EXPANDED ASEAN GUIDE ON

AI GOVERNANCE AND ETHICS – GENERATIVE AI

Executive Summary

What this Expanded Guide is for

This document supplements and supports the ASEAN Guide on AI Governance and Ethics (2024) with policy considerations related to generative AI (Gen AI). It provides a view of the opportunities and risks of Gen AI and recommends a range of policy actions for ASEAN to support its responsible adoption. These recommendations emphasise the importance of promoting the numerous benefits of Gen AI alongside thoughtful, proportional, and regionally interoperable measures that ensure its safety.

Potential Risks of Gen AI

This Guide aims to provide guidance on addressing the six Gen AI risks identified in the ASEAN AI Guide (2024):

- Mistakes and anthropomorphism,
- Factually inaccurate responses and disinformation,
- Deepfakes, impersonation, fraudulent and malicious activities,
- Infringement of intellectual property rights,
- Privacy and confidentiality,
- Propagation of embedded biases.

This Guide also explores frontier and systemic risks posed by long-term evolution in the capabilities of highly advanced Gen AI models, but which are not widespread in ASEAN at this time.

Policy Recommendations for Addressing Gen AI Risks

This Guide defines a range of policy recommendations for ASEAN to consider for promoting the trusted and responsible use of Gen AI in the region while addressing the risks mentioned above. These policy recommendations draw on global leading practices and emphasise the importance of a coordinated, pro-innovation regional response to this new technology.

1. Accountability
2. Data
3. Trusted Development and Deployment
4. Incident Reporting
5. Testing and Assurance
6. Security
7. Content Provenance
8. Safety and Alignment Research & Development
9. AI for Public Good

Use Cases

This Guide illustrates the implementation of its policy recommendations with four detailed use cases. These feature public and private institutions in ASEAN that have taken steps to implement practices aligned with AI governance and ethics.

PhoGPT,
VinAI (Vietnam)

Project Moonshot,
the AI Verify Foundation
(Singapore)

Responsible AI
Internal Programme,
Accenture (ASEAN-wide)

ThaiLLM, BDI, NSTDA,
VISTEC and collaborators
(Thailand)

Contents

01

Introduction

01 Introduction

1.1 Context

A major advancement occurred in the field of Artificial Intelligence (AI) in the mid-2010's with the development of the first Generative AI ("Gen AI") systems, which instead of the descriptions or predictions of earlier AI systems could draw on a large dataset to create "content". This content could take the form of text, images, audio, videos, computer code, and a wide variety of other materials.

Fundamentals of Gen AI

Traditional AI, sometimes called predictive or diagnostic AI, is trained on data to complete tasks like classifying data or providing recommendations. Gen AI is a family of technologies that is distinct because it uses some of these same fundamental machine learning approaches, in combination with a large dataset, to **produce content.** It is a recent innovation in AI technology that builds on the past several decades of innovation in **machine learning** and **deep learning.** The outputs of modern Gen AI can sometimes resemble authentic content, such as text that appears to be written by a human, images that resemble real photographs, or audio that resembles human speech. It typically does so in response to user instructions, in the form of a short written statement called a **prompt.**

The core technologies powering most Gen AI are called **foundation models.** These are probabilistic statistical models that are trained on content and designed to generate it. This training is typically supplemented with techniques that further adjust foundation models to improve quality or safety. This can be done by third parties. One or more foundation models working together, combined with the enabling infrastructure, user interface, additional safety measures, and application layer that modifies how they are to be used, is called a **Gen AI system.**

While Gen AI systems can sometimes imitate human behaviours, they are not creative, self-aware, or intelligent. Instead, they are powerful tools for inferring statistical trends from a dataset and applying those trends to produce new content that resembles its training dataset. Some Gen AI systems continue to update their datasets and improve over time or evolve in response to their use.

Gen AI shares many of the same characteristics and challenges of advanced forms of traditional AI, such as having outputs that are hard to explain or having the potential to inadvertently learn undesirable behaviours from training data.

Gen AI represents a leap forward in AI capabilities and an enormous economic and social opportunity for ASEAN member states, with its total economic opportunity across the greater APAC region estimated to reach nearly S$6 trillion through 2038.[1] Gen AI can automate repetitive tasks, accelerate creative endeavours, and improve the personalisation and responsiveness of public services. Used correctly, it can improve experiences for consumers, employees, and citizens, and can boost living standards and contribute to solving key regional challenges. The capability of some, but not all, Gen AI systems to interact with user prompts written in "natural language" – the usual language and phrasing that people use – rather than computer code also presents an opportunity to democratise access to AI for users and small businesses without the need for extensive technical knowledge. The long-term potential of Gen AI is only beginning to be realised.

Gen AI, with its ability to create new content, also raises challenges that policymakers should look to address. The ASEAN Guide on AI Governance and Ethics (2024) – henceforth, the ASEAN AI Guide (2024) – highlighted six risks of Gen AI for this expanded Guide to explore:

01

Mistakes and anthropomorphism

02

Factually inaccurate responses and disinformation

03

Deepfakes, impersonation, fraudulent and malicious activities

04

Infringement of intellectual property rights

05

Privacy and confidentiality

06

Propagation of embedded biases

These risks, along with long-term frontier or systemic risks posed by the development of highly advanced Gen AI models, are discussed in this expanded Guide. These risks are key considerations informing several policy recommendations designed to mitigate them and to promote the positive use of Gen AI.

Policymakers have an important role to play in the emerging Gen AI economy. The widespread and growing use of Gen AI by both companies and individuals has magnified its potential to impact these users both positively and negatively, and the range of potential impacts, illustrated above, extends across traditional policy domains. Policymakers can help shape the ongoing adoption and use of Gen AI such that they encourage innovation and maximise its positive economic and personal benefits while enabling developers, deployers, and users to adopt responsible practices that address ethical and legal risks. It is only with thoughtful and proactive engagement between the public and private sectors that the full potential of Gen AI will be realised.

The ASEAN AI Guide (2024) highlighted the importance of several measures, focused on trusted development and deployment, for addressing Gen AI risks; this expanded Guide adopts these suggestions and supplements them with a holistic approach rooted in the goal of establishing a trusted Gen AI ecosystem.

This expanded Guide recommends an ASEAN-wide approach to managing Gen AI risks and promoting Gen AI opportunity that strikes a fair balance between the significance of economic growth, innovation, safety, and regional harmonisation.

This expanded Guide is meant to be used in conjunction with the ASEAN AI Guide (2024). Unless otherwise specified, the definitions, recommendations, and practices set out in that document continue to apply here.[*]

1.2 Objectives and Target Audience of this Expanded Guide

This Guide is designed as a resource for ASEAN policymakers to understand the primary challenges that they might encounter when working on subjects related to Gen AI, as well as potential strategies for addressing them.

This Guide summarises the challenges and opportunities of Gen AI for policymakers in nine ecosystem dimensions:

1. Accountability
2. Data
3. Trusted Development and Deployment
4. Incident Reporting
5. Testing and Assurance
6. Security
7. Content Provenance
8. Safety and Alignment Research & Development
9. AI for Public Good

Gen AI policy recommendations in Section 3 are organised corresponding to these nine dimensions. Each dimension is defined in detail there.

[*] Refer to the ASEAN AI Guide (2024) at: https://asean.org/book/asean-guide-on-ai-governance-and-ethics/

1.3 Guiding Principles for the Gen AI Framework

The ASEAN AI Guide (2024) set out seven guiding principles for fostering trust in AI and ensuring the ethical design, development, and deployment of AI systems. These principles continue to be relevant; however, Gen AI introduces additional considerations.

- **Transparency and Explainability:**

  Transparency in AI involves clear disclosure of its usage, how it is involved in decision-making, the kind of data used, and purpose of its use. Understanding whether AI is used in turn enables users to make informed decisions about whether to use that system. **Explainability** is the ability to articulate AI's decision-making process. Proportionate explainability techniques or effective substitutes, such as well-documented repeatability or traceability, can allow users to understand the factors contributing to an AI's output.[†]

  Notwithstanding that some advanced traditional AI models can also be difficult to explain, Gen AI systems may be more complex, with behaviours and outputs that can be unclear or challenging to explain (e.g. so-called "black box" algorithms).

  It is important to build public trust in the use of Gen AI by ensuring that users are aware that they are interacting with Gen AI and of how data is being used. Developers and deployers should consider what information may be useful to share with users to enable them to make informed decisions about their interactions with Gen AI, bearing in mind that the same transparency considerations applicable to traditional AI in the ASEAN AI Guide (2024) remain relevant to this technology. Notably, such disclosure will also need to be balanced against other competing considerations, such as the need to protect proprietary or commercially sensitive information. Developers and deployers can consider options for sharing this information in ways that are useful and understandable to their users while at the same time protecting their own legitimate interests.

  Several explainability techniques are being developed to address these challenges, and post-hoc explanation methods, which analyse models after they have generated their output, are becoming increasingly sophisticated.

  Despite these advancements, providing explanations for AI model behaviours that are useful to users remains difficult. This is especially true for Gen AI models, whose

[†] ASEAN Guide on AI Governance and Ethics (2024), p. 11-12.

outputs are often more complex than traditional AI outputs and are likely to be used in a wider range of contexts. This richness, combined with the inherent complexity of the neural networks used in Gen AI, makes providing satisfying explanations more challenging. While organisations should ideally aim to develop or choose Gen AI models that are more explainable, this may not always be possible. Other options, as discussed in the ASEAN AI Guide (2024), could include improving traceability. However, this may also have its limits given the inherent nature of Gen AI.

Ongoing research is crucial to develop more robust and comprehensive tools for transparency, explainability, and traceability.

- **Fairness and Equity:**

  Fairness and equity in AI focuses on having safeguards in place to ensure that AI does not exacerbate or amplify existing discriminatory or unjust impacts across different demographics. The design, development and deployment of AI systems should not result in unfair bias or discrimination.[*]

  Foundation models underpinning Gen AI systems are often trained on large swathes of internet-sourced data, which may reflect existing social biases. There are also concerns over whether such data is fully representative of different perspectives or cultures due to the potential for oversampling well-connected populations in common languages in internet-sourced data. This may result in the underrepresentation or overrepresentation of certain groups in its output.

  Gen AI, if not managed well, can magnify societal biases through a variety of mechanisms, including by “learning” biases in training datasets, receiving biased feedback during fine-tuning, or incorporating biases from user inputs during inference. Organisations are encouraged to pay attention to all sources of bias, including data quality, the representativeness of the data used to train foundation models, bias in fine-tuning, and the impact of a biased organisational context. They should be attentive to the impact and representativeness of Gen AI outputs across demographics and support these findings through testing and evaluation.[^]

[*] ASEAN Guide on AI Governance and Ethics (2024), p. 12-13.

[^] A variety of techniques exist amid an emerging state of the art. A common metric for evaluating LLM fairness are BOLD (Bias in Open-Ended Language Generation Dataset) statistics, which contains a standard set of 23,679 internet-sourced prompts related to profession, gender, race, religious belief, and political ideology, and measures the sentiment of an LLM’s response to each prompt. RealToxicityPrompts is another widely used library containing over 100 thousand prompts, which when paired with a common detector of toxicity in text, can be used to benchmark the toxicity of LLM behaviour in response to those prompts. These techniques continue to evolve.

- **Security and Safety:**

  Safety in AI involves risk assessments, as well as mitigation for developers, deployers and users, to reduce risks to an acceptable level (the details of which may be determined by deployers or developers, bearing in mind their use case and context). **Security** in AI is about ensuring the confidentiality, integrity and availability of AI systems, including against malicious attacks.[†]

  Organisations should take a risk-based approach to managing AI safety and security. Many in the Gen AI industry mitigate risks using technical methods designed to train generative models to minimise outputs likely to be unsafe. These methods include fine-tuning techniques like Reinforcement Learning from Human Feedback[2] (RLHF) and Constitutional AI,[3] which uses AI models and human-supplied evaluation criteria to produce feedback for Gen AI systems. Current methods of evaluating Gen AI models for safety can involve red teaming[4] to expose unsafe behaviours. However, further research is needed to establish and validate reliable common practices for evaluating and mitigating the full range of safety concerns, particularly given the sometimes-unpredictable nature of Gen AI outputs.

  To enhance security, stakeholders should stress the significance of secure data collection and storage at the application level, and AI developers should adopt measures to safeguard any retained copyrighted works used in training Gen AI systems and put in place a response plan for data breaches. They should evaluate the security risks tied to Gen AI and consider risks associated with “shadow IT”[5]. Measures should be taken to guard against “prompt injection”[6], “data poisoning”[7], and other kinds of Gen AI-specific attacks. A useful reference for a taxonomy of relevant attacks is NIST’s AI 100-2 E2023.[8]

- **Human-centricity:**

  Human-centricity in AI focuses on creating systems to benefit society and improve individual well-being, enhancing life quality while preventing harm.[*]

  Gen AI has the potential to impact human lives and livelihoods both positively and negatively. Maximising the positive impacts and minimising the negative ones is not an automatic outcome of developing the technology; it will depend on conscious effort by AI actors. It is therefore crucial to prioritise human interests and values in the design, development, deployment, and operation of Gen AI.

[†] ASEAN Guide on AI Governance and Ethics (2024), p.13.
[*] Ibid, p. 14.

The absence of a human-centric approach can lead to employment disruptions, reduced social benefits, and various forms of exclusion, which underscores the importance of emphasising human values, dignity, and welfare in the development and deployment of AI systems.

Human-centric development is essential because Gen AI not only has the capacity to generate content, but can also potentially imitate human innovation. Gen AI should be used to enhance human creativity and promote collaboration between humans and AI, thereby expanding the creative potential of individuals instead of reducing it.

- **Privacy and Data Governance:**

  Privacy in AI safeguards personal data from unauthorised access and misuse. **Data governance** in AI includes the proper management (collection, use and disposal) of data used to train foundation models or as input in Gen AI systems. This would include measures to ensure personal data privacy and protection, or to protect the quality and integrity of data throughout their life cycle.[†]

  Gen AI has raised privacy, intellectual property and personal data protection issues due to the practice of gathering, or “scraping”, training data from the internet, which may include personal data. Questions have been raised over the proper legal basis for processing such personal data under applicable personal data protection laws.

  The ability for Gen AI to “memorise” and release training data as output increases the risks of accidental disclosure of personal data to other users. Malicious actors may intentionally attempt to induce models to disclose sensitive personal information from their training datasets. They may also use Gen AI systems to create inauthentic images, videos, and text that violate an individual’s privacy.

  In line with this principle, organisations should follow a privacy-by-design methodology. This includes a range of practices, such as minimising data collection to what is necessary, data anonymisation where doing so does not unjustifiably jeopardise model effectiveness, introducing safeguards within the model to refuse to disclose or share personal data, and ensuring appropriate grounds or a legal basis to process personal data, such as securing user consent, in line with data protection laws where they apply. When using data and deploying privacy-preserving practices, organisations should consider the industry consensus across jurisdictions.

[†] ASEAN Guide on AI Governance and Ethics (2024), pp.14-15.

To facilitate proper data governance, it is important that organisations put in place measures where possible to establish data traceability such as having broad awareness over the sources of and types of data used to train Gen AI models, and how such data may have been curated or moved across the organisation. Organisations should be able to internally verify the accuracy and integrity of the data they use, as well as their compliance with applicable privacy regulations. Organisations should also be transparent, within reasonable commercial bounds, as to the types of training data sources and how that data was processed before training. Such transparency will need to be balanced against the need to protect commercially sensitive or proprietary information.

- **Accountability and Integrity:**

  There needs to be human **accountability** and control in the design, development and deployment of Gen AI systems. This includes having clear internal governance mechanisms for proper management oversight of the Gen AI system's development and deployment.*

  Notably, the development and deployment of Gen AI involves multiple layers in the tech stack. Control over the Gen AI system may also shift between deployers, developers and users across the Gen AI development and deployment chain. This has an impact over how responsibility can be shared between different stakeholders, and the role that each stakeholder has to play in governance mechanisms for proper management oversight.

  As is often the case with innovative systems, their outcomes in a real-world context can differ from the original intentions of their creators, highlighting the importance of dynamic testing to address unforeseen behaviours. This also highlights the importance of clear, adaptable terms and conditions to manage accountability as systems evolve in ways that may not have been fully anticipated during development.

*ASEAN Guide on AI Governance and Ethics (2024), pp. 15-16.

- **Robustness and Reliability:**

  Gen AI systems should be sufficiently **robust** to cope with errors during execution, unexpected or erroneous input, or potential changes in their operating environment that may interact with the Gen AI system in an adversarial manner. Where possible, Gen AI systems should also work **reliably** and have consistent results for a range of inputs and situations.[†]

  However, in some cases Gen AI systems can produce inconsistent results due to factors like biases in training data, data quality issues, or model architecture. This makes it challenging to assess their output for consistency or for system robustness.

  Bearing this in mind, developers and deployers should aim to create measurable quantitative benchmarks and/or standards for acceptable outcomes, acknowledging that the behaviour of Gen AI systems will not be deterministic but will instead fall within a certain range. Developers and deployers should both consider implementing guardrails to prevent their Gen AI systems from responding to out-of-context or erroneous prompts that address subjects outside their intended or designed operation.

[†] ASEAN Guide on AI Governance and Ethics (2024), p.16.

02

Gen AI Risks

02 Gen AI Risks

2.1 New or Enhanced Risks

Gen AI, in addition to the opportunities it presents, also carries risks that may require new approaches to governance to address. The development, deployment, and use of Gen AI systems can raise ethical, legal, and societal issues that may be new, representing risks that are not typically associated with traditional AI, or enhanced, representing risks associated with traditional AI whose impacts or mitigations may change in the context of Gen AI.

Gen AI risks can be mitigated by system developers, deployers, and users, as well as through policy actions that support good AI governance. Many Gen AI risks are already being addressed effectively by emerging norms and practices in industry and society. This guide's policy recommendations are designed to support stakeholders in addressing the risks below.

The six unique Gen AI risks described by the ASEAN AI Guide (2024) are the basis for this paper:

01 **Mistakes and anthropomorphism:**
Gen AI systems can make highly coherent and persuasive mistakes, often referred to as "hallucinations," such as providing incorrect medical advice or generating vulnerable software code.

02 **Factually inaccurate responses and disinformation:**
Gen AI systems can amplify false or misleading information, shaping public perception and eroding trust in reliable sources.

03 **Deepfakes, impersonation, fraudulent and malicious activities:**
Gen AI systems pose risks of impersonation or misinformation by creating realistic content like deepfakes and phishing emails, making it harder to prevent identity theft, identify deception and protect confidential information.

04 **Infringement of intellectual property rights:**
The development and use of Gen AI systems may lead to legal repercussions if copyrighted works are used as data to train the systems without an appropriate legal basis, or if the generated content too closely resembles existing works.

05 **Privacy and confidentiality:**
Gen AI systems can occasionally memorise and reproduce specific training data, or otherwise allow malicious actors to reconstruct sensitive information through their prompts. Employees may also inadvertently disclose confidential data during interactions.

06 **Propagation of embedded biases:**
Gen AI systems can inherit and reflect biases from their training data, leading to biased or toxic outputs that reinforce stereotypes.

2.2 Frontier and Systemic Risks for Future Consideration

Policymakers should also be cognisant of frontier risks alongside the transformative benefits and capabilities of such Gen AI systems, which will prepare them to effectively respond to future technological progress in AI. In addition to the six risks above, which are present-term risks with well-documented occurrence and mitigations, scholars in the field have identified several potential long-term impacts, sometimes called frontier risks. Frontier risks are primarily concerned with the risks surrounding the use, controllability, and value-alignment of highly advanced Gen AI systems.

With regard to Gen AI use, one concern is that dangerous information (such as information related to CBRNE[*] weapons) could, potentially be accessed using Gen AI systems, increasing the capabilities of malicious actors (e.g., to create biological weapons or conduct effective cyber-attacks). There are additional concerns around how Gen AI systems can be used to increase the scale and sophistication of scams and fraud or cause other harms (e.g., generate non-consensual deepfake pornography).

There is also concern around the controllability of AI systems with the advent of Gen AI, with the expert community divided over when and whether AI could become fully autonomous and independent of human constraint. There is broad consensus that current technology does not have the capability to pose such risks. However, there is ongoing scientific debate on how plausible such scenarios are, when they might occur, and how difficult it is to mitigate them.

In the longer term, there is also concern as to the risk of misalignment of AI systems as this technology progresses, where agentic, self-improving AI systems able to work autonomously without human oversight pursue such goals in a way that harms human interests.

In addition, the widespread development and adoption of Gen AI technology poses several long-term systemic risks, ranging from potential labour market impacts to privacy risks and environmental effects. In particular, there is concern over the growing use of computing power in general-purpose AI development and deployment and the corresponding rapid increase in energy usage. This could lead to further increases in CO2 emissions and water consumption, both of which are known to have a negative impact on the environment.

Above all, experts have noted that it is difficult to assess the downstream societal impact of Gen AI; there is currently insufficient research to produce rigorous and comprehensive risk assessment methodologies in this domain. There is need for continued study as the capabilities of Gen AI progress so that policy makers can better understand how to manage the frontier and systemic risks around Gen AI.

[*] CBRNE is an acronym for Chemical, Biological, Radiological, Nuclear, and high yield Explosive

03

Policy Recommendations

03 Policy Recommendations

This section is intended for ASEAN policymakers.

This section provides recommendations to regional policymakers at the ASEAN level. Recommendations to national policymakers are not included in this exercise.

3.1 Accountability

Like most software, Gen AI involves multiple layers in its technology stack, reinforcing the trend towards specialisation and division of labour across the value chain. For example, organisations are likely to choose to partner with a software provider or explore open-source options to access Gen AI models. They may also choose to license whole software systems from their partner organisations, such as chatbots powered by LLMs, or use Gen AI in a software-as-a-service (SaaS) model. Such value chains can sometimes be complex. For example, Gen AI systems can integrate multiple models from diverse providers, such as chatbots powered by LLMs alongside image-generating diffusion models, to create more powerful and versatile solutions.

Given the multiple layers in and complexity of the Gen AI value chain, accountability across the AI ecosystem becomes important. Developers, cloud service providers, deployers, and users have a shared responsibility for driving ethical AI use, ensuring transparency, and building trust. Discussions on how to drive accountability for Gen AI are ongoing and evolving in many countries. As a region and potential common market, ASEAN will benefit from developing a common approach towards accountability for Gen AI.

Useful models to draw lessons from include looking to shared responsibility frameworks in other domains, particularly in cloud computing. Here, a shared responsibility framework demarcates the roles of software or service providers and their customers, and sets out clear service standards for measuring them, such as quantifiable benchmarks. Shared responsibility frameworks are a useful model to reflect the fact that both upstream value chain participants and end deployers have a role to play in establishing accountability over the development and deployment of Gen AI technologies. This also sets expectations on the responsibilities that various players have across the value chain, including users.

Areas for ASEAN to explore:

(a) Developing a common understanding around shared responsibility

ASEAN can encourage and facilitate **collaboration among developers, deployers, regulators, cloud providers, and civil society** to create a common understanding on what shared responsibility would entail. These can include organising physical or virtual forums to develop an understanding of current best practices, risk mitigation strategies, and existing incident reporting frameworks across the value chain relationships in Gen AI development and deployment. Such platforms can also include discussion on emerging trends and new risks, to ensure that policymakers are aware of recent developments. Potential outputs from such forums can include a voluntary framework or set of principles for all stakeholders in the Gen AI value chain, including model creators and cloud providers. It would also be useful to consider the role that end users have to play within this framework. Where relevant and useful, such output should also align with existing international developments and best practices, e.g., the Hiroshima Process International Code of Conduct for Organisations Developing Advanced AI Systems.

3.2 Data

Data is the key building block in all machine learning approaches, enabling models to learn patterns and generate outputs. Gen AI is built on data: by learning from ever-larger datasets of real content, Gen AI has leapt forward in its capabilities.

The data that Gen AI systems learn from is at the core of their value, and the nature of that data presents both an opportunity and a challenge. For ASEAN to reap the benefits of Gen AI, there is a need to ensure access to high-quality data in sufficient volumes. For example, access to data in ASEAN languages can improve the performance of Gen AI for these languages. This in turn will unlock significant economic value in the region and make the benefits of Gen AI more widely accessible.

However, there is also a need to manage the risks surrounding the use of datasets collected on the internet to train Gen AI. The use of such data opens the door to rich and diverse insights but also requires responsible practices to manage considerations like usage rights, data provenance and accuracy, and personal data protection. A balanced approach will help organisations in ASEAN unlock the transformative potential of generative AI while ensuring compliance, creativity, and user trust.

Areas for ASEAN to explore:

(a) Facilitating data sharing to develop ASEAN-relevant models

To improve the representativeness of datasets for the region, ASEAN can work together to promote **the collation of high-quality open datasets and support for industry data sharing.** Access to more high-quality datasets that include both human-generated and synthetic data can unlock value across the Gen AI landscape by including more market participants who may not otherwise have the capabilities to source the required datasets. Publicly shared data, combined with data sharing or sales among industry actors, offers opportunities to level the playing field – especially for smaller firms – and minimise risks like feedback loops. ASEAN Member States can collaborate to share expertise and develop best practices for creating a robust data ecosystem, such as a compendium of machine-readable data sources from across ASEAN and in regionally relevant languages.

BIG DATA INSTITUTE

สวทช
NSTDA

VISTEC
VIDYASIRIMEDHI
INSTITUTE OF SCIENCE AND TECHNOLOGY

BDI, NSTDA, VISTEC and collaborators - ThaiLLM *(see detail in Appendix)*

Thai-language applications of Gen AI have been constrained by the limited performance and cultural sensitivity of leading LLMs developed in Western countries. ThaiLLM is a public sector initiative to train an open-source LLM specifically on a corpus of Thai-language text, with the specific goal of supporting chatbots and other AI applications in Thailand.

ThaiLLM aspires to become a shared national infrastructure, allowing LLM developers to contribute their own Thai-language text datasets under open-source licenses and collaboratively improve the tool's performance. ThaiLLM will empower Thai startups by lowering costs, increasing access to AI capabilities, and reducing reliance on foreign AI solutions, fostering widespread adoption of Gen AI in Thailand.

VinAI - PhoGPT *(see detail in Appendix)*

Vietnamese-language applications of popular LLMs have faced limited performance and a lack of cultural relevance, impeding the adoption of AI there. VinAI's PhoGPT addresses these challenges by having been trained on a dataset of over 102 billion words of Vietnamese text from sources like Wikipedia, books, legal documents, and news articles. This improved performance on Vietnamese-language applications creates opportunities to increase access to AI and has helped encourage the development of AI skills in the country.

As an open-source model, PhoGPT encourages innovation, collaboration, and broader access to Vietnamese-language AI technology for the collective good of the industry. VinAI plans to extend PhoGPT to other underrepresented ASEAN languages.

AI Singapore - SEA-LION

AI Singapore is the country's national-level AI R&D programme which aims to anchor deep national capabilities and catalyse the use of AI across the enterprise economy. It has developed a family of open-source LLMs called SEA-LION (South East Asian Languages in One Network) that better understand Southeast Asia's diverse contexts, languages, and cultures. It was designed to address the unique needs of the region, which are often not fully addressed by leading LLMS developed in Western countries. It did so by significantly expanding the proportion of South-East Asian content compared to leading models, sourced from the internet and contributed directly by their media partners.

The latest version of SEA-LION supports the various national languages of South-East Asian countries including Bahasa Indonesia, Burmese, Chinese, English, Filipino, Khmer, Lao, Malay, Tamil, Thai, and Vietnamese. The model is also expanding its capabilities into major regional dialects, such as Javanese and Sundanese in Indonesia, and Visayan and Illocano in the Philippines. The latest version of the SEA-LION model is fine-tuned for higher performance in Bahasa Indonesia, Thai, and Vietnamese.

A balanced combination of public and private data availability fosters innovation and promotes a healthy, competitive Gen AI environment, especially by facilitating access for startups and researchers who may not otherwise have access to it.

Early forays into this effort have already begun with the development of ThaiLLM, PhoGPT, and SEA-LION (see box stories above). A common repository of ASEAN data can help support the improvement and extension of these models to be even more relevant to ASEAN.

(b) Developing a common approach to personal data protection and data governance in AI

Promoting regional coordination and a **common approach towards the use and sharing of personal data across ASEAN** will help improve access to relevant data that can be used to train Gen AI and foster user trust in such use of their personal data. This includes tools such as the ASEAN Framework on Personal Data Protection. In addition, effort could be made to leverage platforms such as the ASEAN Data Protection and Privacy Forum (ADPPF) to facilitate shared learning and collaboration, with a view to harmonising personal data protection standards. ADPPF serves as an annual forum where national authorities discuss developments in data protection, share knowledge, and implement relevant initiatives under the ASEAN Framework on Digital Data Governance. These collaborative efforts provide a valuable platform for relevant privacy authorities to share information on and address GenAI-related data protection challenges using regional frameworks.

Apart from personal data protection, attention could be paid towards developing regional **guidelines or common approaches towards data handling, storage, and governance for Gen AI** to offer a practical way for organisations and regulators to address challenges around data quality, usage, and compliance. This can be supported by developing clear, measurable regional benchmarks on data quality to enable stakeholders to assess data-related risks and build confidence in their systems. ASEAN can also provide guidance on approaches for organisations to consider when managing cross-border data flows related to AI, respecting the diversity of regulatory environments and data sovereignty rules among ASEAN member states. Where relevant, ASEAN should take reference from existing data governance practices from bodies like ISO, NIST, and the OECD to reduce global fragmentation and support cross-border trade.

3.3 Trusted Development and Deployment

Gen AI systems can consist of one or more foundation models as well as other software components. Its development, deployment, and use are also supported by an overarching governance structure of roles, responsibilities, and oversight within each organisation in the value chain. The effective operation of all these elements together to ensure proper governance or oversight over the full system lifecycle, from design to deployment to decommissioning, is important to build trust in the development and deployment of Gen AI.

This in turn requires the integration of best practices for safety, ethics, and functionality into the governance process. A number of these practices were outlined with traditional AI in mind in the ASEAN AI Guide (2024) and remain relevant to Gen AI development and deployment. However, given the specific risks and concerns around Gen AI, a greater focus on safety best practices, particularly in areas relating to development, disclosure, and evaluation, could improve governance and oversight over the specific challenges around the development and deployment of Gen AI.

In Gen AI development, several safety best practices are commonly accepted. These include techniques such as reinforcement learning, which enhances model quality through human or automated feedback, grounding methods that ensure that outputs remain contextually appropriate, and the tailored design of application components that optimise model performance. These techniques serve as essential "guardrails," providing technical safeguards to ensure that models meet ethical, safety, and functional standards. Additionally, transparency features and user empowerment tools are critical for fostering user trust and confidence in Gen AI systems. Guardrails are usually applied in a risk-sensitive fashion, with riskier Gen AI applications being subject to more guardrails.

Deployment best practices focus on model evaluation. Evaluating Gen AI systems can involve comparing their performance to standard industry benchmarks or engaging in adversarial testing, commonly known as red teaming. A variety of other techniques exist, such as RAGAs evaluations for systems based on retrieval-augmented generation. These practices all aim to identify whether the system meets their desired level of safety before being put into production. Model evaluation at the point of deployment is particularly useful when a Gen AI model or system has been procured from a third party.

Other deployment best practices include technical guardrails that do not require changes to the underlying model, such as the implementation of input/output filtering, user training, human moderation, or continuous system monitoring.

Notably, it is also important to have transparency around safety best practices. This is akin to "food or ingredient labels" by providing relevant information to downstream deployers and end users, so that they can make informed decisions about the deployment or use of the Gen AI model or system. The level of disclosure should take into account the need to protect proprietary information.

In this regard, there is benefit to ASEAN coming together as a region to align on its approach. This will help ASEAN member states promote a cohesive and interoperable approach to the development and deployment of Gen AI that supports the flow of Gen AI-powered goods and service across the region.

Areas for ASEAN to explore:

(a) Establishing guidelines for the development and deployment of Gen AI models and/or applications

Organisations developing and deploying Gen AI bear the primary responsibility for managing their risks and can benefit significantly from specific technical and operational guidance tailored to Gen AI systems. Guidance at the ASEAN level on safety practices for Gen AI can leverage regional expertise and assist organisations in navigating these responsibilities. Areas that can be covered by the **guidance include safety best practices across the model lifecycle, from development through deployment.**

When developing guidelines, ASEAN should remain aware that work is also being done at the global level. Bearing in mind its existence as an active part of the wider global economy, ASEAN should prioritise alignment and **interoperability with the work of bodies like ISO, NIST, and the OECD to reduce fragmentation and support cross-border trade.** In addition, bearing in mind the wide variety of use cases for Gen AI, these guidelines should function as a horizontal baseline that can be adapted to sector-specific considerations.

(b) Coalescing around common disclosure elements

The increasing prevalence of third-party Gen AI models and systems, and especially the use of proprietary or open-source models by organisations in ASEAN, **underscores the critical need for upstream developers to share important information with their downstream deployers.** A careful balance in disclosure is required to consider both the need of deployers to understand the tools that they are using and the need for developers to protect trade secrets and intellectual property. In this regard, ASEAN can also look at guidelines around common disclosure elements for generative AI models and/or applications.

accenture

Accenture Responsible AI Internal Programme *(see detail in Appendix)*

Accenture deploys a range of Gen AI systems in ASEAN countries, both internally in the firm and on behalf of its clients. Effective governance measures are crucial for ensuring that the firm's use of Gen AI adheres to its principles and to its technical and operational guidelines.

Accenture's Responsible AI Internal Programme facilitates the organisation's responsible use of Gen AI. It started by building a governance foundation, which formalised C-Suite sponsorship of the organisation's responsible AI programme, implemented necessary oversight, principles, policies, and standards, and assembled a multi-disciplinary team to own the programme.

There are three elements to delivering the programme:

- Conducting RAI Risk Assessments for each use case,
- Systemic Enablement for RAI Testing to embed risk-based controls that target mitigations to the riskiest use cases, roll out role-specific training, and apply organisational common benchmarks,
- Ongoing Monitoring and Compliance to automate organisation-wide quality assurance, monitoring, and red-teaming of Gen AI solutions.

The programme has supported the delivery of thousands of AI solutions and has delivered training to over 750,000 employees.

3.4 Incident Reporting

Gen AI systems, like all software systems, can occasionally have episodes where they fail to perform as intended or cause unintended harm. When they create harm of sufficient severity, these are often referred to as "incidents"; this often requires incident management activities to repair, resolve and even improve the functioning of a Gen AI system to avoid future occurrences. Incident reporting is an established practice that can support the continuous improvement of Gen AI systems through developing insights and proposing remediations.

Incident reporting and response can take several forms. Vulnerability reporting may take place before incidents happen; it includes cases where software product owners adopt vulnerability reporting as part of an overall proactive security approach. When incidents happen, reporting and response typically begin internally, where organisations have policies and procedures for their employees to report incidents through proper channels, along with appropriate tools and training to manage the incident through an incident management process. Externally, it refers to the capability to report incidents to key stakeholders outside an organisation when required or needed. Such stakeholders can include regulatory authorities, users, customers, or value chain participants.

It is important to note that Gen AI also often operates in an environment where existing incident reporting requirements that are technology agnostic may apply (e.g., personal data breach notifications or cybersecurity breach notifications).

When considering whether incident reporting structures and processes are necessary, it is necessary to consider both existing incident reporting requirements for Gen AI models and whether there is a need to streamline them. Reporting should also proportionately balance the benefits of providing comprehensive reports with the practicality of doing so, such as by tiering the need for reporting to the severity of incidents.

ASEAN can support organisations in the region by developing key concepts that can support organisations in understanding what is needed for incident reporting.

Areas for ASEAN to explore:

(a) Creating a common understanding around incident reporting and incident management

To support interoperability and foster a robust AI safety ecosystem, organisations deploying AI systems would benefit from having a common understanding around what incident reporting and incident management consist of, especially for incidents affecting safety, human well-being, or service access. **This potentially includes how incidents are defined, tracked, documented, and reported.**

A common understanding for incident tracking and management could help the region carry forward dialogue on this area. Commonly accepted incident reporting terminology in other domains, like cybersecurity, should be drawn upon for reference.

This could also be aligned with work that is being done at the international level to reduce fragmentation. **This can include referencing existing work by organisations like the OECD,** particularly its **Expert Group on AI Incidents, such as the definition of what an "AI incident" means and related terminology.**[9]

3.5 Testing and Assurance

Third-party testing, and eventually, formal auditing of Gen AI systems has been highlighted by some scientists and governments as key to supporting their responsible operation. Third-party testing and assurance complements and extends the benefits of the internal model evaluation that is a part of "Trusted Development and Deployment". Using third-party organisations to assess models, in addition to internal assessment, can build user trust by ensuring impartiality and credibility through an independent robust process. This draws on existing practices in a number of industries, such as finance and healthcare, where external auditors supplement organisational oversight functions.

However, defining a testing methodology for Gen AI that is reliable and consistent to complement internal testing is a work in progress. In addition, there is a need to also have sufficient numbers of independent entities to conduct such testing.

Areas for ASEAN to explore:

(a) Developing regionally applicable benchmarks and testing tools

There is a need to have common benchmarks and methodologies to reduce friction when testing across different models or applications. This matters more for ASEAN, which has an interest in facilitating regional business and reducing friction in compliance and testing.

Given that testing methodologies are still being developed, ASEAN has a unique opportunity to lead in developing standardised regional evaluation metrics and benchmarks that align with the region's characteristics and priorities, supporting safe and **reliable AI deployment for the region.** By establishing these benchmarks, ASEAN can empower model developers, deployers, users, and regulators across the region to quantitatively validate and assess Gen AI systems for safety, fairness, and performance **according to an established regional benchmark that can facilitate testing to address concerns relevant to ASEAN.**

One example of regional benchmarks can include both qualitative and quantitative indicators relevant to ASEAN's unique linguistic, cultural, and societal contexts. The development of open, regionally representative benchmarks reflective of ASEAN's languages, cultures, and values would support meaningful testing of Gen AI, ensuring models are not only technically sound but also culturally relevant and inclusive. These benchmarks should be adapted to sector-specific considerations, where those exist.

To enable assurance and evaluation processes, ASEAN should consider **agreeing on a list of evaluation tools and techniques preferred for use across the region.** This will go some way in encouraging companies to test Gen AI by providing them with more certainty on the value that testing will bring. This will also facilitate the standardisation of the developing audit and evaluation ecosystem by enabling the comparable assessment of assessment providers. ASEAN may also wish to consider opportunities to build capacity and support for organisations in adopting these metrics.

A·I· VERIFY FOUNDATION

AI Verify Foundation Project Moonshot *(see detail in Appendix)*

Project Moonshot is an open-sourced toolkit designed to streamline the testing of LLM applications by facilitating benchmarking and red-teaming at scale. It provides performance scores and reports tailored for non-technical audiences, with the goal of facilitating effective governance and oversight from organisational leadership.

Project Moonshot helps organisations using AI to:

- Select the right industry-leading benchmarks,
- Systematically scale their approach to validation and red teaming,
- Communicate safety information to non-technical stakeholders,
- Incorporate considerations relevant to ASEAN's cultural context.

The toolkit integrates industry-leading benchmarks, including both open-source and exclusive ones from partners like MLCommons and the Beijing Academy of AI. In addition to benchmarking, Project Moonshot automates red-teaming, traditionally a human-driven process, using algorithmic methods and LLMs to detect inappropriate content and enhance model robustness. The open-source nature of the project, which is freely available for download and integration into organisational software ecosystems, facilitates widespread, inclusive adoption.

Project Moonshot is developed by the AI Verify Foundation, a public-private partnership based in Singapore.

3.6 Security

The development and deployment of Gen AI systems introduces unique cybersecurity challenges that call for innovative and adaptive approaches to safeguard both AI and organisational infrastructures. Ensuring secure Gen AI deployment means addressing not only the protection of the AI system itself, but also reinforcing safeguards against unauthorised access to organisational systems that might lead to disruption or data loss. The opaque nature of Gen AI adds additional complexity. It makes it more challenging for defenders to detect and prevent attacks, and designing appropriate security controls to handle diverse inputs can be demanding and requires new approaches and skills.

While established cybersecurity best practices remain valuable and relevant, Gen AI as a novel technology may also require tailored security to address its unique considerations. This includes safeguards against novel threat vectors such as adversarial machine learning. Continuous advancements in cybersecurity approaches tailored to Gen AI will be essential to keep pace with this technology's fast-moving landscape.

ASEAN can use its convening power to support its member states, industry, and other organisations in the region in the sharing of best practices and enhancing collaboration around the security of Gen AI systems.

Areas for ASEAN to explore:

(a) Supporting and coordinating measures to promote vulnerability detection

Vulnerability detection initiatives – such as bug bounties and ethical hacking incentives – will play a crucial role in pre-empting and mitigating potential Gen AI security issues. ASEAN can support its members, industry, and other organisations in the region by sharing best practices for implementing such programmes and encouraging regional cooperation to foster interoperable initiatives. **Drawing on the experiences and best practices of both the public and private sectors** will reinforce the effectiveness of these measures. A key focus area should be on facilitating vulnerability reporting across both public and private sector boundaries.

The Republic of Korea's Ministry of Science and ICT established the AI Ethical Impact Assessment Framework in 2023 to identify and manage the ethical impacts of AI-based services. Its purpose is to support companies' voluntary commitment to practicing AI ethics and reliability, while providing standards for users to use AI in an ethical and responsible manner. To ensure the validity and reliability of the results, the actual assessment will be conducted by the Korea Information Society Development Institute (KISDI), a national research institute in Korea. Through pre-evaluation of the ethical impacts of AI products and services, the AI Ethical Impact Assessment will derive implications for management, institutional, and policy measures that maximise positive impacts and minimise negative ones. In 2024, the AI Ethical Impact Assessment will be piloted for AI-based video synthesis services.

(b) Promoting security knowledge-sharing among AI ecosystem stakeholders

Regular information-sharing among key ASEAN-region stakeholders will be critical for the timely and proactive development of Gen AI-specific security practices. This can include information on adversarial tactics, techniques and case studies for Generative AI. It is also important to explore and invest in new tools and security safeguards (e.g., studying digital forensics tools for Generative AI). This collaborative approach should **focus on securing Gen AI solutions through knowledge exchange on secure implementation practices and technological advancements.** ASEAN Member State cybersecurity authorities, national AI Safety Institutes, industry, civil society and academic participants, and other ecosystem players should explore open and accessible security tool development to further empower the regional AI community.

(c) Establishing guidelines on security by design

Dynamic, Gen AI-specific technical and operational guidance can support organisations in addressing the security considerations inherent to Gen AI. This guidance should include **insights into Gen AI system security, broader IT infrastructure security in Gen AI deployments, and should consider frontier risks related to security,** such as Gen AI's impact on cybersecurity (e.g., Gen AI-produced phishing or code). A continually evolving library of common security threats and taxonomy can aid cybersecurity authorities, AI Safety Institutes, industry, and civil society or academic participants in evaluating their existing security measures, contributing to a robust, secure-by-design AI environment across ASEAN.

In developing security-by-design guidelines for Gen AI, ASEAN should recognise that comprehensive guidance already exists at the global level. To enhance compatibility, ASEAN should align with well-established frameworks and standards from bodies like ISO, NIST, and OECD where relevant. This approach will help reduce regulatory divergence, facilitate cross-border trade, and reduce operational costs for businesses. These guidelines should be adapted to sector-specific considerations, where those exist.

Cyber Security Agency of Singapore Guidance on the Security of AI Systems

The Cyber Security Agency of Singapore (CSA) has worked with industry and international partners to develop guidelines ("Guidelines on Securing AI Systems") and a companion guide ("Companion Guide on Securing AI systems").[*] These were launched at the Singapore International Cyber Week in October 2024. The documents seek to support system owners in understanding the risks of AI and to provide practical advice and recommendations on how to secure AI systems throughout their lifecycle. System owners can use these documents, alongside other resources, to make informed decisions about adopting AI securely. CSA will keep these documents "live" and update them regularly in response to changes in the AI security landscape.

- Guidelines on Securing AI Systems. This document articulates the broad risk management principles and desired outcomes of securing AI. It provides principles-level guidance that can be an evergreen approach for system owners to secure AI.
- Companion Guide on Securing AI Systems. The Companion Guide offers practitioners a comprehensive reference and community resource when building their own AI security plans. It is meant to complement the Guidelines with a set of practical, specific and actionable measures to support system owners in adopting the Guidelines. It is not prescriptive, and curates practical measures, security controls and best practices from industry and academia.

[*] More details on the two documents can be found at the following link: https://csa.gov.sg/Tips-Resource/publications/2024/guidelines-on-securing-ai

3.7 Content Provenance

The outputs of Gen AI can be challenging to distinguish from original content produced by humans, such as text by an author, art by a digital artist, or photos from a camera. Deepfakes – a Gen AI technique that uses deep learning to generate realistic but artificial images, audio, or videos – have legitimate and transformative applications in marketing, historical reconstruction, education, and training, but also present opportunities for misuse, including in impersonation and the spread of misinformation and disinformation.[*] The increased accessibility of Gen AI tools has lowered the technical barriers to creating such content, thus expanding their potential for use both by legitimate, well-intentioned users and malicious actors. Without clear labelling practices or advanced content provenance tools, it can be difficult to determine when media has been generated by AI.

AI-generated content is increasingly challenging to distinguish from original content, and this has important implications for public trust in media and information integrity. For instance, malicious actors could use legitimate Gen AI tools to spread misinformation or disinformation, including to undermine electoral processes. There is a need for technical solutions, such as digital watermarking and cryptographic provenance, to help businesses and consumers easily discern and differentiate between original and AI-generated content. Notably, while these solutions are designed to be tamper-resistant, they are not foolproof; some watermarks may still be vulnerable to modification and some identifying metadata could still be removed from files. Bearing this in mind, technical solutions alone may not be sufficient and would likely need to be complemented by additional governance and enforcement mechanisms.

The Coalition for Content Provenance and Authenticity (C2PA) is a global consortium including industry, civil society, and media establishing digital content provenance measures. The C2PA has worked to define uniform technical standards for establishing cryptographic provenance metadata for digital media and to coordinate the adoption of those standards. Organisations around the world, including C2PA, continue to develop innovative solutions to the challenge of content provenance and work towards setting a global standard. For example, technology companies such as Google and Meta are implementing AI labelling on platforms such as YouTube and Instagram.

ASEAN is well-positioned to leverage its member states, industry, civil society, and academia to develop a response to the challenge of establishing content provenance. There is widespread acknowledgement of and support from key stakeholders for the need for AI-generated content to be identifiable as such. Developing a regional approach to doing so and improving regional capability to understand and participate in technical developments in this area will ensure that ASEAN can adopt best practices in line with global developments.

[*] See the ASEAN Guideline on Management of Government Information in Combatting Fake News and Disinformation in the Media. https://asean.org/book/asean-guideline-on-management-of-government-information-in-combating-fake-news-and-disinformation-in-the-media/

Areas for ASEAN to explore:

(a) Supporting development and capabilities around content

The adoption of content provenance technologies (e.g., cryptographic provenance or digital watermarking) should be accompanied by policies and enforcement measures. A holistic approach will enable governments, companies, organisations, and individuals to more effectively determine the origin of a piece of content to create a trustworthy media environment.

ASEAN's role is to use its convening power to encourage member states, industry, civil society, and academia to share best practices on content provenance technologies. Improving regional capabilities and understanding can help facilitate learning around how domestic policies and enforcement mechanisms could supplement these technologies. Developing **a regional repository of real world examples of techniques and approaches towards establishing content provenance would also enable users to learn more about content provenance** and its use. ASEAN should aim to align its approach with global norms and adopt widely-recognised solutions to the problem of content provenance wherever possible.

3.8 Safety and Alignment Research & Development

Many safety and alignment issues have been raised regarding Gen AI systems, and on some of these issues there is clear consensus in the policy and scientific communities that more research is needed. Many of these systems' effects are not yet well-understood, and technical methods to guarantee system behaviours are still being debated extensively by the expert community.

This underscores the need for dedicated research focused on both the social and technical dimensions of Gen AI's impact, in the present and future, and incorporating organisational, social, economic, and environmental dimensions. Such research will support organisational activities and policy measures to improve the safety and alignment of Gen AI and will form the basis of ongoing consultations between stakeholders.

ASEAN's role is to promote sharing around research initiatives by convening researchers and other stakeholders in regional knowledge-sharing fora to carry forward discussions on AI safety and alignment research priorities for the region.

Areas for ASEAN to explore:

(a) Sharing on AI safety and alignment research

A key insight from scientific literature is the need for sustained research on AI safety. Sharing ongoing work on safety and alignment research across ASEAN will empower member states, industry, citizens, and society to study the impacts and opportunities of Gen AI that are identified as key priorities for the region. ASEAN can use its role to **promote or convene platforms to facilitate the sharing of insights on AI safety and alignment research across member states and to help identify regional priorities for areas of further research.**

Where ASEAN Member States establish AI Safety Institutes (AISIs) or other national bodies with a mandate to study issues related to AI safety (such as industry networks, academic bodies, or digital agencies which include an AI research mandate), platforms for dialogue amongst these institutes can enhance their efforts by **developing beneficial partnerships and preventing duplicative work.**

In addition, some countries in Asia, including ASEAN member states, have established national AISIs or equivalent organisations. These organisations are natural partners to engage on discussions for research priorities for ASEAN. They may also be involved in ongoing wider discourse around AI safety at the international level. ASEAN should leverage such institutes, especially those established in the region, to plug into international conversations on AI safety.

3.9 AI for Public Good

Gen AI is a powerful, transformative tool with the potential to significantly improve people's lives. Citizens can use Gen AI to accelerate and improve daily tasks, to help them find information, or to improve their access to skills like coding. Governments can use Gen AI to make citizen services more personal or responsive, or to improve the efficiency of their administrative functions so they can serve their citizens faster and at less expense. Companies can use Gen AI to create a range of new products and services – most of which may not have been invented yet – as well as to improve speed and efficiency, which can reduce prices and improve experiences for their customers. Gen AI has the potential to unlock growth opportunities that will contribute to ASEAN's agenda for economic development and the improvement of the standards of living of its citizens. This technology has a powerful upside if ASEAN, its members, industry, society, and its citizens take the right steps to harness it for good.

ASEAN has a role to play in improving regional capacity to understand the potential of Gen AI to improve public sector delivery. This in turn can help member states make informed choices on public use of the technology and promote socially positive Gen AI use.

Areas for ASEAN to explore:

(a) Creating a compendium of Gen AI use cases

ASEAN can create a regional compendium of responsible Gen AI use cases by highlighting instances in which member states, industry, civil society, and academia are using Gen AI in an exemplary way. This compendium, especially when promoted by the member states and accessible in multiple languages, can serve as a knowledge reference that will help highlight the ways that Gen AI can be used responsibly and effectively.

A particular **focus on public sector use cases is likely to be helpful** to facilitate the use of AI for the public good. The compendium can also be a basis for further exchange and learning between member states on public sector applications of Gen AI, which in turn can help improve public service delivery within ASEAN member states through the responsible adoption of this technology.

(b) Promoting awareness and education on Gen AI

ASEAN can serve as the vehicle for educational measures on Gen AI that **promote awareness and skill development,** such as by launching publicity campaigns, skill-building workshops and training initiatives, or hosting online resources for citizens and business-owners to engage with. ASEAN can cooperate with member states, industry, as well as Dialogue Partners and Development Partners to design and deliver these awareness and education activities. By facilitating coordinated education and awareness across the region that offers a common message and baseline of skill acquisition, ASEAN can improve public access and understanding of Gen AI and help direct its use for socially positive purposes.

Awareness-building activities can **promote digital literacy and ensure that citizens are aware of the risks of Gen AI.** Helping citizens learn how to spot AI-generated content or use provenance markings, especially in lower-opportunity communities that may not otherwise have access to such information, is an opportunity for ASEAN to draw on its regional expertise to help all its citizens.

Educational initiatives can also help to ensure that the **workforces of ASEAN member states are prepared for the changes to work that Gen AI creates.** Offering trainings, or coordinating trainings offered by different bodies in different countries, can ensure that citizens are equipped with the skills to work with Gen AI as it becomes more prevalent and more crucial for regional growth.

04

Conclusion

04 Conclusion

The new challenges of Gen AI do not outweigh the substantial opportunities presented by this technology to improve people's lives, create economic opportunity, and empower new groups in society. This expanded Guide, supplementing the ASEAN Guide on AI Governance and Ethics that was published in February 2024, provides a policy-focused view of the risks and potential actions to promote responsible, risk-aware Gen AI use that ASEAN as a region should consider addressing as it continues to navigate seizing opportunities and managing challenges presented by Gen AI. This Guide is collaboratively developed by all ASEAN Member States and will serve as a foundation for ASEAN to take forward future work. It is intended to accompany the ASEAN Guide on AI Governance and Ethics (2024) as part of ASEAN's internal discourse on next steps for how the original Guide can be supplemented to leverage the opportunities and risks presented by Gen AI.

ASEAN Member States are recommended to apply, on a voluntary basis, the recommendations in this Guide. Nothing in this Guide may be interpreted as replacing or changing any party's legal obligations or rights under any ASEAN Member State's laws.

Appendix

Appendix: Use Cases

Illustrating the principles and recommendations of this expanded Guide in more detail, four use cases exemplify how organisations in ASEAN are tangibly approaching the issue of AI governance and ethics. These use cases were submitted by ASEAN Member States and feature examples of both public and private institutions.

The four use cases included in this expanded Guide are, in alphabetical order:

PhoGPT,
VinAI (Vietnam)

Project Moonshot,
the AI Verify Foundation
(Singapore)

Responsible AI
Internal Programme,
Accenture (ASEAN-wide)

ThaiLLM, BDI, NSTDA,
VISTEC and collaborators
(Thailand)

These use cases are documented in detail below.

PhoGPT, VinAI

VinAI is a subsidiary of the Vingroup conglomerate focused on developing innovative AI-based products and services, ranging from generative AI solutions and smart mobility to smart cameras. Beyond commercial software development, VinAI is also an active research institution that publishes academic papers and releases open-source software based on its work. Contributing to the open-source community and training young Vietnamese talent through the first AI residency programme in South-East Asia have helped VinAI improve the overall capability of its national AI ecosystem.

In 2023, VinAI launched an LLM called PhoGPT, the first open-source LLM developed for Vietnamese language and culture, as well as an accompanying technical paper. PhoGPT was developed in response to the success of other LLMs abroad – such as OpenAI's series of GPT models – that have been trained and tuned predominantly on English-language datasets that reflect Western cultural attitudes. This has limited the performance and relevance of Vietnamese-language applications of LLM technology, and in turn, presented a challenge to both its adoption and to the emergence of a Gen AI ecosystem in Vietnam.

While not intended to compete commercially with larger state-of-the-art models, PhoGPT was designed as a contribution to a growing body of open-source software and literature that is building the capabilities of Gen AI models that are relevant to the Vietnamese culture and language.

Features of PhoGPT

PhoGPT-4B, the base version of the model, is an LLM consisting of 3.7 billion parameters trained on a dataset of 102 billion words of Vietnamese-language text sourced from Wikipedia, medical texts, books, legal documents, news articles, and other web content. It was trained over a period of three months on a relatively small number of processors, making it much less costly to develop than leading commercial models. This focus on extensive Vietnamese-language data allows PhoGPT to perform well on a wide range of general topics in Vietnamese, ranging from research to public service and commercial uses. VinAI has expressed interest in extending the functionality of PhoGPT to other ASEAN languages, which also are frequently underrepresented in leading LLMs.

In performance comparisons, PhoGPT has demonstrated strong results, ranking ahead of many of the world's leading LLMs at the time of its release in Vietnam-specific tests of accuracy. The achievement of encouraging levels of performance on a relatively small investment in computing power suggests promising opportunities for the development of LLMs in other emerging economies.

PhoGPT-4B-Chat is a specialised version of the base model fine-tuned for conversational purposes. It has been trained on 70,000 examples of instructional prompts and an additional 290,000 conversations, making it ideal for Vietnamese-language chatbot applications.

Open-Source Deployment

PhoGPT is an open-source model that VinAI has made freely available online. This allows developers and researchers to access, modify, and integrate the model into their projects, and to retrain, fine-tune, or modify the model as required. This fosters innovation and collaboration, as well as promoting transparency and validation by allowing users to assess and rectify potential limitations in the model.

Open-source distribution also allows PhoGPT to reach the widest possible audience, including researchers and small firms that may not have the resources to license a model. This promotes the general adoption of Gen AI technology in Vietnamese-language applications. Contributing to and advancing the state of the open-source AI ecosystem in Vietnam has strong benefits for the sector overall and for private companies in the country; a healthy research ecosystem with access to advanced, culturally and linguistically relevant tools will support all industry participants through access to technology and expertise.

Value and Impact

PhoGPT's promising early results have given Vietnamese AI researchers an important stepping stone towards enhanced AI capabilities in models that are culturally and linguistically relevant, providing other institutions with a freely available model as a basis for ongoing progress. VinAI's research has supported a number of local universities in advancing their own work on Vietnam-specific Gen AI technologies, and its research on AI has helped facilitate its ongoing AI Residency programme, which has trained over 100 young Vietnamese AI scientists. It has also proven practically useful. An international technology firm has collaborated with VinAI to include PhoGPT in its AI software platform, making it available to its user base of companies in Vietnam and around the world.

The development of a locally relevant LLM represents a meaningful step towards an ASEAN data ecosystem (supporting the Data dimension) where ASEAN firms have access to high-quality datasets, and Gen AI tools built on those datasets, that are pertinent to their own linguistic and cultural contexts. This democratises access to Gen AI capabilities and will allow greater participation in the benefits of AI technology in the region.

Project Moonshot, AI Verify Foundation

The AI Verify Foundation[*] was established in June 2023 by Singapore's Infocomm Media Development Authority (IMDA), bringing together over 150 organisations across the global AI ecosystem. Spanning technology providers, innovative companies, non-profits, and foundations, it represents a unique effort to facilitate ecosystem collaboration on testing and governing AI.

In June 2024, IMDA worked with close industry partners from the Foundation and beyond to build an open-source LLM Evaluation Toolkit called Project Moonshot.[†]

A variety of factors makes testing for AI risks a challenge:

- **Numerous Benchmark Options** – Lack of clarity on how to approach the enormous number of available LLM benchmarks,
- **Scaling Challenges** – Difficulty of systematically selecting, validating, and red-teaming models at scale
- **Communication barriers** – Difficulty in communicating technical safety information to (non-technical) stakeholders,
- **Cultural Sensitivity** – Lack of tools and benchmarks sensitive to ASEAN's regional values, culture, and linguistic context.

Project Moonshot's Solution

Project Moonshot is a toolkit is designed to help organisations streamline the testing process for their LLM applications by effectively deploying benchmarking and red teaming at scale. Performance on each assessment is scored and presented in a report that is designed for a non-technical audience to use.

Benchmarking Capabilities

Benchmarking is an effective and industry-standard way of validating LLMs. Project Moonshot is curating industry-leading benchmarks into a single tool which can be integrated into leading cloud services platforms with minimal effort. These benchmarks take the form of question-answer pairs for which a score can be given. In addition to industry-leading open-source benchmarks, Moonshot includes a range of high-quality closed source benchmarks made available through its partnership with MLCommons and several other leading organisations such as the Beijing Academy of AI.

[*] AI Verify Foundation. "AI Verify Foundation." https://aiverifyfoundation.sg/
[†] AI Verify Foundation. "Project Moonshot." https://aiverifyfoundation.sg/project-moonshot/

Red-Teaming Capabilities

In addition to benchmarking, LLM application developers and users can also use red teaming to ensure that models are robust against efforts to generate inappropriate content and mitigate risks like the generation of harmful content.

Project Moonshot helps standardise and automate red-teaming practices, which traditionally required a fully human process to develop creative ways to break the system and were difficult to scale. The toolkit contains attack modules, which uses both algorithmic methods and LLMs to generate effective red-teaming prompts relevant to the model being tested.

Users Impacted

1. Organisations validating their LLMs before release,
2. Organisations choosing a LLM model for their context or use case,
3. Organisations with LLM applications looking to strengthen their guardrails.

Value and Impact

Project Moonshot represents an effective mobilisation of the industry to create an open and culturally relevant baseline assessment for model safety and performance in Singapore. It has taken a step towards mobilising international leading practices on the development and deployment of LLMs and represents a common hub for sharing open tools (supporting the **Testing and Assurance** dimension). The toolkit is being actively deployed by industry players, reflecting its practical value.

Taking these steps as part of a public-private consortium like the AI Verify Foundation has helped to ensure that they will be relevant to organisations using Gen AI. It has also ensured that the developers of Project Moonshot have been able to receive guidance, inputs, and feedback from industry on the toolkit’s continued improvement; its development is ongoing and has targeted the addition of new benchmarks, languages, and functionalities.

Project Moonshot is built in a modular architecture that enables developers to pick and choose the packages that they would like to integrate into their own testing pipelines. Multiple integration options – via API, command line, or with a web interface – are available. The project is open-source and can be accessed online.*

*AI Verify Foundation. “Project Moonshot.” https://aiverifyfoundation.sg/project-moonshot/

Responsible AI Internal Programme, Accenture

accenture

Accenture is a global professional services firm specialising in digital, cloud, and security solutions across strategy, consulting, technology, and operations. These services are supported by the world's largest network of Advanced Technology and Intelligent Operations centres and a workforce of approximately 750,000 people across 120 countries, with a large and longstanding presence in six ASEAN member states.

Accenture has long recognised the potential for AI to help transform how people live and work, and the need to responsibly develop, design and deploy this fast-growing technology for both its clients and for internal use. It is vital to scale this technology in responsible, ethical ways, and put AI governance and the responsible use of AI into practice to mitigate any potential risks. For this reason, Accenture started its Responsible AI Internal Programme in 2022 to ensure that the firm has the necessary tools to protect its organisation and its community, and to foster trust and confidence in its client interactions.

Accenture's Responsible AI Principles

Accenture's approach to developing and deploying AI solutions is founded on a set of principles that are applied to its own operations as well as its collaborations with clients, partners, and suppliers. These principles are:

- Human by design
- Fairness
- Transparency / Explainability / Accuracy
- Safety
- Accountability
- Compliance / Data Privacy / Cybersecurity
- Sustainability

Vision for Accenture's Responsible AI Internal Programme

Accenture's Responsible AI Internal Programme started by focusing on its overall AI governance approach: Formalising C-Suite governance and sponsorship of the overall programme, creating a governance framework, implementing key principles, policies, and standards, supporting cross-use case supervision, and forming an internal multi-disciplinary programme team.

When developing and deploying AI systems or tools, the programme focuses on the following three elements:

- **Conduct AI Risk Assessment:** Developing standard screening and assessment procedures for performing initial risk assessments and regulatory reviews of the firm's AI.
- **Systemic Enablement for RAI Testing:** Institutionalising the approach into its RAI compliance programme, implementing standards for AI procurement, embedding controls into technology/processes/systems, and developing benchmark testing tools and persona-based training for both traditional AI and Gen AI.
- **Ongoing Monitoring and Compliance of AI:** Enabling ongoing monitoring and compliance through quality assurance programmes, monitoring capabilities, incident remediation, and red teaming.

The programme has sought to innovate the delivery of its controls, itself using AI tools to automate and scale its operations. For instance, system owners can interact with a Gen AI chatbot to resolve issues or queries related to elements of risk assessment. By continuing to responsibly automate its risk assessment and testing activities, the programme aims to embrace new technology wherever possible to make its operations more efficient and to improve the rigour of its risk mitigation.

In order to build a responsible AI culture, the programme has also rolled out mandatory AI responsibility, ethics, and compliance trainings for team members working with AI, with the type of training tiered based on the requirements of their role. Basic responsible AI training has been rolled out across the firm's 750,000 people, with deeper AI ethics training available for the 30,000 most directly involved.

Value and Impact

Accenture's programme is an example of how the private sector can be empowered to build AI solutions in a manner sensitive to risk by cultivating an organisational understanding of the benchmarks, tools, leading practices, and capabilities required to implement effective controls (supporting the **Trusted Development and Deployment** dimension).

Accenture's Responsible AI Internal Programme has facilitated the evaluation of thousands of client engagements, Accenture assets, and internal applications, ensuring adherence to ethical guidelines and mitigating potential risks. When high-risk systems are identified, they are actively supported by the programme, ensuring that resources are concentrated on mitigating the most impactful potential use cases. The programme and its training component in particular have also allowed the organisation to improve its institutional knowledge of AI risks and mitigations – the foundations of an effective and universal responsible AI culture.

ThaiLLM, BDI, NSTDA, VISTEC and collaborators

BIG DATA INSTITUTE

สวทช NSTDA

VISTEC VIDYASIRIMEDHI INSTITUTE OF SCIENCE AND TECHNOLOGY

The Thai government is investing around USD 3 million (S$3.8 million) to develop the Thai Large Language Model (ThaiLLM) in an initiative spearheaded by the Big Data Institute (BDI), a public agency that manages government data, and its collaborators from public, private and academic sectors such as the National Science and Technology Development Agency (NSTDA), Vidyasirimedhi Institute of Science and Technology (VISTEC), the National Electronics and Computer Technology Center (NECTEC), AI Entrepreneur Association of Thailand (AIEAT), AI Association of Thailand (AIAT), Chulalongkorn University, and Mahidol University. This initiative is meant to serve as a common AI infrastructure for Thailand, reflecting two related challenges in the Thai AI ecosystem: that many of the leading LLMs are trained on predominantly English-language datasets that reflect Western cultural attitudes, and that small or emerging companies in Thailand may lack the capabilities to support the development of Thai-language LLMs of their own.

These constraints have limited the extent to which Thai organisations in the public and private sectors have been able to deploy Gen AI solutions, despite the importance in Thailand of sectors highly suited to Gen AI augmentation like health, travel, and the environment.

Establishing a Central Infrastructure for Collaborative Gen AI Development

BDI, NSTDA, VISTEC and their collaborators established ThaiLLM, a Gen AI model capable of generating meaningful and natural text similar to human-produced content. This initiative aims to create an open-source, foundational tool and open data for organisations in Thailand to use in chatbots and other Gen AI applications.

In the long term, it is intended for ThaiLLM to serve as a common national infrastructure where all LLM developers can contribute their datasets. This infrastructure will be managed under open licenses and will be open source, enabling all stakeholders to benefit from its growing body of Thai-language training material. To support this goal, ThaiLLM seeks to provide trainings for skills required to increase the number of AI users and AI professionals in Thailand.

Value and Impact

ThaiLLM looks to foster the development of Thai startups by lowering costs, increasing access to Gen AI capabilities, and decreasing dependence on foreign AI solutions. It intends to create a common, publicly funded LLM infrastructure that will scale over time as participants continue to share content, building the foundation of more widespread adoption to come.

The creation of a locally tailored large language model (LLM) marks significant progress toward establishing an ASEAN-centred data ecosystem (supporting the Data dimension). By providing access to high-quality datasets and AI tools that reflect the region's unique linguistic and cultural contexts, this initiative expands the availability of generative AI capabilities. As a result, it enhances the region's ability to engage with and benefit from AI technology, fostering greater inclusivity across ASEAN.

Appendix: Methodology

This document was developed on the basis of an extensive literature review and reflects extensive consultations conducted between the ASEAN Member States. This reflects the guiding principles of this Guide's policy recommendations, which advise ASEAN Member States to pursue an approach to AI Governance and Ethics that prioritises both regional coordination and alignment with an emerging global consensus. The principles, risks, and recommendations that this Guide provides are designed to balance interoperability with global developments with the needs of ASEAN as a region.

The literature review that facilitated the development of this document was extensive and attempted to capture a portion of the key documents with influence on the global consensus on AI governance and ethics.

The below list, while not exhaustive, highlights several of the reports, frameworks, legal instruments, and examples of global thought leadership that were referenced when developing this Guide:

- **AI Seoul Summit**
  - International Scientific Report on the Safety of Advanced AI: Interim Report (2024)
  - Seoul Declaration for Safe, Innovative and Inclusive AI by Participants Attending the Leaders' Session of the AI Seoul Summit (2024)
- **AI Verify Foundation**
  - Cataloguing LLM Evaluations (2023)
  - Model Artificial Intelligence Governance Framework (Second Edition) (2020)
  - Model AI Governance Framework for Generative AI (2024)
- **Centre for Data Ethics and Innovation (UK)**
  - The Roadmap to an Effective AI Assurance Ecosystem (2021)
- **Council of Europe**
  - Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (2024)
- **European Commission (EU)**
  - EU AI Act (2024)
- **Google**
  - An AI Opportunity Agenda for ASEAN (2024)
- **Group of Seven**
  - Hiroshima Process Code of Conduct for Organisations Developing Advanced AI Systems (2023)
- **Infocomm Media Development Authority (Singapore)**
  - Generative AI: Implications for Trust and Governance (2023)
- **International Standards Organisation**
  - ISO/IEC 42001:2023 – Information Technology – Artificial Intelligence – Management System (2023)

- **Meta**
    - Open Loop US Program: Red-Teaming & Synthetic Content (2024)
- **Microsoft**
    - Global Governance: Goals and Lessons for AI (2024)
- **Ministry of Foreign Affairs (China)**
    - Global AI Governance Initiative (2023)
    - Shanghai Declaration on Global AI Governance (2024)
- **Monetary Authority of Singapore (Singapore)**
    - Emerging Risks and Opportunities of Generative AI for Banks (2024)
- **National Information Security Standardisation Technical Committee (TC260) (China)**
    - AI Safety Governance Framework (v1.0) (2024)
- **National Institute of Standards and Technology (US)**
    - Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations (AI 100-2 E2023) (2024)
    - Artificial Intelligence Risk Management Framework (AI RMF 1.0) (2023)
    - Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (NIST AI 600-1) (2024)
- **Office for Artificial Intelligence (UK)**
    - A Pro-Innovation Approach to AI Regulation (2023)
- **Organisation for Economic Cooperation and Development**
    - Stocktaking for the Development of an AI Incident Definition (2023)
    - Initial Policy Considerations for Generative Artificial Intelligence (2023)
    - Recommendations of the Council on Artificial Intelligence (2024)
- **Stanford Institute for Human-Centered Artificial Intelligence**
    - Artificial Intelligence Index Report (2024)
- **United Nations**
    - Resolution on Seizing the Opportunities of Safe, Secure and Trustworthy Artificial Intelligence Systems for Sustainable Development (A/RES/78/265)
    - Resolution on Enhancing International Cooperation for AI Capacity Building (A/RES/78/311)
- **United Nations Educational, Scientific and Cultural Organisation**
    - Recommendation on the Ethics of Artificial Intelligence (2021)
- **US-EU Trade and Technology Council**
    - TTC Joint Roadmap on Evaluation and Measurement Tools for Trustworthy AI and Risk Management (2022)
- **White House (US)**
    - Blueprint for an AI Bill of Rights (2022)
    - Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (2023)

In addition to the findings of this body of literature, the perspective of subject matter experts in Gen AI development and deployment from Accenture was consulted. This Guide was collaboratively developed by ASEAN Member States on the basis of dialogue and the exchange of ideas.

References

[1] Accenture. “Gen AI-powered reinvention: APAC’s opportunity to outpace the competition”. https://www.accenture.com/content/dam/accenture/final/accenture-com/document-3/Accenture-Gen-AI-Powered-Reinvention.pdf

[2] Bai, Y., et al. “Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback”. https://arxiv.org/abs/2204.05862

[3] Bai, Y., et al. “Constitutional AI: Harmlessness from AI Feedback”. https://arxiv.org/pdf/2212.08073

[4] Meta Open Loop. “Red-Teaming & Synthetic Content”. https://www.usprogram.openloop.org/site/assets/files/1/openloop_us_phase1_report_and_annex.pdf

[5] IBM. “What is Shadow IT?”. https://www.ibm.com/topics/shadow-it

[6] Liu, Y., et al. “Prompt Injection attack against LLM-integrated Applications”. https://arxiv.org/pdf/2306.05499

[7] OWASP. “LLM03:2025 Supply Chain.” https://genai.owasp.org/llmrisk/llm032025-supply-chain/

[8] Vassilev, A., Oprea, A., Fordyce, A., and Anderson, H. “Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations”. https://csrc.nist.gov/pubs/ai/100/2/e2023/final

[9] OECD. “Defining AI Incidents and Related Terms”. https://www.oecd-ilibrary.org/science-and-technology/defining-ai-incidents-and-related-terms_d1a8d965-en