalised Gen AI Framework states that this should be complemented by ongoing monitoring efforts to detect malfunctions before they are noticed by end-users.

d) Testing and Assurance

The finalised Gen AI Framework clarifies that "How to Test" also includes specifying the scope of testing to complement internal testing, on top of defining a reliable and consistent testing methodology.

e) Content Provenance

The finalised Gen AI Framework now further acknowledges that technical solutions alone may not be sufficient, and will likely have to be complemented by enforcement mechanisms.

f) Safety and Alignment Research & Development (R&D)

The finalised Gen AI Framework clarifies that the potential risks include present risks (e.g. bias, hallucination) and future catastrophic risks. Also, references to additional resources (which would be invested to AI research) in the Gen AI Framework are specified to include computational resources and access to models.

g) AI for Public Good

Under the section of Sustainability, the finalised Gen AI Framework now refers to the "resource requirements" instead of "power requirements" of generative AI. Energy and water were added as examples of resource requirements. Examples of generative AI that would leave carbon footprint include model training and inference.

Project Moonshot

5. Project Moonshot is one of the first LLM toolkits in the world to bring both benchmarking and red-teaming together to assist AI developers, compliance teams and AI system owners to evaluate their LLM system. The key features of Project Moonshot are:

a) Ready accessibility

The toolkit is open sourced on GitHub (here), providing ready access for the standardised testing of LLMs from popular model providers such as OpenAI, Anthropic, Together, or HuggingFace. Users simply need to provide their API Key to begin testing. For those working with models hosted on custom servers or developing their own LLM applications, the toolkit facilitates integration through Model Connectors, which are designed for minimal coding, making it easier to set up and deploy.

b) Comprehensive benchmarking

The toolkit offers a diverse set of benchmarks to measure the LLM application's performance in the categories of Capability, Quality, and Trust Safety. These benchmarks include well-recognised standards such as Google's BigBench and HuggingFace's leaderboards, alongside more specialised tests tailored to specific domains, such as Tamil Language and Medical LLM benchmarks. A full list of the tests has been set out on GitHub (here).

c) Advanced Red-teaming

The toolkit simplifies the process of red-teaming by providing an easy to use interface which allows for the simultaneous probing of multiple LLM applications, and equips the user with red-teaming tools such as prompt templates, context strategies, and attack modules.

6. Project Moonshot aligns closely with the emphasis under the finalised Gen AI Framework to have robust governance to ensure the responsible development and deployment of AI technologies. Under the "Trusted Development and Deployment" dimension, the finalised Gen AI Framework indicates that the two main approaches to evaluating generative AI today include benchmarking (testing models against datasets of questions and answers to assess performance and safety), and red teaming (where a red team acts as an adversarial user to "break" the model and include safety, security and other violations). Developers of LLMs are encouraged to make use of the toolkit for standardised testing under Project Moonshot.

Conclusion

7. The launch of the finalised Gen AI Framework, together with Project Moonshot, represents a significant development in Singapore's commitment to foster a responsible and secure AI ecosystem.

8. While the finalised Gen AI Framework is non-binding (as opposed to formally enacted laws or regulations), there is no doubt that regulators would take guidance from the framework to craft future policies and regulations. Coupled with Project Moonshot, which offers a robust toolkit to allow stakeholders to rigorously assess their AI applications, it reinforces Singapore's commitment to robust AI governance.

9. If you have any question on any aspect of the Gen AI Framework or Project Moonshot, please contact our Mr. Tan Tee Jim, SC (tanteejim@leenlee.com.sg) or Mr. Basil Lee (basillee@leenlee.com.sg).

About Lee & Lee

Lee & Lee is one of Singapore's leading law firms being continuously rated over the years amongst the top law firms in Singapore. Lee & Lee remains committed to serving its clients' best interests, and continuing its tradition of excellence and integrity. The firm provides a comprehensive range of legal services to serve the differing needs of corporates, financial institutions and individuals. For more information: visit www.leenlee.com.sg.

The following partners lead our departments:

Kwa Kim Li Managing Partner kwakimli@leenlee.com.sg	Quek Mong Hua Litigation & Dispute Resolution quekmonghua@leenlee.com.sg	Owyong Thian Soo Real Estate owyongthiansoo@leenlee.com.sg
Tan Tee Jim, *SC* Intellectual Property tanteejim@leenlee.com.sg	Adrian Chan Corporate adrianchan@leenlee.com.sg	Louise Tan Banking louisetan@leenlee.com.sg