CONCORDIA AI
安远AI

State of AI Safety in Singapore

July 2025

About Concordia AI

Concordia AI is a social enterprise with a mission to ensure that AI is developed and deployed in a way that is safe and aligned with global interests. It is not affiliated with, nor funded by, any government or political organisation. The views expressed in this report are those of the authors alone. Concordia AI received no financial support from any government or corporate entities for the research, writing, or publication of this report.

Authors

This report is written by Jonathan Lee, Jason Zhou, Kwan Yee Ng, and Brian Tse.

How to cite this report

Jonathan Lee, Jason Zhou, Kwan Yee Ng, and Brian Tse, "State of AI Safety in Singapore," Concordia AI, July 2025, https://concordia-ai.com/wp-content/uploads/2025/07/State-of-AI-Safety-in-Singapore-2025.pdf.

Testimonials

"The State of AI Safety in Singapore report provides a comprehensive overview of the AI safety and governance landscape in Singapore and contributes to the growing global conversation on AI safety. It will serve as a valuable reference for practitioners and academics interested in AI governance and safety research in Singapore and beyond."

Professor Simon Chesterman,
National University of Singapore

"This report is an insightful overview of Singapore's evolving AI safety landscape and the rigorous testing frameworks that support it. The report also serves as a useful input for identifying technical research areas that are key to building trust in AI systems. It highlights Singapore's commitment to both the governance and technical research in this area."

Professor Kwok-Yan Lam,
Nanyang Technological University

"By analysing the crucial issues of AI safety, this report makes a compelling case for deeper interdisciplinary collaboration and will be a very useful resource for researchers and practitioners alike."

Professor Heng Wang,
Singapore Management University

"AIDX Tech is proud to have contributed to this report. Its clear overview of Singapore's expanding AI assurance ecosystem provides a valuable roadmap of testing and resources, empowering AI developers and AI adopters to accelerate the deployment of trustworthy AI."

Dr. Yifan Jia,
Founder, AIDX Tech

Acknowledgements

We would like to thank the following individuals and organizations for their support and feedback to the report (in last name alphabetical order):

- Simon Chesterman, National University of Singapore
- Shaun Ee, Institute for AI Policy and Strategy
- Yifan Jia, AIDX Tech
- Mohan Kankanhalli, National University of Singapore
- Tristan Koh, AI Singapore
- Kwok-Yan Lam, Nanyang Technological University
- Tze Yun Leong, National University of Singapore
- Erica Liaw, AI Governance and Safety, Infocomm Media Development Authority of Singapore
- Calissa Man, Independent
- Clement Neo, Nanyang Technological University
- Zilan Qian, University of Oxford
- Thayalini Selvaraj, National University of Singapore
- Saad Siddiqui, Safe AI Forum
- Robin Staes-Polet, The Future Society
- Zhi Xuan Tan, National University of Singapore
- Marie Teo, Tony Blair Institute for Global Change
- Heng Wang, Singapore Management University
- Edward Yee, FAR.AI

All views and conclusions expressed remain solely those of the authors.

Table of Contents

Executive Summary

While achieving AI's global benefits and managing its risks requires broad international cooperation, current global debates concentrate disproportionately on the few nations building state-of-the-art AI systems. Singapore shows that smaller, resource-constrained states can still influence emerging safety norms. Since introducing the Model AI Governance Framework in 2019—one of the world's first—the city-state has focused on building pragmatic, industry-ready tools and multilingual safety evaluations that translate high-level principles into day-to-day practice. This report, current to early July 2025, provides the first full survey of Singapore's AI safety ecosystem, organized into five domains.

Domestic Approach

- **Singapore relies on voluntary frameworks and targeted legislation instead of a broad or national AI-specific law.** The Model AI Governance Framework, first issued in 2019 for traditional AI and updated in 2024 for generative AI, provides broad voluntary guidelines for industry, while legislation is targeted and focuses on specific AI risks, such as new penalties for AI-generated election deepfakes. There is no clear move toward a national AI law at the moment.
- **Policy instruments emphasize downstream testing and assurance rather than model-level controls.** Initiatives such as the "Starter Kit for Safety Testing of LLM Applications" and the "Global AI Assurance Pilot" provide deployers with dedicated test cases and specific guidance on how to test different components of generative AI applications for safety risks. Because testing and evaluations are less well-explored at the application level than at the model level globally, Singapore's focus on deployment testing positions the country to fill an important gap in global AI safety practice.

International Approach

- **Singapore plays an outsized convening role in global and regional AI governance.** It has actively engaged in global AI governance discussions since 2018, contributing at the United Nations and global AI safety summits, among other fora. It leverages its neutral foreign policy stance to convene international AI events such as the Singapore Conference on AI and uses its diplomatic platforms to amplify the voices of smaller states. As Chair of the Association of Southeast Asian Nations (ASEAN) Digital Ministers' Meeting in 2024 and through the Digital Forum of Small States (Digital FOSS) initiative, Singapore has promoted inclusive dialogue and capacity-building so that developing countries can help shape global AI norms.
- **Bilaterally, Singapore embeds AI governance clauses in trade and digital agreements and encourages interoperability through 'crosswalks' that map international governance frameworks onto each other.** Recent digital economy agreements have included provisions for

building AI governance systems and sharing best practices between partners. The AI Verify Testing Framework streamlines compliance by mapping to NIST AI Risk Management Frameworks and ISO/IEC 42001, allowing businesses to meet AI safety obligations through a single testing process.

Industry

- **Singapore's home-grown models prioritize training on regional languages, with safety features still at an early stage and slated for further development.** The SEA-LION and MERaLION model families focus on Southeast Asian languages and dialects rather than frontier capability; current safeguards are limited to basic toxicity evaluations and the SEA-Guard prompt filter, which is in the early stages.
- **Singapore is a vibrant assurance hub, hosting major foreign general-purpose AI developers and both local and international AI safety testing and assurance providers.** Leading US and Chinese technology companies and global frontier start-ups maintain Singapore offices or partnerships on AI safety testing, bringing expertise to Singapore. Meanwhile, local and international AI assurance companies form a comprehensive ecosystem by providing testing and assurance services across model, application, and organizational levels.

Technical Research

- **Academic research on AI safety is expanding with Singaporean universities serving as the primary drivers.** Most publications come from NUS, NTU, SMU, and SUTD, with support from A*STAR, GovTech and the Singapore AI Safety Institute. Their research center on robustness, multimodal safety, unlearning, and agent behavior, and the "Singapore Consensus on Global AI Safety Research Priorities" (May 2025) provides a set of technical research areas—risk assessment, safety-by-design development, and post-development control—providing a roadmap for future collaborations toward underexplored risk areas.

Public Opinion

- **Reliable public opinion data on AI safety are limited and focused on near-term concerns.** No dedicated national survey has probed Singaporeans' views on AI risks. Few global polls include Singapore, those that do reveal public concern about misinformation, data privacy, cybersecurity, and reduced human interaction.

Introduction

How can a small nation without frontier AI capabilities meaningfully contribute to global AI safety? This question has gained importance as global AI governance discussions increasingly center on the actions of a few major powers, potentially overlooking the approaches of smaller yet strategically well-positioned countries. Despite Singapore's early move into AI governance—issuing the world's first comprehensive national AI governance framework in 2019—international understanding of Singapore's approach remains limited. This gap matters because Singapore represents a different model in the AI ecosystem. Rather than competing on raw computational power or pursuing frontier AI model development, Singapore has built notable strengths in AI assurance, multilingual safety testing, and practical governance frameworks that bridge cutting-edge global standards with regional needs.

Singapore's approach has evolved significantly since its pioneering 2019 Model AI Governance Framework. From that first official publication, the city-state has developed a range of governance instruments and safety infrastructure. In the first half of 2025 alone, Singapore convened the second Singapore Conference on AI (SCAI), which produced the "Singapore Consensus on Global AI Safety Research Priorities"; released a Joint Testing Report with other national AI Safety Institutes (AISIs); launched a generative AI application testing pilot and sandbox; and advanced its open source model line-up with MERaLION-2.

Structural factors such as limited land area, high energy costs, and the absence of local frontier AI labs make large-scale model training impractical in Singapore. However, the nation has turned these constraints into opportunities by focusing instead on later stages of the AI lifecycle. Singapore provides independent AI model testing frameworks, conducts application-level safety testing initiatives ("Global AI Assurance Sandbox"), and issues practical guidelines for AI deployers ("Starter Kit for Safety Testing of LLM-Based Applications").

In addition, Singapore has focused on multilingual testing and has backed a home-grown AI model pipeline that puts linguistic diversity first. The country has four official languages—English, Mandarin Chinese, Malay, and Tamil—and is situated amid a diversity of other languages in Southeast Asia. Its SEA-LION family of language models, along with the multimodal MERaLION, covers 13 Southeast Asian languages and even local dialects or accents. Beyond models, Singapore has also organized multilingual red-teaming exercises and other projects to address the under-researched challenges of non-English-language AI safety.

Singapore's influence extends well beyond its borders. In the region, SEA-LION's architecture underpins Thailand's Wangchan-LION and Indonesia's Sahabat-AI large language models, while the "ASEAN Guide on AI Governance and Ethics" (later expanded to cover generative AI) draws heavily on Singapore's Model AI Gov-

ernance Frameworks. Building on these regional contributions, Singapore has emerged as a trusted convenor in global AI governance; since 2018 it has participated in United Nations discussions as well as the global AI Safety Summits. Its reputation for diplomatic neutrality and reputation as a reliable international host enables the country to host high-profile convenings at home—including the Singapore Conference on AI—while initiatives such as the Digital Forum of Small States (Digital FOSS) AI Playbook ensure that smaller nations share in setting international norms. These initiatives showcase how a small but diplomatically active city-state can shape AI safety norms by leveraging its unique strengths in domestic governance, international collaboration, third-party testing communities, and local research in AI safety science.

To address this knowledge gap and document Singapore's distinctive contributions to global AI safety, this report provides the first comprehensive analysis of Singapore's AI safety ecosystem. We examine Singapore's approach to AI safety and governance across five key areas:

1. **Domestic Approach:** We analyze Singapore's multi-layered governance strategy—comprising voluntary frameworks, targeted legislation, national standards, and testing/evaluation initiatives—and show how these measures address different categories of AI risks.
2. **International Approach:** We examine how Singapore's international reputation enables it to convene high-profile AI events and contribute to global AI governance through multilateral fora, regional initiatives, and bilateral collaborations.
3. **Industry:** We explore Singapore's homegrown general-purpose AI models and thriving third-party AI assurance market, as well as the role of foreign AI developers in the local ecosystem.
4. **Technical Research:** We map the landscape of AI safety research in Singapore across universities, government agencies, and research institutes, outlining key research themes in AI safety.
5. **Public Opinion:** We analyze available survey data on Singaporean public attitudes toward AI risks.

Through the report, we aim to provide the following outcomes to our readership:

1. **For policymakers:** Provide insights and takeaways from Singapore's pragmatic AI governance approaches that could inform policy discussions globally.
2. **For AI researchers:** Highlight opportunities for collaboration in AI safety science with Singapore's institutions and initiatives.
3. **For industry (AI developers and third-party AI assurance suppliers):** Demonstrate Singapore's value as a regional hub for independent AI assurance, testing, and certification.
4. **For our general readership:** Offer perspective on how a small, multilingual state can contribute meaningfully to global AI safety practices.

Given the breadth of topics covered, some initiatives are only summarized rather than described in full detail. Readers are encouraged to use this report as a starting point for deeper exploration into specific areas of interest.

Scope of the Report

This report concentrates on the risks posed by advanced AI systems with broad capabilities, which have been referred to as "generative AI" in Singapore's 2023 paper "Generative AI: Implications for Trust and Governance,"[a] and as "general-purpose AI (GPAI)" in the more recent "International AI Safety Report" (2025) and the "Singapore Consensus on Global AI Safety Research Priorities" (2025).[2,3] This report discusses both "generative AI" and "GPAI"; in both cases, this refers to "systems that can perform or can be adapted to perform a wide range of tasks." We analyze three risk categories that arise from these systems: malicious use (e.g. fake content, cyber offence, and biological and chemical attacks); malfunctions (e.g. reliability issues, bias, and loss of control); and systemic risks (e.g. labor market risks, global AI divides, and environmental risks).

Accordingly, our analysis is limited to the risks that arise from these generative AI or GPAI systems. We do not examine sector-specific rules that govern narrower applications, such as the Monetary Authority of Singapore's FEAT principles for finance,[b] the Road Traffic Act's autonomous vehicle regulations,[5] guidelines on healthcare provision,[6] or military uses of AI. Many other AI requirements are embedded in broader industry guidelines, and cataloguing every such provision lies beyond this report's scope. Instead, we focus on developments most relevant to the safety of generative and general-purpose AI.

Sources used in the report include official government publications, standards-setting documents, peer-reviewed papers, industry reports, and the latest public opinion polling. Because all material is drawn from public information, the picture is necessarily partial. Some initiatives may be underway but not yet public, and details in fast-moving areas (e.g. model releases, consultation papers, pilot projects, standards) can quickly change or be rescinded. Our research cut-off was early July 2025; developments after that date fall outside the scope of this report. Readers should therefore view the report as a snapshot rather than a definitive record and consult primary sources for the latest information.

a. In this paper, the term "generative AI" is used to collectively represent both foundational models that have the ability to perform a wide range of tasks, and task-specific models.[1]

b. Fairness, Ethics, Accountability and Transparency principles. [4]

Domestic Approach

Key takeaways

- Singapore entered the AI governance conversation early with its Model AI Governance Framework, which was among the first comprehensive national AI governance frameworks when published in 2019 for traditional AI. It was updated in 2024 to address generative AI.
- There is no national AI-specific legislation in Singapore. Instead, Singapore relies on voluntary frameworks and narrowly-targeted legislation (deepfake provisions in election and online-safety regimes) to address generative AI harms.
- Most governance initiatives concentrate on risks that arise during deployment (e.g. providing testing frameworks and assurance for AI applications) rather than at the model level.
- Singapore has issued its first national AI standard (SS ISO/IEC 42001) in 2025, adapting the international ISO/IEC 42001:2023 to the local context.
- A consistent thread in Singapore's governance efforts is explicit testing and guidance for Southeast Asian languages and culturally diverse user groups, to reduce bias and broaden safe adoption.

Several public bodies across policy, regulatory and research play a role in AI governance. The Ministry of Digital Development and Information (MDDI), renamed from the Ministry of Communications and Information (MCI) in July 2024, sets digital policy and oversees, inter alia, the Infocomm Media Development Authority (IMDA), Personal Data Protection Commission (PDPC), Government Technology Agency (GovTech) and the Cyber Security Agency of Singapore (CSA). IMDA regulates the media and infocomm sector, which include spearheading AI regulations in Singapore. PDPC provides guidelines on the use of personal data for AI deployments, while CSA tackles AI cybersecurity risks. GovTech, the government's technology arm, translates these principles into day-to-day practice for public servants through resources such as the Responsible AI Playbook.[7] The Singapore Standards Council sets new AI standards through its AI Technical Committee (AITC). Together, these government agencies form the core of Singapore's AI governance architecture.

This section first discusses voluntary AI governance frameworks and AI safety testing and assurance initiatives, as they form the core of Singapore's approach to AI governance. Then, we examine hard regulations specific

to safety risks from generative AI or GPAI, which are currently confined to clearly defined harms from AI-generated content, such as deepfake election advertising or sexually explicit synthetic images.[a] Lastly, we explore Singapore's national standards on AI governance.

1.1 Voluntary Frameworks

Singapore's earliest official publication on AI governance was in June 2018, in the "Discussion Paper On Artificial Intelligence (AI) And Personal Data – Fostering Responsible Development And Adoption Of AI" drafted by the Personal Data Protection Commission of Singapore (PDPC).[8] This kickstarted a consultation process with industry and government agencies,[9] and led Singapore to publish the voluntary Model AI Governance Framework (MGF), first for traditional AI in 2019 and 2020, and later on for generative AI.

In January 2019, IMDA and PDPC released the first Model AI Governance Framework to guide companies to implement AI responsibly.[10] It translated principles into practices for companies, in four key areas:

1. Internal governance structures and measures;
2. Determining the level of human involvement in AI-augmented decision-making;
3. Operations management; and
4. Stakeholder interaction and communication.

The framework highlighted two "high-level guiding" principles: (1) Human-centricity and (2) Decision-making that is "Explainable, Transparent, and Fair," while highlighting risks such as unintended discrimination, opacity, and systemic bias. A harm-severity matrix (Figure 1.1) defines when humans must retain control in safety-critical contexts; for a high severity, high probability risk scenario, a human-in-the-loop decision-making model is recommended.

IMDA and PDPC released an update to the MGF in January 2020,[11] which focused on providing more practical guidance for industry implementation. The update included examples illustrating how various organizations had implemented AI governance practices and was accompanied by companion guidance documents such as "Companion to the Model AI Governance Framework - Implementation and Self-Assessment Guide for Organizations,"[12] and "Compendium of Use Cases - Practical Illustrations of the Model AI Governance Framework."[13]

Changes in the 2020 update included the addition of particular interpretability tools, scenario-based testing and adversarial testing to improve robustness, and a call for "regular model tuning." The Implementation and Self-Assessment Guide provided questions that companies can use to evaluate their compliance with the MGF, such as asking companies if they can explain how the AI model arrives at an outcome and suggesting

a. There are sector-specific regulations related to AI, such as the Fairness, Ethics, Accountability and Transparency (FEAT) principles in the financial sector, and the Road Traffic Rules which cover autonomous driving; these regulations are outside the scope of the report and will not be discussed in detail.

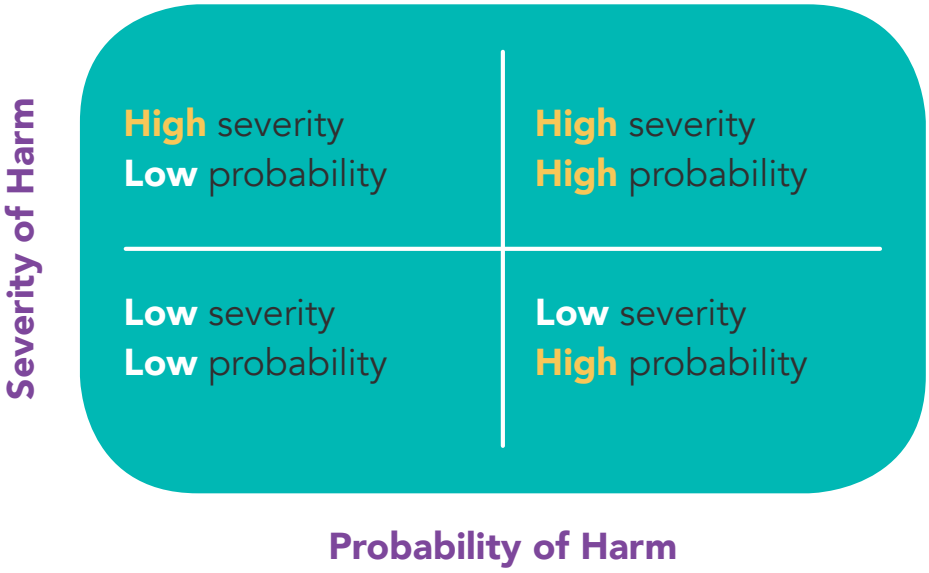

Figure 1.1: Harm Assessment Matrix in MGF

several strategies to improve model explainability. The Compendium included use cases that illustrate how Singapore-based organizations had implemented or aligned their AI governance practices with the MGF.

As the earlier MGF editions focused on traditional AI, Singapore updated the framework in 2023 to cover generative AI. In June 2023, IMDA released a "Discussion Paper on Generative AI - Implications for Trust and Governance" that defined "Generative AI models" as models with the ability to generate text and other media types, which includes both task-specific models and foundation models that can perform a wide range of tasks, effectively covering general-purpose large models.[14] The discussion paper argued that generative AI poses novel risks not present in traditional AI, such as hallucination, controllability, misalignment, and cyber misuse. It also set out an initial approach for addressing generative AI risks while acknowledging potential "existential risks from powerful AI."

The Discussion Paper was followed by the "Model AI Governance Framework for Generative AI" (MGF-GenAI), which was finalized in May 2024.[15] MGF-Gen AI sets out a systematic and balanced approach for tackling generative AI risks. It outlines nine dimensions that must be addressed together to foster a trusted ecosystem, as no single intervention will be a silver bullet. This included the "Trusted Development and Deployment" dimension, which expanded upon the principles of the earlier MGF. The following table (Table 1.1) elaborates on the suggested measures under each of these nine dimensions.

Table 1.1: The nine dimensions of the MGF-GenAI framework

MGF-GenAI framework	
Dimension	**Framework Measures**
Accountability	Setting up the right incentive structure for different players in the AI system development life cycle to be responsible to end-users
Data	Ensuring data quality and addressing potentially contentious training data in a pragmatic way, as data is core to model development
Trusted Development and Deployment	Enhancing transparency around baseline safety and hygiene measures based on industry best practices, in development, evaluation and disclosure
Incident Reporting	Implementing an incident management system for timely notification, remediation and continuous improvements, as no AI system is foolproof
Testing and Assurance	Providing external validation and added trust through third-party testing, and developing common AI testing standards for consistency
Security	Addressing new threat vectors that arise through generative AI models
Content Provenance	Transparency about whether content is AI-generated or human-produced, as a useful signal for end-users
Safety and Alignment Research & Development (R&D)	Accelerating R&D through global cooperation between AI Safety Institutes, to improve model alignment with human intention and values
AI for Public Good	Harnessing AI to benefit the public by democratizing access, improving public sector adoption, upskilling workers, and developing AI systems sustainably

MGF-GenAI significantly broadened Singapore's AI governance approach in two key ways. First, it expanded coverage across the whole AI system lifecycle, addressing not just AI developers but also application developers and deployers. Second, it dramatically widened the scope of the AI risks it addressed, incorporating risks such as dangerous capabilities, value misalignment, and autonomous replication.

1.2 AI Safety Testing and Assurance

Another aspect of Singapore's AI governance approach is AI safety research, testing, and evaluations. In May 2022, Singapore introduced "AI Verify,"[16] which consisted of:

1. A testing framework aligned to globally recognized governance principles;[b] and
2. A software toolkit for conducting technical tests for Fairness, Explainability, and Robustness and for generating testing reports to help companies build trust with their stakeholders.

b. The 11 governance principles are transparency; explainability; repeatability/reproducibility; safety; security; robustness; fairness; data governance; accountability; human agency and oversight; and inclusive growth, and societal and environmental well-being.[17]

As Singapore updated MGF to cover generative AI in 2023, IMDA also extended AI Verify to cover generative AI risks through the launch of Project Moonshot in May 2024.[18] Project Moonshot is an open source evaluation toolkit to assess the safety and capability of large language models (LLMs) and downstream applications through benchmarking and red-teaming. It now includes more than 100 datasets that can be used to benchmark LLM performance on trust and safety issues such as bias, toxicity, hallucination. It also includes red-teaming modules that can be used for adversarial testing.

In June 2023, IMDA set up AI Verify Foundation (AIVF) to manage open sourcing efforts and to lead in growing an AI testing and assurance community. To date, AIVF has around 200 member organizations, including "premier" members like Amazon Web Services (AWS), Dell, Google, IBM, Microsoft, Red Hat, Resaro, and Salesforce.[c]

One of AIVF's first initiatives was the "Cataloguing LLM Evaluations" paper (Catalogue) with IMDA in October 2023.[20] The Catalogue introduced a taxonomy of the LLM evaluation landscape and recommended a baseline set of evaluations that an LLM should be tested on pre-deployment. The paper defined "frontier models" using the Frontier Model Forum's definition of "large-scale machine learning models that exceed the capabilities currently present in the most advanced existing models, and can perform a wide variety of tasks." AIVF adopted the extreme risks taxonomy from the research paper "Model evaluation for extreme risks," which outlines both dangerous capabilities (e.g., offensive cyber capabilities, weapons acquisition, self-replication, and persuasion) and alignment risks (e.g., power-seeking behavior and resistance to shutdown).[21] In conjunction with the Catalogue, AIVF launched the "Generative AI Evaluation Sandbox",[22] which drew on the Catalogue to set out common baseline methods for assessing LLMs. This brought together model developers, application developers, and third-party testers, allowing them to evaluate AI systems through concrete use cases.[d]

In May 2024, AIVF signed a memorandum of intent with MLCommons to co-develop a common set of safety-testing benchmarks for generative AI models.[23] The partnership—backed by the MLCommons AI Safety working group of researchers, industry engineers, standards experts, and civil-society advocates—aimed to establish an internationally accepted baseline that developers and policymakers can use to evaluate models across languages, cultures, and risk contexts. The AILuminate v1.0 benchmark (English) was released in December 2024, with subsequent release (v1.1 French model) in February 2025, and a proof-of-concept in May 2025 for the Chinese version, packaged with an evaluator model, which AIVF is developing, in collaboration with NUS. [24e]

IMDA and AIVF further recognized that while ongoing generative AI testing primarily focused on model-level testing, it was also important to test generative AI applications. In February 2025, IMDA and AIVF launched the "Global AI Assurance Pilot" (Pilot) to codify emerging norms and best practices around technical testing

c. Premier members form the AIVF's governing committee and are involved in setting AIVF's strategy.[19]

d. These players include key model developers like Google, Microsoft, Anthropic, IBM, NVIDIA, Stability.AI and Amazon Web Services (AWS); app developers with concrete use cases like DataRobot, OCBC, Global Regulation Inc, Singtel, and X0PA.AI; and third-party testers such as Resaro.AI, Deloitte, EY, and TÜV SÜD.

e. Ghosh et al., 2025.[25] The paper mentioned AIVF is leading an effort to include Chinese-language evaluation and bolster "APAC adoption".

of generative AI applications. 16 specialist AI testers from around the world were paired with 17 deployers of real-world generative AI applications (from 10 different industries including finance, healthcare, recruitment, and the public sector).[26] The Pilot targeted three specific outcomes:

1. Developing testing norms that could inform future standards for technical testing of generative AI applications;
2. Creating AI testing tool roadmaps to guide open source and proprietary testing software (including AIVF's Project Moonshot); and
3. Establishing foundations for a viable assurance market and potential future accreditation programs.[27]

The pilot's findings also informed IMDA's "Starter Kit for Safety Testing of LLM-Based Applications" (Starter Kit), which was released for public consultation in May 2025. The Starter Kit aimed to provide organizations with "practical step-by-step reference" for how to think about and test for common risks.[28] It focused on four key risks commonly encountered in generative AI applications—hallucination, undesirable content, data disclosure, and vulnerability to adversarial prompts. The Starter Kit offered guidance on a structured approach to testing, from output to components, and recommended tests and testing methodologies for each of the risks (Table 1.2).

The Starter Kit brought together insights from the companies that participated in the Global AI Assurance Pilot, workshops with industry, and collaboration with government agencies such as the Cyber Security Agency of Singapore (CSA). By July 2025, seven baseline tests had been made available on Project Moonshot for organizations to access and execute, with more to be included later.

In July 2025, building on the successful Pilot, IMDA and AIVF launched the "Global AI Assurance Sandbox" to continue connecting application developers and testing providers to perform testing on generative AI applications. The Global AI Assurance Sandbox referenced the four risk dimensions of the Starter Kit: hallucination, undesirable content, data disclosure, and vulnerability to adversarial prompts.[29]

While AIVF focuses on collaboration with and outreach to the industry, government-led research to advance the science of AI safety is conducted by the Singapore AI Safety Institute (AISI). The Digital Trust Centre (DTC), a national research center at the Nanyang Technological University, was designated as the Singapore AISI in 2024. The DTC was initially set up with a S$50 million grant in 2022 by IMDA and the National Research Foundation to conduct research in trust technologies such as privacy protection solutions. It now incorporates AI safety in its scientific research.[30]

Table 1.2: Overview of Starter Kit

Overview of Starter Kit				
	Hallucination	**Undesirable Content**	**Data Disclosure**	**Vulnerability to Adversarial Prompts**
Output Tests				
Baseline Tests Public benchmarks + Red teaming	Test tendency to produce incorrect output with regard to basic facts (e.g. general knowledge, Singapore) or basic tasks (e.g. fact-finding, summarisation)	Test tendency to produce common types of socially harmful (e.g. toxic and hateful), legally prohibited or crime-facilitating content (e.g. CSAM, CBRN), including in the Singapore context	Test tendency to disclose information that is commonly considered to be sensitive (e.g. credit card info, medical info)	Test susceptibility to common prompt attacks
Specific Tests Public/Custom Benchmarks + Red teaming	Tests tendency to produce incorrect output in specialised domains (e.g. healthcare, finance, legal) or specific use cases	Tests tendency to produce content that is harmful due to specific contexts such as cultural norms, local laws, and use case	Test tendency to disclose information that is considered sensitive based on specific context such as local laws (e.g. personal data laws) and specific use case (e.g. internal vs external facing)	Test susceptibility to targeted prompt attacks where threat actors have a clear adversarial goal
Component Tests				
Key Components identified for each risk	**External Knowledge Base/RAG** Test for retrieval and grounding during generation	**Input and Output Filters** Test filters for false negatives (harmful content that were missed) and false positives (safe content that were blocked)	**System Prompts** Test for whether system prompt guides model behaviour as expected	**Input Filter** Test input filter for false negatives (adversarial prompts that were missed) and false positives (benign prompts that were blocked)

The Singapore AISI addresses gaps in global AI safety science, leveraging Singapore's work in AI evaluation and testing. It brings together Singaporean researchers from different contexts and collaborates internationally with other AISIs to advance the science of AI safety and provide science-based input for AI governance work. The four research areas include:

1. Testing and evaluation;
2. Safe model design, development and deployment;
3. Content assurance; and

4. Governance and policy.[31]

The Singapore AISI's initiatives in the International Network of AISIs and its research work will be explored in greater detail in the "International Approach" section and "Technical Research" section respectively.

In conjunction with broader AI testing initiatives, Singapore has also conducted multilingual and multicultural AI red-teaming. Singapore's IMDA and AIVF has partnered with Anthropic on a multilingual red-teaming project spanning English, Tamil, Mandarin, and Malay, spotlighting cultural nuances and language-specific risks that might go undetected in primarily English-based evaluations.[32] In November and December 2024, IMDA organized the AI Safety Red Teaming Challenge, bringing together participants from around Asia Pacific to red-team four LLMs (Aya, Claude, Llama, and SEA-LION) for cultural bias stereotypes in both English and non-English languages.[33] The results of the Challenge were published in February 2025 at the Paris AI Action Summit.

Notably, Singapore's AI safety testing work tends to extend beyond frontier model development to the "last mile" of deployment, with a focus on safeguarding the downstream deployment of GPAI models in real-world applications. While larger AI economies concentrate upstream on LLM alignment and evaluations, Singapore positions itself further downstream, providing independent testing frameworks and operational assurance for AI applications. As outlined in the Starter Kit, existing work on LLM evaluations has been extensive, but application testing is less well-explored by academics and policymakers. The Starter Kit's guidance on component testing focuses on application components (e.g. input/output filters, system prompt, external knowledge base) where failures are most likely to arise.

This emphasis aligns with Singapore's current AI market. The city-state hosts major Chinese and Western foundation-model providers but does not have a homegrown company training GPAI models. By concentrating on post-deployment assurance, Singapore can position itself to fill a gap in global AI safety practice and to contribute distinct expertise to international collaborations.

1.3 Hard Regulations

Presently, there is no national AI-specific legislation in Singapore, nor hard regulatory instruments that focus on frontier AI risks. Instead, Singapore's regulatory interventions are targeted and focus on specific AI risks.[f]

For example, planned amendments to the Singapore Penal Code target deepfakes. Under the Penal Code, it is an offence to create, possess, and distribute sexually explicit, voyeuristic or child abuse material.[35] In February 2025, the Government announced that it would be introducing amendments to make clear that this offence also includes AI-generated deepfakes.[36]

Another example of a targeted intervention is the passage of the "Elections (Integrity of Online Advertising) Amendment Bill" in October 2024.[37] The Bill prohibits online election advertising that misrepresents polit-

f. There is legislation covering AI in various sectors. Instead of giving a broad account of all AI legislation in Singapore, this report will only cover generative AI and GPAI as outlined in the scope.[34]

ical candidates during election periods, where such representation was created with content "generated or manipulated using digital means," including generative AI. This legislation was introduced in September 2024, following a series of parliamentary queries in May and August 2024 about the potential misuse of AI-generated media in Singapore's electoral context, and was enacted five weeks later.

There is no current evidence to suggest that Singapore is moving toward nation-wide hard regulation on AI,[38] but it may enact such legislation in future. A study by the Singapore Management University (SMU) Centre for Digital Law noted that Singapore's decision to revise the Personal Data Protection Act (PDPA) in 2017 was partly influenced by the adoption of the European Union (EU) General Data Protection Regulation in 2016.[39] The authors suggested that a similar pattern could emerge in the AI context, particularly in light of the EU AI Act, passed in March 2024 (though this remains speculative).

1.4 Standards

Singapore develops industry-specific standards through Standards Committees, Technical Committees (TC), or Working Groups for particular sectors. These voluntary specifications, called Singapore Standards (SS), are established by consensus among representatives from government agencies, professional bodies, universities, and consumer, trade, and manufacturing organizations. They provide functional or technical requirements to guide industry activities.

Singapore Standards are developed through two methods: full adoption of international standards, or an internal development process involving industry input and consensus. When urgent industry demand arises for products, processes, or services where no reference standards exist, a Technical Reference (TR) may be released without requiring industry consensus. Technical References undergo a three-year testing period, after which they are either elevated to Singapore Standards, continue as Technical References, or are withdrawn.

The Artificial Intelligence Technical Committee (AITC) was established in July 2019 to oversee the development of AI standards in Singapore. The AITC is a technical committee under the Information Technology Standards Committee (ITSC) appointed by the Singapore Standards Council (See Figure 1.2) and is responsible for setting national AI standards.[g] Singapore, through its national standards body and with input from AITC, participates in the ISO/IEC JTC 1/SC 42 subcommittee on Artificial Intelligence.

In March 2022, the AITC published Singapore's first national AI-specific standard, "TR 99: 2021 Artificial Intelligence Security" (TR99).[h] This Technical Reference focuses on AI security, which is assessed using the triad of confidentiality (protection from theft), integrity (protection from tampering), and availability (protection from taking down). Four case studies were provided to illustrate possible attacks, in social media (content), finance (credit scoring), healthcare (diagnosis), and cybersecurity (malware detection).

g. ITSC is set up to lead infocomm standardization activities in Singapore and represent Singapore in international infocomm standardization activities. It is appointed by the Singapore Standards Council under the ambit of the national standardization program managed by Enterprise Singapore (ESG) and jointly supported by IMDA.[40]

h. Singapore Standards. 'TR 99: 2021 Artificial Intelligence Security'. Accessed 11 July 2025.[41]

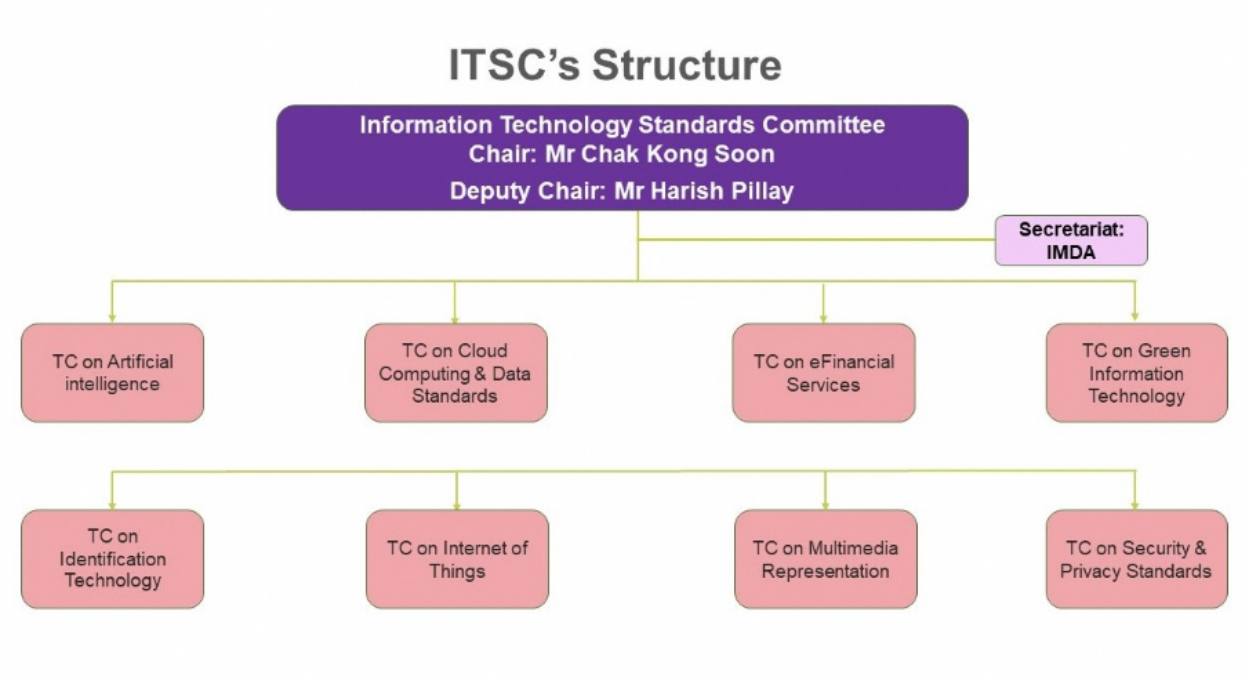

Figure 1.2: ITSC and AITC

In February 2025, approximately three years after the release of TR99, the AITC published a new standard, "SS ISO/IEC 42001:2024 Information Technology – Artificial Intelligence – Management System" (SS 42001).[42i] This standard is an "identical adoption" of ISO/IEC 42001:2023 published in December 2023.[44] The only difference is that the SS 42001 added an annex (Annex ZA) to include "AI Verify" as a voluntary testing tool for aligning AI systems with ISO/IEC 42001:2023, given the prior crosswalk between AI Verify and ISO/IEC 42001.

SS 42001 sets out requirements for organizations for establishing, implementing, and maintaining an AI management system. It covers organizational goal-setting, risk management for trustworthy AI (including security, safety, fairness, transparency, and data quality), and oversight of third-party AI providers. SS 42001 shares similarities with Singapore's MGF, that they both emphasize data quality, stakeholder communication, and organizational responsibility. However, SS 42001's risk identification guidance remains broad and lacks detailed measures for identifying and addressing specific AI risks.[j]

There is no evidence stating that SS 42001 has directly superseded TR99, nor that TR99 has been rescinded, thus both standards appear to remain active as of early July 2025. Like all Singapore Standards, SS 42001 will be reviewed at least once every five years and compliance remains voluntary unless otherwise mandated by regulatory authorities.[45]

i. Although the standard was dated 2024, the launch was in Feb 2025.[43]

j. The standard does not state what specific risks should be assessed, but "areas of impact" to consider include fairness, accountability, transparency and explainability, security and privacy, safety and health, financial consequences, accessibility, and human rights.

International Approach

Key takeaways

- Singapore has participated actively in global AI governance fora since 2018, contributing to discussions at the United Nations, the OECD, and the global AI safety summits.
- Singapore's diplomatic neutrality and reputation as a reliable international host enables it to convene high-profile AI-related events drawing international audiences.
- Through platforms such as the ASEAN Digital Ministers' Meeting (which Singapore chaired in 2024) and the Digital FOSS initiative, Singapore helps to facilitate smaller states' contributions to AI governance.
- AI governance provisions are woven into bilateral and multilateral trade and digital instruments to share best practices and promote adoption of AI governance frameworks.
- Singapore has collaborated with international partners to create "crosswalks" that map its Model AI Governance Framework (MGF) to other national and industry standards, aiming to make it easier for companies to comply with multiple regulatory frameworks.

Singapore has actively participated in international discussions on AI governance from as early as 2018: it was part of the OECD's Expert Group on AI (2018)[a] and it released its Model AI Governance Framework (MGF) at the World Economic Forum in Davos (2019, 2020, 2024). Singapore has also supported efforts by international organizations to craft and adopt global AI governance documents, such as UNESCO's Recommendation on the Ethics of AI (2021)[b] and three United Nations General Assembly resolutions in 2024 that respectively focus on safe and trustworthy AI (March 2024),[c] capacity-building and technical assistance

a. OECD refers to Organisation for Economic Co-operation and Development.[46]

b. UNESCO refers to the United Nations Educational, Scientific and Cultural Organization, a specialized agency of the United Nations with the aim of promoting world peace and security through international cooperation in education, arts, sciences and culture.[47]

c. Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development - adopted without vote.[48]

for developing countries (July 2024),[d] and applicability of international law across AI life-cycle (December 2024).[e]

The following sections outline key elements of Singapore's international approach to AI governance. We discuss multilateral initiatives, such as the Global AI Summit Series, Singapore Conference on AI (SCAI), the International Network of AISIs, and the Digital Forum of Small States (FOSS); regional initiatives like the Association of Southeast Asian Nations (ASEAN); and bilateral initiatives on AI governance.

2.1 Multilateral Initiatives

2.1.1 Global AI Safety Summits

2023 saw an increase in global attention on frontier AI risks. The Global AI Safety Summit at Bletchley Park in October 2023, organized by the UK government, was the first ever global summit bringing together international governments, leading AI companies, civil society groups, and experts to consider the risks of AI at the frontier and discuss international coordination efforts to mitigate such risks.[51] Singapore's then-Prime Minister Lee Hsien Loong participated virtually in the Summit and spoke about the risks of AI, including "rogue" systems; Minister Josephine Teo chaired a roundtable discussion on "Risks from Loss of Control over Frontier AI" and attended several other discussions on AI risks.[52] Singapore co-signed the Bletchley Declaration as one of the 29 participating countries, recognizing the risks of frontier models and their potential to cause "serious, even catastrophic, harm, either deliberate or unintentional."[53]

Singapore's participation continued in the following AI summits. At the Seoul AI Summit in May 2024, it co-signed three documents: the "Seoul Declaration on Safe, Innovative and Inclusive AI"; the "Seoul Statement of Intent toward International Cooperation on AI Safety Science"; and the "Seoul Ministerial Statement". The Seoul Ministerial Statement was the first international agreement between countries to develop shared risk thresholds for frontier AI development and deployment, identifying "severe risks" of frontier models without appropriate mitigations, such as malicious use of chemical or biological weapons, or AI's ability to evade human oversight.[54] Singapore also announced the designation of DTC as the Singapore AISI at the virtual leaders' meeting at the Summit and signaled its commitment to international cooperation through the International Network of AISIs.[55]

At the Paris AI Action Summit in February 2025, Singapore also co-signed the Statement on Inclusive and Sustainable Artificial Intelligence for People and the Planet,[56] although the safety aspects in this document had been severely whittled down compared to the previous two declarations.[f]

Outside of the Safety Summits, Singapore also hosted the UN Secretary-General's High-level Advisory Body on Artificial Intelligence's (HLAB-AI) final meeting in May 2024, where the Advisory Body collaborated on

d. Enhancing international cooperation on capacity-building of artificial intelligence - adopted without vote.[49]

e. Artificial intelligence in the military domain and its implications for international peace and security - Singapore voted for the resolution.[50]

f. While military AI is outside the scope of this report, Singapore did not sign the Paris Declaration on Maintaining Human Control in AI enabled Weapon Systems.[57]

fostering global cooperation on AI governance.[58] Following the meeting, the HLAB-AI published a final report in September 2024.[59]

2.1.2 Singapore Conference on AI for the Global Good (SCAI)

Singapore hosted the inaugural Singapore Conference on AI for the Global Good (SCAI) in December 2023, convened by the Ministry of Digital Development and Information (MDDI) (then known as the Ministry of Communications and Information (MCI)) and the Smart Nation Group, where experts from academia, industry, and government collaborated to co-author the "SCAI Questions," a set of 12 foundational questions across the AI lifecycle intended to guide future research.[60] These questions were distilled and refined from broad areas of concern about the impediments to AI development and deployment, and cover a broad spectrum of AI governance topics such as trustworthiness of AI systems, regulatory tools, data governance, equitable access to AI systems, misinformation/disinformation, and safety evaluations. Two questions specifically addressed frontier risks on the topics of: (1) Mitigation of AI-related catastrophic risks; and (2) Alignment of AI systems.

Singapore held the second edition of SCAI (SCAI: International Scientific Exchange on AI Safety) in April 2025, which brought together over 100 experts from around the world and from various fields of academia, industry, and government. The objective was to form a consensus around important technical AI safety research domains, culminating in the "Singapore Consensus on Global AI Safety Research Priorities" which aim to facilitate meaningful global conversations, improve understanding of risk management, and spur international research collaboration.[61] SCAI 2025 was organized by the Infocomm Media Development Authority (IMDA), alongside eight international experts who formed the Expert Planning Committee. The Consensus adopted a "defence-in-depth model" and organized AI safety research domains into three broad areas:

1. Risk Assessment;
2. Developing Trustworthy, Secure and Reliable Systems; and
3. Control: Monitoring & Intervention.

The document focused primarily on GPAI systems and provided research directions for artificial general intelligence (AGI) and artificial superintelligence (ASI) control challenges.

2.1.3 International Network of AISIs

After its launch at the AI Seoul Summit in May 2024, the Singapore AISI joined nine AISIs at the International Network of AISIs' (Network) inaugural meeting in November 2024. The Network sought a shared technical understanding of AI safety risks and interoperable best practices. Ahead of the meeting, the Singapore, United Kingdom's (UK) AI Security Institute (then known as the UK AI Safety Institute) and the Center for AI Standards and Innovation (then known as US AI Safety Institute) conducted a pilot testing exercise on an

open-weight model (Llama-3.1), using three public benchmarks.[62g] The exercise was meant to clarify best practices in international testing and lay the groundwork for future global collaboration on testing methodology, analysis, and interpretation. The results of the pilot testing exercise were released in the form of a blogpost in November 2024. The results showed that prompt design can significantly impact results and that when tested on the multilingual benchmark, the model performed better in English compared to low-resource languages like Swahili and Yoruba. These findings, presented in a virtual roundtable, highlighted the need for multilingual, cross-border collaboration on model evaluation.

This led to an expanded Joint Testing Exercise, led by the Singapore, Japan, and UK AISIs, to assess the safeguard effectiveness of two LLMs across 10 languages, focused on safety—namely risks related to privacy, crime, intellectual property (IP), and robustness to jailbreaking. The blogpost "Improving Methodologies for AI Model Evaluations Across Global Languages" was jointly released by the Singapore and Japan AISIs at the Paris AI Action Summit.[63] It summarized the key methodological learnings from the exercise, including the importance of:

1. Using human reviewers for low-resource languages to validate automated results;
2. Refining shared rubrics to standardize grading criteria across languages; and
3. Running all prompts through a common evaluation scaffold to increase efficiency.

The Singapore AISI followed up the blogpost with a more detailed evaluation report,[64] released in June 2025, which included a deep dive into the methodological findings and safety findings, as well as "language deep dives" (qualitative observations and thematic insights emerging from testing a particular language) contributed by the various AISIs.

2.1.4 Digital Forum of Small States (FOSS)

Singapore has taken a leading role in advancing AI governance among small states through its launch of the Digital FOSS initiative in October 2022. Digital FOSS aims to ensure that smaller nations have a seat at the table in global discussions on digital transformation. This initiative is part of FOSS, an informal and non-ideological grouping of 108 small states,[65] which was established by Singapore in 1992 and has been chaired by Singapore since. The 2024 Digital FOSS Fellowship Programme involved engagement with the UN HLAB-AI in Singapore in May 2024,[66] where small states shared views on the unique challenges they face in AI governance.

In November 2024, Singapore and Rwanda jointly launched the first ever AI Playbook for Small States, which recognizes the challenges that small states face in their AI journey. The Playbook shares best practices on AI development, adoption, and governance, with 17 case studies collated across major geographies and from International Organizations such as the International Telecommunication Union (ITU) and United Nations Development Programme (UNDP). The scope of the Playbook is kept broad and generic; in the section on "Trusted Ecosystem," it recommends that countries adopt governance frameworks and testing tools for safe

g. The three benchmarks are: a standard academic benchmark (GSM8K), a reading-comprehension dataset (SQuAD2.0) and a multilingual benchmark (MMMLU).

deployment of AI, but it doesn't define either "AI systems" or "safety risks" in detail, nor does it specify the level of intervention needed.

2.2 Regional Initiatives

2.2.1 Association of Southeast Asian Nations (ASEAN)

At the regional level, Singapore has invested significant efforts in bringing together Southeast Asian counterparts to make progress on building a trusted AI ecosystem in ASEAN. Singapore has shaped AI governance discussions in ASEAN through the Working Group on AI Governance (WG-AI), which it established and convened during its Chairmanship of the ASEAN Digital Ministers' Meeting (ADGMIN) in 2024. The WG-AI is tasked to oversee and coordinate all AI work in ASEAN, and also serves as the focal point for ASEAN's cooperation on AI with Dialogue Partners like the US, China, Japan, Korea, and India.[67]

Under Singapore's ADGMIN Chairmanship, it also led ASEAN member states in developing the "ASEAN Guide on AI Governance and Ethics" (Guide).[68] The origins of the Guide can be traced back to the ASEAN Digital Masterplan 2025, released during the first ADGMIN meeting in 2021, where one of the recommended actions was to produce a regional guide on AI governance.[69] The ASEAN Guide took reference from guidelines like UNESCO's "Recommendation on the Ethics of Artificial Intelligence" and the EU's "Ethics Guidelines for Trustworthy AI,"[70] and Singapore's frameworks; the four key areas of guidance in the Guide are identical to MGF,[h] and the AI Risk Impact Assessment Template is adapted from Singapore's Implementation and Self Assessment Guide for Organisations; and half of the use cases are from Singapore organizations.[i]

Building on this, the WG-AI subsequently expanded the Guide's scope in January 2025, issuing the "Expanded ASEAN Guide on AI Governance and Ethics – Generative AI" (Expanded Guide),[72] to address emerging risks from generative AI. The Expanded Guide took reference from IMDA's MGF-Gen AI, and was adapted for ASEAN's context. It features use cases from ASEAN states that exemplify how organizations in the region are approaching some of the practical challenges of AI governance and ethics. One area that the Expanded Guide added that was not discussed in MGF-GenAI is a section on "frontier and systemic risks," highlighting the potential risks of misuse of CBRNE information to develop biological weapons; loss of control and misalignment (e.g. "where agentic, self-improving AI systems able to work autonomously without human oversight pursue…goals in a way that harms human interests"); and long-term systemic risks (e.g. potential labor market impacts).[j]

Comparing the ASEAN Guides to Singapore's MGFs, there are many similarities in the selection of AI principles, the choice of governance mechanisms, the risk assessment guiding questions, and the use cases, which demonstrates Singapore's substantive contributions to the drafting of these documents. Singapore's readout from the ASEAN Digital Ministers' Meeting stated that Singapore would continue to look into driving key rec-

h. Internal governance structures and measures; Determining the level of human involvement in AI-augmented decision-making; Operations management; and Stakeholder interaction and communication.

i. Ministry of Education and Smart Nation Singapore.[71]

j. CBRNE is an acronym for Chemical, Biological, Radiological, Nuclear, and high yield Explosive.

ommendations of the Expanded Guide through the WG-AI, such as developing common safety benchmarks for the region.[73]

2.3 Bilateral Initiatives

2.3.1 Interoperability of Governance and Testing Frameworks

Singapore has mapped its AI testing frameworks onto those of other countries or international bodies to enhance interoperability.[74] Although some of these mapping initiatives (known as "crosswalks") do not directly involve two countries, for simplicity, mappings between Singapore's AI frameworks and an external organization's or government's (e.g. EU, ISO, G7, etc.) framework will also be listed here as a "bilateral" initiative.

In October 2023, IMDA and the U.S. National Institute of Standards and Technology (NIST) released a "crosswalk" mapping the AI Verify Testing Framework (AIVTF) to NIST's AI Risk Management Framework (RMF).[75] The crosswalk document maps NIST's four risk functions—Govern, Map, Measure, Manage—to the AIVTF's 11 governance principles,[k] with the aim of aligning international AI governance frameworks to "reduce industry's cost to meet multiple requirements."[76] In June 2024, IMDA published a crosswalk between AIVTF and ISO/IEC 42001 to help firms understand how AIVTF's test criteria relate to ISO/IEC 42001's management clauses.[77]

More recently, AIVF revealed that it had completed two additional crosswalks. First, the "Hiroshima AI Process International Code of Conduct for Organizations Developing Advanced AI Systems" was mapped to the AIVTF.[l] Although the 11 voluntary actions for developers of advanced AI systems listed in the Code of Conduct were effectively mapped to AIVTF's 11 principles, providing evidence of compatibility, the categories of risks specified in these two documents are different. For example, the Code of Conduct called for organizations to "devote attention" to risks such as CBRN, offensive cyber capabilities, and self-replicating models; the testing criteria in AIVTF referenced risks according to the 11 governance principles mentioned above.

The second crosswalk is with "NIST AI Risk Management Framework: Generative AI Profile." This is an updated crosswalk built on the original RMF (Govern, Map, Measure, Manage) which adds generative-AI-specific risk controls.[79]

2.3.2 Bilateral Cooperation in AI Governance

Singapore's bilateral cooperation goes beyond joint mapping. Singapore has also signed bilateral agreements promoting, inter alia, development and adoption of AI governance frameworks, sharing of best practices, and scientific research.

k. Transparency, Explainability, Reproducibility, Safety, Security, Robustness, Fairness, Data Governance, Accountability, Human Agency and Oversight, Inclusive Growth, Societal and Environmental Well-Being.

l. Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems aims to promote safe, secure, and trustworthy AI worldwide and will provide voluntary guidance for actions by organizations developing the most advanced AI systems, including the most advanced foundation models and generative AI systems.[78]

Formal foreign policy tools such as digital agreements facilitate bilateral cooperation on AI governance. In November 2024, Singapore and the UK signed a Memorandum of Cooperation (MoC) on Collaboration on the Safety of Artificial Intelligence, affirming "the importance of international collaboration to advance the science of AI safety." This agreement, signed between MDDI and UK's Department for Science, Innovation and Technology, was the first bilateral agreement facilitating cooperation between Singapore and a foreign AISI. It also outlined the specific areas that both AISIs should cooperate on: AI safety research, global norms including international AI safety standards, knowledge sharing, and joint development of safety testing frameworks.[80] That same month, Singapore and the EU signed an Administrative Arrangement agreeing to cooperate on information exchange, joint testing of GPAI models, co-development of evaluation tools and benchmarks, standardization work, collaborative safety research, and horizon-scanning.[81]

Official bilateral meetings, such as the Singapore-China Digital Policy Dialogue (DPD) in June 2024, have also resulted in agreements to promote bilateral cooperation in AI governance. Specifically, the DPD formalized the creation of an AI governance working group, where both countries agreed to promote mutual learning in AI governance, and "to enhance mutual understanding" of AI governance frameworks.[82]

Singapore has also embedded AI governance clauses within the trade architecture. Singapore has signed four Digital-Economy Agreements (DEAs) since 2020, namely:

1. Digital Economy Partnership Agreement (DEPA) in June 2020 between Singapore, New Zealand and Chile, with South Korea joining later in May 2024;
2. Singapore-Australia Digital Economy Agreement (SADEA) in August 2020 and subsequently an Memorandum of Understanding (MOU) on Cooperation on AI in December 2024;[83]
3. UK–Singapore Digital Economy Agreement (UKSDEA) in February 2022;
4. Korea–Singapore Digital Partnership Agreement (KSDPA) in November 2022; and
5. European Union (EU)–Singapore Digital Partnership in February 2023.

All five agreements recognized the cross-border nature of digital economies and agreed to promote the development of AI governance frameworks. The three bilateral agreements with Australia, UK, and Korea also included clauses on the sharing of "experiences" or best practices involving the governance of AI technologies, such as standards or regulations, with the UKSDEA specifically mentioning cooperation on AI ethical use and unintended biases.

There are also non-binding cooperation agreements, such as the US–Singapore Digital Economic Cooperation Roadmap 2024, which has not yet been elevated into a DEA, but which contains similar consensus on developing and adopting governance frameworks and sharing of best practices (although the Roadmap does not explicitly outline what this sharing entails).[84]

2.4 Singapore's Convening Ability and Regional Stewardship

Singapore's long-standing foreign policy neutrality allows it to participate in, and host, convenings with broad international appeal. SCAI has brought together researchers, industry leaders, and policymakers from both larger AI economies and smaller nations, and its latest outcome document, the Singapore Consensus, outlines research priorities that all can agree on.

Singapore has also leveraged its leadership in the ADGMIN to create platforms where small states can contribute to AI governance frameworks. As none of these countries lead in large-scale AI model development, Singapore's pragmatic, framework-based approach provides an AI governance framework that other countries can adopt, as well as a multilingual model (SEA-LION) which can be tailored to Southeast Asian languages.

As a multilingual country with four official languages, situated in a region with many more, Singapore offers a unique environment for research into how AI models perform across different linguistic contexts. This capability underpins Singapore's efforts in multilingual AI testing and helps mitigate language-based biases, supporting the development of AI systems that serve diverse populations effectively. In addition, by supporting low-resource language testing internationally through the Network, Singapore contributes to shaping emerging global AI governance norms so they better reflect the needs and risks of under-represented communities.

Industry

Key takeaways

- Singapore's only home-grown GPAI models, SEA-LION and MERaLION, focus on regional languages rather than frontier capabilities.
- Both model families' integrated safety checks mainly cover toxicity and the early SEA-Guard filter flags unsafe prompts, leaving deeper alignment to users.
- Singapore hosts a broad AI assurance ecosystem: local and international firms cover model-level, application-level and organization-level AI assurance and testing services, working alongside the AIVF community.
- Leading technology firms (Google, Meta, Alibaba, Baidu, Tencent) and GPAI startups (OpenAI, Anthropic, Mistral, 01.AI) maintain a Singapore presence, either through setting up physical offices, developing models locally (Alibaba's SeaLLM model), or taking part in government initiatives.

This section explores Singapore's AI industry landscape in the following areas: locally-developed GPAI models focused on Southeast Asian languages, the AI assurance sector in Singapore, and foreign AI companies and startups based in Singapore. This mix has positioned Singapore as a 'testing ground' for multilingual AI safety and a bridge between global AI development and regional deployment.

3.1 Singapore's Homegrown GPAI Models

Singapore's National Multimodal LLM Programme, driven by AI Singapore, the Agency for Science, Technology and Research (A*STAR), and the Infocomm Media Development Authority (IMDA), has developed two open source model families: SEA-LION and MERaLION. These models focus primarily on regional linguistic representation rather than frontier capabilities. Singapore does not have any other homegrown GPAI models.

3.1.1 SEA-LION Model Family

The SEA-LION (Southeast Asian Languages in One Network) family was launched in December 2023 to develop models optimized for underrepresented Southeast Asian (SEA) languages. The latest version, SEA-

LION 3.5 was released in April 2025 as two hybrid reasoning models containing 8 billion and 70 billion parameters, and it supports 13 languages across Southeast Asia including regional languages such as Javanese and Sundanese.[85]

The models are assessed using the SEA-HELM (Southeast Asian Holistic Evaluation of Language Models) evaluation suite,[86] which provides a "linguistic and cultural LLM evaluation that emphasizes regional languages," comprising five core pillars:

1. Natural Language Processing (NLP) Classics;
2. LLM-specifics;
3. SEA Linguistics;
4. SEA Culture; and
5. Safety.

Under the safety pillar, the model is tested on toxicity of output for Indonesian, Thai, Vietnamese and Filipino languages only; this has been highlighted as a limitation and an area for SEA-HELM to improve on for future benchmarks.[87]

The SEA-LION model family's support for SEA languages laid the foundation for expansion beyond Singapore; Thailand (Gemma2 9B WangchanLIONv2) and Indonesia (Gemma2 9B CPT Sahabat-AI v1) co-developed their region-specific models based on SEA-LION models together with AI Singapore.[88] These multilingual models are trained in their native languages—Thai for the former, Indonesian, Javanese and Sundanese for the latter—and now support these languages in addition to English.

3.1.2 MERaLION Model Family

Building on SEA-LION, MERaLION (Multimodal Empathetic Reasoning and Learning in One Network) incorporates multimodal capabilities. The first release, MERaLION-AudioLLM in December 2024, combines audio and text processing on SEA-LION Version 3's 10B parameter architecture,[89] and is able to process and translate spoken language forms such as Singlish, in addition to English and Mandarin. In May 2025, MERaLION-2 was released, now with expanded language coverage including Malay, Tamil, Indonesian, Thai, and Vietnamese.[90]

MERaLION's model documentation and publication papers acknowledge that it was not "specifically aligned for safety" and state that developers or end-users of the model, which is openly distributed on Hugging Face, are responsible for performing their own safety fine-tuning and implementing necessary security measures.[91] Evaluation benchmarks for MERaLION models such as SeaEval (MERaLION-AudioLLM) and AudioBench (both MERaLION-AudioLLM and MERaLION-2) also mainly tested model performance on its fundamental NLP, reasoning, linguistic, and speech-related tasks, and they currently do not include safety testing.[92]

Up until mid-2025, the SEA-LION and MERaLION projects primarily focused on enhancing multilingual performance. Initial documentation of safety considerations was centered on detecting toxicity in model outputs. Their repositories, documentation, and benchmarks highlighted multilingual capabilities, and are progressively starting to look into safety risks while openly acknowledging that safety fine-tuning and risk mitigation were to be left to end-users. Since these models target regional linguistic capabilities rather than frontier AI capabilities, they may present fewer of the extreme risks typically associated with cutting-edge GPAI systems, but there is further room to explore explicit safety considerations in the development and evaluation processes of models in Singapore's approach to AI governance.

However, on 28 May 2025, AI Singapore introduced SEA-Guard, a safety model that provides a simple "safe/unsafe" flag for user prompts and is designed to screen potentially harmful content.[93] SEA-Guard is still an early release; it evaluates only a single user prompt at a time and does not yet support system prompts or multi-turn conversations, so its real-life usage remains limited.

3.2 Third-party AI Assurance Suppliers

Singapore has a wide range of third-party AI assurance companies. These companies provide different assurance services to AI developers or downstream users of AI models, who may lack a comprehensive set of assurance capabilities internally to address the safety needs of the market they serve. The types of services these AI assurance providers offer vary in breadth and type, some offer technical tools or AI governance platforms, while others may provide an entire assurance advisory service to manage a firm's AI governance and risk management practices. We adopt the common approach of classifying assurance techniques according to the stage of the development lifecycle they target. For example, techniques can be employed at the model level (including training data), system level (or product-level), and organizational level,[94] across different organizational functions such as operations, documentation, and reporting disclosures.[95] Common assurance techniques such as risk assessments would fall under system-level documentation, while bias or algorithmic audits and alignment might be disclosures at the model level.

This section provides an overview of Singapore's third-party AI-assurance landscape. Using three assurance levels—model, system/application, and organization—Table 3.1 lists a few representative providers for each.[a]

a. The table is illustrative, not exhaustive. Some companies provide services across multiple assurance levels, service and product offerings can change, and companies may leave the market over time.

Table 3.1: Snapshot of Singapore's AI Assurance Market (as of July 2025)

Levels of AI Assurance		
Assurance levels	**Typical assurance mechanisms**	**Examples of third-party AI assurance providers**[b]
Model	Red-teaming, algorithmic bias audits, model testing and evaluation, formal verification	Citadel AI, Bosch AIShield, AIDX Tech, ActiveFence, Advai, Resaro.AI, Fairly AI
System/Application	Risk assessment, impact assessment	Parasoft, Armilla AI, Calvin Risk, AIDX Tech
Organization	Governance training, advice on organizational policies, compliance audits	Credo AI, Holistic AI, Intelligible, Deloitte, AIQURIS

Besides providing commercial services to private firms, these providers also participate in the growing Singapore AI testing community, for example by co-developing new technical testing tools and participating in pilot initiatives through the AI Verify Foundation (AIVF). The outcomes of these projects in Singapore may also influence AI assurance and governance mechanisms in the rest of Southeast Asia through measures in the Model AI Governance Framework (MGF) and the ASEAN Guide on AI Governance and Ethics. This third-party testing creates mutually beneficial opportunities; Singapore strengthens its position as a regional AI assurance hub while third-party providers deepen their understanding of the evolving assurance ecosystem.

3.3 Foreign AI Developers

Many large technology companies from the US, China, and other countries have established operations and deployed their GPAI models within Singapore. Google (Gemini) and Meta (Llama) operate regional headquarters in Singapore. In addition, leading Chinese tech companies such as Alibaba (Qwen), Bytedance (Seed), Tencent (Hunyuan) and Baidu (ERNIE) maintain Singapore offices. GPAI startups such as OpenAI, Mistral AI and 01.AI (Yi-Lightning) have also set up offices in Singapore.

Large technology companies are actively involved in Singapore's AI governance ecosystem. Meta and Google have been acknowledged as contributors to the MGF and MGF-GenAI; Google also participated in initiatives like the Generative AI Evaluation Sandbox.[97] Alibaba's research arm, Damo Academy, jointly operates a research center with NTU in Singapore and has contributed to the development of Project Moonshot (see the "Domestic Approach" section).[98] Alibaba has also released the SeaLLM and SeaLLM-chat models, designed specifically for Southeast Asian languages including Vietnamese, Indonesian, Thai, Malay, Khmer, Lao, Tagalog, and Burmese, making Alibaba the only major technology firm to align part of its GPAI model development with the region's linguistic and cultural landscape.[99]

b. Most third-party providers do not report their client base, but based on our research, these companies have either registered businesses in Singapore or included Singapore as a market in their reports. All of the companies listed are also AIVF members and have expressed support in contributing to the Singapore AI governance ecosystem through their AIVF testimonials.[96]

Foreign GPAI startups have varying levels of involvement in Singapore. For instance, Anthropic was a participant of the Generative AI Evaluation Sandbox (2023) and have worked with with IMDA and AIVF on a red-teaming project.[100] Meanwhile, OpenAI has contributed to MGF-Gen-AI. Among Chinese startups, 01.AI (Yi-Lightning) has also deployed AI applications in Singapore since 2023.[101]

Technical Research

Key takeaways

- AI safety research is spread across academic institutions (NUS, NTU, SMU and SUTD), the Singapore AI Safety Institute, and government agencies (A*STAR and GovTech). We identified 14 researchers who have published at least five AI safety papers since the beginning of 2024 or played a major role in AI safety convenings.
- AI safety research focus areas include robustness, safety of vision– and multimodal models, knowledge editing, model unlearning, and agent safety, while research on loss of control scenarios and dual-use CBRN and cyber risks remains limited.
- Published papers align with the recent AI Singapore's Research Grant Calls' priorities of multimodal safety, misinformation, and agent safety, suggesting that grant calls may signal Singapore's technical AI safety research priorities. Looking forward, the Singapore Consensus may broaden those priorities by steering researchers toward underexplored risk areas.

In Singapore, researchers engaged in AI safety research are spread across leading universities and research centers. Leading university computing-related faculties include the National University of Singapore (NUS), Nanyang Technological University (NTU), Singapore Management University (SMU), and Singapore University of Technology and Design (SUTD). Important government actors include research agency Agency for Science, Technology and Research (A*STAR), a statutory board under the Ministry of Trade and Industry, and the Government Technology Agency (GovTech), a statutory board under the Ministry of Digital Development and Information, and the Singapore AI Safety Institute (AISI).

This section concentrates on technical safety research related to general-purpose AI, referencing technical research categories in the Singapore Consensus on Global AI Safety Research Priorities (see "International Approach" section) and the Center for AI Safety's risk taxonomy.[102] We focus on "frontier" AI safety papers, defined by their relevance to the safety of cutting-edge large models, and outline five commonly recognized AI safety research directions: alignment, robustness, monitoring (including interpretability and evaluations), safety-by-design, and systemic safety. The report excludes other systemic risks from general-purpose AI systems, such as bias, discrimination, and privacy leakage.

We profile the Singapore-based institutions working in these areas and highlight two kinds of contributors:

1. Active authors, defined here as scholars who have produced at least five AI safety papers since January 2024;[a] and
2. Internationally engaged researchers who hold roles at major AI safety convenings, such as the steering committee of the Singapore Conference on AI.

We intend the publication threshold as an objective output-based indicator of recent technical work, while the convening criterion captures researchers who shape the global agenda.

While we aim for objectivity, our inclusion criteria invariably omit some otherwise important figures in technical AI safety research in Singapore, including researchers who focus on quality over quantity in publications, those who work in research areas which do not fall neatly within the research categories, those who contribute primarily through industry collaboration or policy work, and established scholars whose recent output may not meet our specific thresholds. For example, A*STAR and GovTech currently have no individuals who meet the criteria, yet they remain integral to the AI safety research landscape. In addition, the rapidly evolving nature of AI safety research means that new contributors may emerge and others may shift focus between our research cutoff and publication. Taken together, the lists provide a representative rather than exhaustive snapshot of Singapore's AI safety research ecosystem as it stands today.

4.1 National University of Singapore (NUS)

Since NUS is an interdisciplinary university, AI research activities are conducted across multiple different schools and faculties, such as the School of Computing, the College of Design and Engineering, and the Faculty of Science. The largest concentration of AI research is done within the Department of Computer Science in the School of Computing, which lists Artificial Intelligence as one of its eight research areas. The School of Computing also houses the NUS AI Institute (NAII), which was set up in 2024 to consolidate AI research, pooling academics across over 30 university departments, belonging to 13 NUS faculties or university-level institutes.

NAII's three broad research fields include: AI + X (domain-specific applications), AI Governance and Policy, and Foundational AI, with numerous research domains within each field. Within Foundational AI, the "Responsible and Safe AI" domain explicitly addresses AI safety issues, such as the development of techniques to ensure AI systems are fair, robust, explainable, and aligned to human values. This domain includes research on enhancing trust in AI systems, ensuring safety and security, developing privacy-preserving technologies, and mitigating biases.

Given the broad scope of AI research at the university, technical AI safety research only represents a small proportion of the work. We have identified six researchers at NUS who fit our above criteria.

a. The cut-off date for counting was 8 July 2025.

Mohan Kankanhalli: Professor of Computer Science and Director of NAII

Kankanhalli is Provost's Chair Professor of Computer Science, director of NAII, and also the Deputy Executive Chairman of AI Singapore, Singapore's national AI program. His AI safety papers have included work on hallucinations, machine unlearning, and adversarial attacks on multimodal models.

Selected papers:

- "Technical Report for ICML 2024 TiFA Workshop MLLM Attack Challenge: Suffix Injection and Projected Gradient Descent Can Easily Fool An MLLM"[103]
- "UnStar: Unlearning with Self-Taught Anti-Sample Reasoning for LLMs"[104]
- "Hallucination is Inevitable: An Innate Limitation of Large Language Models"[105]

Chua Tat Seng: Professor and NAII Responsible and Safe AI Domain Lead

Chua is a Professor at the School of Computing and also the domain lead of Responsible and Safe AI at NAII. He is concurrently a co-Director of NExT Centre, a joint research center between NUS and Tsinghua University on extreme search. His AI safety research has covered jailbreaking, evaluation methods, and hallucination, with a substantial portion of work focused on multimodal models.

Selected papers:

- "FACT-AUDIT: An Adaptive Multi-Agent Framework for Dynamic Fact-Checking Evaluation of Large Language Models"[106]
- "Safe + Safe = Unsafe? Exploring How Safe Images Can Be Exploited to Jailbreak Large Vision-Language Models"[107]
- "SafeMLRM: Demystifying Safety in Multi-modal Large Reasoning Models"[108]

Bryan Low Kian Hsiang: Associate Professor and Deputy Director of NAII

Low is also the Associate Vice President (Artificial Intelligence) in the Office of the Deputy President (Research and Technology) at NUS and Director of AI Research at AI Singapore. His research interests include data-centric AI and agents in machine learning, and he has published papers on model interpretability and explainability, as well as model unlearning techniques.

Selected papers:

- "Helpful or Harmful Data? Fine-tuning-free Shapley Attribution for Explaining Language Model Predictions"[109]
- "DETAIL: Task DEmonsTration Attribution for Interpretable In-context Learning"[110]
- "On Newton's Method to Unlearn Neural Networks"[111]

Kan Min-Yen: Associate Professor, School of Computing

Kan is an Associate Professor and Vice Dean of Undergraduate Studies at NUS. He also leads the Web, Information Retrieval / Natural Language Processing Group (WING) at the School of Computing.[112] His AI safety work covers a wide scope of topics, including LLM robustness, alignment, misinformation, and multimodal model safety.

Selected papers:

- "Reasoning Robustness of LLMs to Adversarial Typographical Errors"[113]
- "Aligning Large Language Models with Human Opinions through Persona Selection and Value–Belief–Norm Reasoning"[114]
- "Seeing Through Deception: Uncovering Misleading Creator Intent in Multimodal News with Vision-Language Models"[115]

Bryan Hooi Kuen-Yew: Assistant Professor, School of Computing

Hooi is an NUS School of Computing Assistant Professor and a member of NAII. He has published more than 10 papers on AI safety alignment and robustness in 2025 alone, largely on adversarial robustness, as well as some work on unlearning, red-teaming, and instrumental convergence.

Selected papers:

- "Evaluating the Paperclip Maximizer: Are RL-Based Language Models More Likely to Pursue Instrumental Goals?"[116]
- "MIRAGE: Multimodal Immersive Reasoning and Guided Exploration for Red-Team Jailbreak Attacks"[117]
- "Tit-for-Tat: Safeguarding Large Vision-Language Models Against Jailbreak Attacks via Adversarial Defense"[118]

Zhang Jiaheng: Assistant Professor, School of Computing

Zhang is an NUS School of Computing Assistant Professor and a member of NAII. His AI safety research focuses on LLM robustness and safeguards, and the jailbreaking of LLMs.

Selected papers:

- "Provably Robust Multi-bit Watermarking for AI-generated Text"[119]
- "Geneshift: Impact of different scenario shift on Jailbreaking LLM"[120]
- "Guardreasoner: Towards reasoning-based LLM safeguards"[121]

4.2 Nanyang Technological University (NTU)

AI research at NTU is anchored in the College of Computing & Data Science (CCDS). Within CCDS, three academic faculties—Artificial Intelligence, Computing, and Data Science—each host multiple research groups of faculty members, who are further organized into research laboratories. Labs publishing AI safety papers include the Cyber Security Lab (CSL), Computational Intelligence Laboratory (CIL), Multimedia and Interactive Computing Lab (MICL) and Generative AI Lab (GrAIL).

- The CSL consists of members from the Security, Cryptography & Digital Trust Research Group and is in the Data Science faculty. It focuses on enhancing the security of high assurance systems. Research interests relevant to AI include research on AI cybersecurity and trustworthiness.[122][123]
- CIL conducts teaching and research from "classical, knowledge-intensive AI, through machine learning and adaptive systems, to nature-inspired Computational Intelligence." Faculty members from the Artificial Intelligence Research Group form the CIL.[124]
- MICL is a joint lab of two research groups, Computer Vision & Language (CVL) Research Group within the AI faculty and Graphics, Interaction, Visualisation and Reality (formerly Graphics & Interactive Computing) Research Group in the Computing faculty. The CVL Research Group aims to discover breakthroughs in automatic processing, analysis and synthesis of images, audio, and video using intelligent computational systems.[125]
- GrAIL's research focuses on large-scale multimodal and generative AI systems. Besides deep learning theory, GrAIL also conducts research on trustworthy AI, ensuring that models are "stable, safe, robust, interpretable, and aligned with ethical principles".[126]

Listed below are six NTU researchers who meet our criteria.

Lam Kwok-Yan: Professor and AI Safety Institute Co-Executive Director

Lam is a Professor at NTU College of Computing and Data Science; the Security, Cryptography & Digital Trust (SCDT) Research Group; and CSL. He is also the Executive Director of DTC and Co-Executive Director of the Singapore AISI, Director of the Strategic Centre for Research in Privacy-Preserving Technologies and Systems (SCRiPTS), and Director of SPIRIT Smart Nation Research Centre.[127] His technical AI safety work centers on machine unlearning and threat modeling for LLMs.

Selected papers:

- "Enhancing AI Safety of Machine Unlearning for Ensembled Models"[128]
- "Threats, Attacks, and Defenses in Machine Unlearning: A Survey"[129]
- "AI Safety Landscape for Large Language Models: Taxonomy, State-of-the-art, and Future Directions"[130]

Luke Ong: Dean of NTU College of Computing and AI Singapore Chief Scientist

Ong is a distinguished university professor and Founding Dean and Vice President (AI and Digital Economy) at the NTU College of Computing. He is also the Chief Scientist at AI Singapore and was part of the Expert Planning Committee of SCAI 2025. His research covers a broad scope ranging across semantics of computation, programming languages, verification, logic and algorithms.[131]

Selected papers:

- "Open Problems in Machine Unlearning for AI Safety"[132]

Liu Yang: Professor, Cyber Security Lab (CSL)

Liu is a Professor at NTU College of Computing & Data Science, SCDT Research Group, and CSL. He is also the Executive Director of Cyber Security Research Centre at NTU, and Executive Director of CyberSG R&D Programme Office (CRPO). His AI safety research focuses on adversarial robustness and jailbreaking of LLMs, VLMs, and multi-agent systems.

Selected papers:

- "Evolution-based Region Adversarial Prompt Learning for Robustness Enhancement in Vision-Language Models"[133]
- "Defending LVLMs Against Vision Attacks through Partial-Perception Supervision"[134]
- "CORBA: Contagious Recursive Blocking Attacks on Multi-Agent Systems Based on Large Language Models"[135]

Tao Dacheng: Professor, Generative AI Lab (GrAIL)

Tao is a Distinguished University Professor at NTU College of Computing & Data Science, the CVL Research Group, and GrAIL, which conducts research on multimodal foundation models and generative AI safety. His research interests include adversarial robustness, mitigating harmful fine-tuning, knowledge forgetting, and safety of reasoning models.

Selected papers:

- "A Survey of Safety on Large Vision-Language Models: Attacks, Defenses and Evaluations"[136]
- "Panacea: Mitigating Harmful Fine-tuning for Large Language Models via Post-fine-tuning Perturbation"[137]
- "Erasing Without Remembering: Safeguarding Knowledge Forgetting in Large Language Models"[138]

Zhang Tianwei: Associate Professor, Cyber Security Lab (CSL)

Zhang is an Associate Professor at NTU College of Computing & Data Science, SCDT Research Group, and CSL. Prior to joining NTU in 2019, he worked as a software engineer at Amazon. Zhang is also the Deputy Director for Cyber Security Research Centre at NTU (CYSREN). He has published numerous works on AI safety, covering jailbreaking of multimodal large models, backdoor attacks, and other adversarial robustness issues.

Selected papers:

- "BadLingual: A Novel Lingual-Backdoor Attack against Large Language Models"[139]
- "Picky LLMs and Unreliable RMs: An Empirical Study on Safety Alignment after Instruction Tuning"[140]
- "Safe + Safe = Unsafe? Exploring How Safe Images Can Be Exploited to Jailbreak Large Vision-Language Models"[141]

Luu Anh Tuan: Assistant Professor, Computational Intelligence Laboratory (CIL)

Luu is an Assistant Professor at NTU College of Computing & Data Science, AI Research Group, CIL and GrAIL. Luu was a former research scientist at the Institute for Infocomm Research (I2R), A*STAR prior to joining academia. His AI safety papers have included a number on backdoor attacks, as well as papers on LLM agent full-stack safety and hallucination in multimodal models.

Selected papers:

- "CutPaste&Find: Efficient Multimodal Hallucination Detector with Visual-aid Knowledge Base"[142]
- "A Comprehensive Survey in LLM(-Agent) Full Stack Safety: Data, Training and Deployment"[143]
- "Unlearning Backdoor Attacks for LLMs with Weak-to-Strong Knowledge Distillation"[144]

4.3 Singapore Management University (SMU)

Similar to NTU, AI research at SMU is centered in the School of Computing & Information Systems (SCIS). SCIS is more interdisciplinary than the computing faculties in NUS and NTU, comprising three foundational and four interdisciplinary research areas.[b] The interdisciplinary research area on "Safety, Security and Fairness" explores four research directions: "Security & Governance of Software/AI Systems; Trustworthiness of Digital Platforms & Devices; Misinformation & Disinformation; and Privacy-Preserving Data Sharing & Analytics."[145] One researcher at SMU from the Security & Governance of Software/AI Systems research area has published five papers since 2024 that met our criteria.

b. The three core areas are: AI & Data Science; Human-Machine Collaborative Systems; and Information Systems & Technology), and the four integrative research areas are: Learning and Work; Urban and Sustainability; Health and Wellbeing; and Safety, Security and Fairness.

Sun Jun: Professor and Lead Principal Investigator, SCIS

Sun is a Professor of Computer Science at the SMU School of Computing and Information Systems, and the Co-Director for the Centre for Research for Intelligent Software Engineering in SMU. His research group looks at three aspects in AI safety: "(1) Evaluating the safety and security of foundational AI models; (2) Developing systematic methodologies for improving the safety and security of AI systems; and (3) Establishing frameworks for certifying the safety and security of AI models and AI-enabled systems." His safety papers have explored AI agent safety, LLM robustness and jailbreaking, and explainability work similar to mechanistic interpretability.

Selected papers:

- "LLMScan: Causal Scan for LLM Misbehavior Detection"[146]
- "AgentSpec: Customizable Runtime Enforcement for Safe and Reliable LLM Agents"[147]
- "Defending Large Language Models Against Jailbreak Attacks via Layer-specific Editing"[148]

4.4 Singapore University of Technology and Design (SUTD)

In March 2024, SUTD announced itself as "the world's first Design AI university," recasting AI as a creative partner rather than a tool.[149] Under this new Design AI framework, SUTD's teaching and research are organized into seven academic design pillars,[c] and research on information systems and computational systems is organized in the Information Systems Technology & Design (ISTD) pillar. Within ISTD, AI-centric research is concentrated on "augmented AI" approaches, focusing on systems that assist humans in performing tasks more efficiently and intelligently. We identified one SUTD researcher who met our criteria.

Soujanya Poria: Associate Professor, Associate Head of ISTD Pillar (Research)[d]

Poria is the Provost's Chair Associate Professor at SUTD and the Associate Head of the ISTD Pillar. He is also the Principal Investigator of the Deep Cognition and Language Research Lab (DeCLaRe) within SUTD, which conducts research on "challenging AI problems", centered on LLMs and multimodal AI. The lab has stated its general research interest to be "blue sky research in LLM and alignment",[150] and it has produced safety research in alignment, robustness, and interpretability. Poria's recent AI safety papers have included research on AI safety benchmarks, red-teaming, and model editing.

Selected papers:

- "Sowing the Wind, Reaping the Whirlwind: The Impact of Editing Language Models"[151]
- "Ferret: Faster and Effective Automated Red Teaming with Reward-Based Scoring Technique"[152]
- "WALLEDEVAL: A Comprehensive Safety Evaluation Toolkit for Large Language Models"[153]

c. They are Architecture & Sustainable Design (ASD), Design & Artificial Intelligence (DAI), Engineering Product Development (EPD), Engineering Systems & Design (ESD), Humanities, Arts & Social Sciences (HASS), Information Systems Technology & Design (ISTD), and Science, Mathematics & Technology (SMT).

4.5 Agency for Science, Technology and Research (A*STAR)

A*STAR is a government agency in Singapore that spearheads "economic-oriented research to advance scientific discovery and develop innovative technology".[155] It houses over 30 research institutes split between its two research focuses: (1) Biomedical Research; and (2) Science and Engineering Research.[156] Three research institutes within A*STAR have published AI safety papers:[e] the Institute of Infocomm Research (I²R), the Institute of High Performance Computing (IHPC), and the Centre for Frontier AI Research (CFAR).[157]

- I²R: I²R is A*STAR's largest ICT research institute. Its mission is to "foster world-class infocomm and media research"[158] to power Singapore's digital economy. Key research capabilities center on machine intellection, aural and language intelligence, visual intelligence, cybersecurity, and communications and networks. I²R works on the MERaLiON model as a key research project under Singapore's National Multimodal Large Language Model (LLM) Programme.[159] One area of AI safety research under I²R is the development of "model-agnostic interpretability tools" to improve understanding of complex deep learning models.[160]
- IHPC: IHPC provides national leadership in computational modeling and simulation; its vision is to be "a global leader in computational modeling, simulation and AI".[161] Core research includes computational AI, which includes physics-based AI, efficient AI, multimodal AI, and autonomous AI. Safety-relevant research areas include improving robustness and interpretability of AI models.
- CFAR: CFAR is an A*STAR center whose mission is to develop next-generation AI technologies and seed cross-disciplinary R&D across the agency. CFAR organizes its work around five research pillars: AI for Science, Theory & Optimisation in AI, Artificial General Intelligence, Resilient & Safe AI (which includes robustness and interpretability), and Sustainable AI.

Several researchers within the respective A*STAR institutes have published papers that correspond to our categories of AI safety research, though none published at least five papers between January 2024 and July 2025.

4.6 Government Technology Agency (GovTech)

The Government Technology Agency (GovTech) was launched to harness tech and drive the Smart Nation initiative after the restructuring of the Infocomm Development Authority (IDA) in 2016 and now functions as a statutory board under the Ministry of Digital Development and Information. It is responsible for the delivery of Singapore's digital government products to the public.[162] GovTech's Data Science & AI Division (DSAID)

d. At the time of writing, Poria is still an associate professor at SUTD, but has announced that he will take on an associate professorship at NTU in the summer of 2025.[154]

e. We don't make a distinction between a research "institute" and other centers or research labs within A*STAR ; all bodies will be listed as "institutes."

develops whole-of-government AI products that government agencies can plug-and-play rather than building from scratch.[f]

While GovTech's main role is to build and deliver national digital platforms and services for the public sector, it has also begun contributing to AI safety research. Since 2024, researchers in GovTech released some AI safety research papers on topics such as, prompt-guardrail methodologies,[164] reducing toxicity in Singlish,[165] and evaluating out-of-knowledge-base robustness.[166]

4.7 Singapore AI Safety Institute

In May 2024, the Digital Trust Centre (DTC)—a national research center that drives research and development in trust technologies in cybersecurity, privacy technologies and trustworthy AI—was designated as the Singapore AISI (see also Domestic Approach and International Approach sections). The Singapore AISI pools researchers across Singapore to conduct research on AI safety evaluations and testing, specifically in the four research areas: (1) Testing & Evaluation; (2) Safe Model Design, Development & Deployment; (3) Content Assurance; and (4) Governance & Policy.[167]

There are five researchers in AISI/DTC who have listed research interests on AI safety,[168] covering safety research topics such as mechanistic interpretability and explainable AI. Singapore AISI Co-Executive Director Lam Kwok-Yan, who is also a professor at NTU, has been mentioned above (see subsection on "NTU"). As the Singapore AISI was recently established, the other researchers have not yet published papers that meet our criteria to be listed in this section.

4.8 Overall Trends

Our analysis reveals that AI safety research in Singapore is concentrated primarily across four universities: NUS, NTU, SMU and SUTD. We identified 14 researchers who have published at least five AI safety papers since the beginning of 2024 or played a major role in AI safety convenings. NUS and NTU had the most featured researchers, proportionate to their total AI research faculty.

4.8.1 Institutions

Institutions vary in their approaches to organizing AI safety research. Only NUS has elevated AI to a university-wide institute through the NUS AI Institute (NAII), while other universities house AI safety research within their respective computing faculties or departments. NUS (through its Responsible and Safe AI domain) and SMU (via its Safety, Security and Fairness research area) have explicitly listed AI safety as a designated research priority.

f. Flagship tools include Analytics.gov (secure data-analytics workspace), Cloak (personal data anonymisation), GovText (NLP pipeline for policy documents), Transcribe (localized speech-to-text) and the Video Analytics System (computer-vision toolkit).[163]

At the same time, safety research outside academia remains sparse. Although some frontier safety research has been published by GovTech and A*STAR's institutes such as I²R and CFAR, their researchers have not met the five-paper threshold as defined in our criteria for inclusion in this report—though admittedly, the threshold is relatively high, and somewhat arbitrary. Output of papers from Singapore-based companies appeared limited, unsurprisingly given the limited research presence of AI companies in Singapore.

4.8.2 Research Focus Areas

Analysis of the 14 researchers' published work reveals distinct patterns in Singapore's AI safety research priorities. Many papers focused on robustness (jailbreaking, backdoor attacks, etc.), with nine researchers publishing work specifically on robustness of large vision LLMs or multimodal models. This interest in the safety of models with visual components may reflect a focus on preventing harms such as deepfakes and non-consensual imagery, prioritizing these misuse cases over threats like biological weapons or cyberattacks.

Research grant priorities reinforce these patterns. In the AI Governance Research Grant Call 2023,[169] one of three research topics that AI Singapore highlighted was "supercharged disinformation" from generative AI, citing deepfakes and speech impersonation as potential harms. The 2024 Grant Call listed discrimination as the only eligible theme,[170] with multimodal model research in detection, monitoring, and prevention of misinformation as one possible research topic within that.

Additionally, at least six researchers showed interest in knowledge editing and unlearning techniques, reflecting a view that removing harmful information from AI models might be a key safety technique. Agent safety was also featured in papers by five of the 14 researchers, demonstrating that Singaporean researchers are engaged in frontier safety research. The recent 2025 Grant Call supports this trend, listing control of agents as one of four research themes alongside discrimination, copyright, and social resilience. Unlike earlier calls, the 2025 Grant Call specifically invites joint proposals between technical and non-STEM researchers for all four themes, reflecting a shift toward encouraging interdisciplinary collaboration in AI safety research.

4.8.3 Notable Gaps and Future Directions

Loss of control of advanced AI systems does not appear to be a topic of emphasis among Singaporean researchers. None of the listed researchers wrote papers on, for example, weak-to-strong generalization or AI supervision of superintelligent systems. However, there were a couple of papers on mechanistic interpretability and instrumental convergence, respectively, and work on issues like machine unlearning could still prove useful for ensuring control over increasingly advanced systems.[g] Future research may focus more on loss of control and dangerous capabilities including dual-use cyber, chemical, biological, and nuclear knowledge, especially following the release of the Singapore Consensus in May 2025 outlining global research priorities for technical AI safety.

g. Mechanistic interpretability: Zhang et al. (2025).[171] Instrumental convergence: Yufei He et al. (2025).[172]

Public Opinion

Key takeaways

- Public opinion data on AI safety in Singapore is limited. No dedicated national survey has examined Singaporeans' views on AI risks, leaving an evidence gap for policymakers.
- The limited public data available come from three global surveys that include Singapore as one of the sample countries; these show heightened concern over misinformation, data privacy, cybersecurity, and loss of human interaction from use of AI systems.
- Fewer Singaporean respondents expressed concern about systemic risks such as unequal access, environmental impact, or labor market disruption.
- None of the surveys addressed perceptions of malicious use or misuse of AI or the longer-term, existential risks associated with advanced AI systems, so public views on those topics remain largely unexamined.

At present, there is no national survey conducted specifically on Singaporeans' perceptions of AI safety risks. While numerous global surveys have explored general public attitudes toward AI, only a subset of these delve into the perceived negative impacts and safety concerns associated with AI systems.[173] Even fewer of these studies include Singapore as a target country, limiting the ability to draw Singapore-specific insights.

However, three recent global surveys included Singapore as a target country with sufficiently large[a] and representative sample sizes: the Ipsos AI Monitor 2024, 2025 Global AI Study conducted by the University of Melbourne in collaboration with KPMG, and 2024–2025 Survey on Generative AI in Media by YouGov. These surveys provide a snapshot of how Singaporeans view the risks associated with AI.

5.1 Global Surveys

5.1.1 Ipsos AI Monitor 2024

The Ipsos AI Monitor 2024 is a 32-country study that examines public attitudes towards AI, including levels of understanding of AI, trust towards AI, and expectations for its future. In Singapore, around 1,000 respondents aged between 21 and 74 were surveyed, primarily drawn from a more urban and educated demographic than the general population.[174]

a. We define a sample size of 1,000 as sufficiently large.

When it comes to AI-related risks, the report provides insights into public concerns about discrimination and disinformation. 60% of Singaporeans expressed trust that AI systems would not discriminate or show bias, a figure higher than the global average of 54%. This trust in AI closely mirrors Singaporeans' trust in humans not to discriminate (59%), again significantly higher than the global average of 45%. While the survey records that 60% of Singaporeans trust AI systems not to discriminate, it does not ask about the level of concern or perceived severity of discrimination-related harm among the Singaporean public, leaving an important aspect of risk perception unquantified.

On disinformation, 38% of surveyed Singaporeans anticipated that AI would exacerbate the problem over the next three to five years, while 30% expected an improvement. These results align closely with the global average, suggesting moderate concern over AI-driven disinformation. The survey did not pose questions on broader AI safety issues such as the risks posed by GPAI or existential threats.

5.1.2 University of Melbourne and KPMG Global Study 2025

A more comprehensive view of Singaporeans' concerns about AI safety emerges from the 2025 KPMG-University of Melbourne Global AI Study,[175] which covered public sentiment in 47 countries. In this survey, approximately 1,000 Singaporean respondents were sampled, reflecting the national population in terms of age, gender, and region.

When asked about potential risks, the top concerns among Singaporeans were cybersecurity threats (88%), loss of privacy or intellectual property (85%), and misinformation or disinformation (85%); the risks most commonly experienced personally were loss of human interaction (65%), misinformation/disinformation (64%), and inaccurate outcomes (61%). This suggests that Singaporeans are worried about potential future harms they have not yet experienced, even though their actual AI interactions primarily involve everyday usability issues like inaccurate responses and reduced human connection. Concerns such as harmful manipulation, system failure, or bias were acknowledged but reported less frequently in daily encounters (see Appendix B for the full table of risks).

5.1.3 YouGov Survey on Generative AI in Media (2024–2025)

Complementing these two reports is a "Survey on Generative AI in Media" by YouGov that focused specifically on generative AI in media across Hong Kong, Indonesia, and Singapore, conducted from December 2024 to January 2025. Among the 1,001 Singaporean respondents, top concerns included misinformation and deepfakes (58%), loss of human touch (50%), and privacy and data use (48%).[176] By contrast, issues such as bias and fairness received the least concern (28%), reinforcing a trend across surveys that Singaporeans are less preoccupied with systemic or ethical biases than with concrete, observable issues like misinformation or privacy breaches (see Appendix A for the full table of risks).

Conclusion

The State of AI Safety in Singapore report provides an up-to-date overview of Singapore's domestic and international approaches to AI governance, its industry and technical AI safety research landscape, and public attitudes toward AI risks. Taken together, these dimensions illuminate how Singapore is embracing opportunities and responding to risks presented by rapid AI development.

Domestically, Singapore's Model AI Governance Framework and AI Verify's testing framework and practical tools set the tone for a voluntary, industry-guided approach, complemented by targeted legislation addressing specific harms such as election deepfakes or online disinformation. On other fronts, we observe the development of the national AI standard SS ISO/IEC 42001 and publications centered on generative AI application testing, such as the Global Assurance Pilot report and the Starter Kit, signaling the importance Singapore places on providing clear, practical guidelines to AI developers. A key metric to watch will be how widely these voluntary frameworks and standards are adopted in practice, and whether emerging challenges prompt any shift toward stronger regulatory measures.

Internationally, Singapore has positioned itself as a neutral facilitator and active participant in global AI governance. Over the past year, Singapore has championed initiatives that include small states in the AI governance conversation, such as the release of the Digital FOSS AI Playbook and the ASEAN Guide. After handing over its ADGMIN chairmanship to Thailand earlier this year, we expect Singapore to continue leveraging this bridge-building role through the ASEAN WG-AI to foster consensus and build capacity among smaller nations. Looking ahead, Singapore's participation in joint testing initiatives through the International Network of AISIs and its track record of convening international experts through SCAI position it to continue facilitating cross-border collaboration on AI safety research, evaluation methodologies, and other practical risk management measures.

Singapore's governance initiatives, with their focus on industry guidance, have allowed Singapore to build a thriving AI assurance market and host regional offices of major global AI players and frontier start-ups, exposing local stakeholders to international best practices. For homegrown models, Singapore's SEA-LION and MERaLION models have so far prioritized multilingual performance, with safety largely left to downstream users, but the recent release of SEA-Guard signals a first step toward embedding safety guardrails directly into the domestic model line-up. As AI capabilities advance, Singapore's assurance ecosystem will need to tackle new challenges—such as real-time agent monitoring and control evaluations—further reinforcing the city-state's role as a regional hub for practical AI safety solutions.

A snapshot of the technical AI safety research in Singapore shows a small community of AI researchers working on safety; common research directions include model robustness, multimodal safety, knowledge editing, and agent behavior. Academic institutions remain central to research output, but we also observe government agencies and the new Singapore AI Safety Institute contributing. Frontier risk areas remain underexplored, but there may be more work in this area following AI Singapore's AI Research and Governance Joint Call (2025) and the recently published Singapore Consensus.

Lastly, public perception presents an information gap. Without dedicated national surveys, evidence of Singaporeans' awareness of and concern about AI risk remains limited. Building a fuller understanding of public sentiment and raising awareness of longer-term challenges could prove critical as AI technologies become more deeply embedded in society.

In summary, Singapore's experience offers a roadmap for other smaller countries seeking to safely harness the benefits of AI. By setting out a clear governance framework, providing practical guidance to AI model and application providers, and organizing international AI convenings, Singapore manages to steer conversations about AI safety. We hope that this report will equip readers with concrete examples of Singapore's AI safety and governance efforts and spark further collaboration toward a safer AI future—for nations large and small alike.

Appendix: Survey Findings

AI Risks in University of Melbourne and KPMG Global Study 2025[177]

Perception of AI Risks		
Types of risks	**% of people concerned about these risks (top 3 bolded)**	**% of people experienced negative outcomes from AI (top 3 bolded)**
Cybersecurity risks	**89%**	54%
Misinformation or disinformation	**85%**	**64%**
Loss of privacy or intellectual property	**85%**	48%
Manipulation or harmful use	84%	47%
Loss of human interaction and connection	83%	**65%**
Human rights being undermined	82%	38%
System failure	82%	50%
Job loss	81%	48%
Deskilling and dependency	81%	48%
Inaccurate outcomes	79%	**61%**
Disadvantage due to unequal access to AI	76%	43%
Environment impact	73%	40%
Bias or unfair treatment	72%	36%

YouGov Survey on Generative AI in Media (2024–2025)[178]

Concerns on Generative AI in Media	
What concerns do you have about the use of generative AI in content creation?	**Percentage of Singaporeans who expressed these concerns**
Loss of human touch	**50%**
Privacy and data usage	**48%**
Misinformation and deepfakes	**58%**
Quality of information	40%
Job displacement	37%
Originality	30%
Intellectual property and ownership	28%
Content moderation	21%
Mass content creation	22%
Bias and fairness	28%

Notes

1 IMDA and Aicadium, "Generative AI: Implications for Trust and Governance," June 2023, accessed July 11, 2025, https://aiverifyfoundation.sg/downloads/Discussion_Paper.pdf

2 Yoshua Bengio et al., *International AI Safety Report*, DSIT 2025/001 (2025), https://www.gov.uk/government/publications/international-ai-safety-report-2025

3 Infocomm Media Development Authority, *The Singapore Consensus on Global AI Safety Research Priorities: Building a Trustworthy, Reliable and Secure AI Ecosystem* (Singapore: Infocomm Media Development Authority, May 2025), https://file.go.gov.sg/sg-consensus-ai-safety.pdf

4 Monetary Authority of Singapore, *Principles to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of Artificial Intelligence and Data Analytics in Singapore's Financial Sector* (Singapore: Monetary Authority of Singapore, 2019), https://www.mas.gov.sg/publications/monographs-or-information-paper/2018/feat

5 Government of Singapore, *Road Traffic (Autonomous Motor Vehicles) Rules 2017*, tex.howpublished: Singapore Subsidiary Legislation No. S 464, August 2017, https://sso.agc.gov.sg/SL/RTA1961-S464-2017?DocDate=20170823

6 Ministry of Health, Singapore and Health Sciences Authority and Integrated Health Information Systems, *Artificial Intelligence in Healthcare Guidelines (AIHGle)* (Singapore: Ministry of Health, Singapore, 2022), https://isomer-user-content.by.gov.sg/3/9c0db09d-104c-48af-87c9-17e01695c67c/1-0-artificial-in-healthcare-guidelines-(aihgle)_publishedoct21.pdf

7 GovTech, "Responsible AI - Responsible AI Playbook," accessed July 11, 2025, https://playbooks.aip.gov.sg/responsibleai/responsibleai/

8 Personal Data Protection Commission of Singapore, "Discussion Paper On Artificial Intelligence (AI) And Personal Data –Fostering Responsible Development And Adoption Of AI," June 5, 2018, accessed July 11, 2025, https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/Discussion-Paper-on-AI-and-PD---050618.pdf

9 Nydia Remolina and Josephine Seah, "How to address the AI Governance discussion? What can we learn from Singapore's AI strategy?," August 2019, https://ink.library.smu.edu.sg/cgi/viewcontent.cgi?article=1000&context=caidg

10 IMDA and PDPC, "Model AI Framework First Edition," January 2019, accessed July 11, 2025, https://ai.bsa.org/wp-content/uploads/2019/09/Model-AI-Framework-First-Edition.pdf

11 IMDA and PDPC, "Model Artifical Intelligence Governance Framework Second Edition," January 2020, accessed July 11, 2025, https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/SGModelAIGovFramework2.pdf

12 IMDA and World Economic Forum, "Companion to the Model AI Governance Framework –Implementation and Self-Assessment Guide for Organizations," January 2020, accessed July 11, 2025, https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/SGIsago.pdf

13 IMDA and PDPC, "Compendium of Use Cases: Practical Illustrations of the Model AI Governance Framework," January 2020, accessed July 11, 2025, https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/SGAIGovUseCases.pdf

14 IMDA and Aicadium, "Generative AI: Implications for Trust and Governance," June 2023, accessed July 11, 2025, https://aiverifyfoundation.sg/downloads/Discussion_Paper.pdf

15 IMDA and AI Verify Foundation, "Model AI Governance Framework for Generative AI," May 30, 2024, accessed July 11, 2025, https://aiverifyfoundation.sg/wp-content/uploads/2024/05/Model-AI-Governance-Framework-for-Generative-AI-May-2024-1-1.pdf

16 IMDA, "Singapore Launches A.I. Verify," Infocomm Media Development Authority, May 25, 2022, accessed July 11, 2025, https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/press-releases/2022/sg-launches-worlds-first-ai-testing-framework-and-toolkit-to-promote-transparency

17 PDPC, "Singapore's Approach to AI Governance," July 11, 2025, accessed July 11, 2025, https://www.pdpc.gov.sg/help-and-resources/2020/01/model-ai-governance-framework

18 AI Verify Foundation, "Project Moonshot," AI Verify Foundation, July 11, 2025, accessed July 11, 2025, https://aiverifyfoundation.sg/project-moonshot/

19 AI Verify Foundation, "Foundation members," AI Verify Foundation, 2025, accessed July 19, 2025, https://aiverifyfoundation.sg/foundation-members/

20 IMDA and AI Verify Foundation, "Cataloguing LLM Evaluations," October 2023, accessed July 11, 2025, https://aiverifyfoundation.sg/downloads/Cataloguing_LLM_Evaluations.pdf

21 Toby Shevlane et al., *Model evaluation for extreme risks*, Issue: arXiv:2305.15324, arXiv:2305.15324, September 22, 2023, accessed July 11, 2025, arXiv: 2305.15324[cs], http://arxiv.org/abs/2305.15324

22 "Generative AI Evaluation Sandbox," Infocomm Media Development Authority, October 31, 2023, accessed July 11, 2025, https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/press-releases/2023/generative-ai-evaluation-sandbox

23 MLCommons, "MLCommons and AI Verify to collaborate on AI Safety Initiative," MLCommons, May 31, 2024, accessed July 16, 2025, https://mlcommons.org/2024/05/mlcommons-and-ai-verify-moi-ai-safety-initiative/

24 MLCommons, "ailuminate.mlcommons.org/benchmarks/," 2025, accessed July 19, 2025, https://ailuminate.mlcommons.org/benchmarks/; MLCommons, "MLCommons Releases AILuminate LLM v1.1, Adding French Language Capabilities to Industry-Leading AI Safety Benchmark," MLCommons, February 11, 2025, accessed July 19, 2025, https://mlcommons.org/2025/02/ailumiate-v1-1-fr/; MLCommons, "MLCommons Announces Expansion of Industry-Leading AILuminate Benchmark," MLCommons, May 29, 2025, accessed July 19, 2025, https://mlcommons.org/2025/05/nasscom/

25 Shaona Ghosh et al., *AILuminate: Introducing v1.0 of the AI Risk and Reliability Benchmark from MLCommons*, arXiv:2503.05731, April 18, 2025, accessed July 19, 2025, arXiv: 2503.05731[cs], http://arxiv.org/abs/2503.05731

26 "Global AI Assurance Pilot," AI Verify Foundation, July 11, 2025, accessed July 11, 2025, https://aiverifyfoundation.sg/ai-assurance-pilot/

27 AI Verify Foundation, "Testing Real World GenAI Systems Main Report," May 2025, accessed July 11, 2025, https://assurance.aiverifyfoundation.sg/report/introduction/

28 IMDA, "Starter Kit for Safety Testing of LLM-Based Applications," May 28, 2025, accessed July 11, 2025, https://www.imda.gov.sg/-/media/imda/files/about/emerging-tech-and-research/artificial-intelligence/large-language-model-starter-kit.pdf

29 AI Verify Foundation, "Global AI Assurance Sandbox," AI Verify Foundation, 2025, accessed July 19, 2025, https://aiverifyfoundation.sg/ai-assurance/

30 "NTU to set up $50m centre to advance digital trust technologies in Singapore," *The Straits Times* (Singapore), June 1, 2022, ISSN: 0585-3923, accessed July 11, 2025, https://www.straitstimes.com/tech/tech-news/ntu-to-set-up-50m-centre-to-advance-digital-trust-technologies-in-singapore

31 AISI, "AI Safety Institute," July 11, 2025, accessed July 11, 2025, https://sgaisi.sg/our-works/

32 Anthropic, "Challenges in Red Teaming AI Systems," June 13, 2024, accessed July 10, 2025, https://www.anthropic.com/news/challenges-in-red-teaming-ai-systems

33 Ministry of Digital Development and Information, "Singapore Announces New AI Safety Initiatives at the Global AI Action Summit in France," Ministry of Digital Development and Information, February 11, 2025, accessed July 11, 2025, https://www.mddi.gov.sg/newsroom/singapore-announces-new-ai-safety-initiatives/

34 Ministry of Digital Development and Information, "Minister Josephine Teo's Comments at CNBC Converge Live 2025," Ministry of Digital Development and Information, March 12, 2025, accessed July 11, 2025, https://www.mddi.gov.sg/newsroom/minister-josephine-teo-comments-at-cnbc-converge-live-2025/

35 AGC, "Penal Code 1871," Singapore Statutes Online, July 11, 2025, accessed July 11, 2025, https://sso.agc.gov.sg/Act/PC1871?ValidDate=20250324&ProvIds=pr377BG-,pr377BH-,pr377BI-,pr377BJ-

36 Singapore Parliament, "Reference from Other Countries' Decisions to Criminalise Creation or Possession of Sexually Explicit Deepfake Images and Videos," February 5, 2025, accessed July 11, 2025, https://sprs.parl.gov.sg/search/#/sprs3topic?reportid=written-answer-na-18838

37 AGC, "Elections (Integrity of Online Advertising) (Amendment) Bill," Singapore Statutes Online, September 9, 2024, accessed July 11, 2025, https://sso.agc.gov.sg/Bills-Supp/29-2024/Published/20240909?DocDate=20240909

38 Jason Grant Allen, Jane Loo, and Jose Luna, "Governing intelligence: Singapor erning intelligence: Singapore's evolving AI go olving AI governance ernance framework," January 2025, https://ink.library.smu.edu.sg/cgi/viewcontent.cgi?article=6527&context=sol_research

39 Jason Grant Allen, Jane Loo, and Jose Luna, "Governing intelligence: Singapor erning intelligence: Singapore's evolving AI go olving AI governance ernance framework," January 2025, https://ink.library.smu.edu.sg/cgi/viewcontent.cgi?article=6527&context=sol_research

40 Infocomm Media Development Authority, "IT Standards Committee (ITSC)," Infocomm Media Development Authority, accessed July 11, 2025, https://www.imda.gov.sg/regulations-and-licensing-listing/ict-standards-and-quality-of-service/industry-committees-and-working-groups/it-standards-committee

41 Singapore Standards, "TR 99: 2021 Artificial intelligence Security," July 11, 2025, accessed July 11, 2025, https://www.singaporestandardseshop.sg/Product/SSPdtDetail/553b4562-ef85-4ebb-bfeb-2b35beb1d29c

42 Singapore Standards, "SS ISO/IEC 42001:2024 Information Technology –Artificial Intelligence –Management System," July 11, 2025, accessed July 11, 2025, https://www.singaporestandardseshop.sg/Product/SSPdtDetail/8925f190-9b21-4af4-b52c-268f7df47727

43 Singapore University of Technology and Design (SUTD), "Launch of Artificial Intelligence Management Systems Accreditation Programme," Singapore University of Technology and Design (SUTD), March 17, 2025, accessed July 11, 2025, https://www.sutd.edu.sg/stories-listing/ai-management-systems-accreditation-programme-launch/

44 ISO, "ISO/IEC 42001:2023," ISO, July 11, 2025, accessed July 11, 2025, https://www.iso.org/standard/42001

45 Singapore Standards, "SS ISO/IEC 42001:2024 Information Technology –Artificial Intelligence –Management System," July 11, 2025, accessed July 11, 2025, https://www.singaporestandardseshop.sg/Product/SSPdtDetail/8925f190-9b21-4af4-b52c-268f7df47727

46 OECD, "List of participants in the OECD Expert Group on AI (AIGO)," July 11, 2025, accessed July 11, 2025, https://oecd.ai/en/list-of-participants-oecd-expert-group-on-ai

47 UNESCO, "Recommendation on the Ethics of Artificial Intelligence," May 16, 2023, accessed July 11, 2025, https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence

48 United Nations, General Assembly, *Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development*, Draft Resolution A/78/L.49 (United Nations, March 2024), https://docs.un.org/en/A/78/L.49

49 *Enhancing international cooperation on capacity-building of artificial intelligence: resolution*, in collab. with UN. General Assembly (78th sess. : 2023-2024), Num Pages: 5 Place: New York, July 1, 2024, accessed July 11, 2025, https://digitallibrary.un.org/record/4054005

50 United Nations General Assembly, *Artificial intelligence in the military domain and its implications for international peace and security: resolution* (December 24, 2024), accessed July 11, 2025, https://digitallibrary.un.org/record/4070018

51 UK government, "AI Safety Summit 2023," April 28, 2025, accessed July 11, 2025, https://www.gov.uk/government/topical-events/ai-safety-summit-2023

52 Hong Siang Ng, "Countries developing 'frontier AI' need to work together for mutual security: PM Lee," CNA, November 2, 2023, accessed July 11, 2025, https://www.channelnewsasia.com/singapore/ai-safety-summit-singapore-pm-lee-frontier-3892476

53 UK government, "The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023," GOV.UK, February 13, 2025, accessed July 11, 2025, https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023

54 UK government, "New commitment to deepen work on severe AI risks concludes AI Seoul Summit," May 22, 2024, accessed July 11, 2025, https://www.gov.uk/government/news/new-commitmentto-deepen-work-on-severe-ai-risks-concludes-ai-seoul-summit

55 IMDA, "Digital Trust Centre designated as Singapore's AISI," Infocomm Media Development Authority, May 22, 2024, accessed July 11, 2025, https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/factsheets/2024/digital-trust-centre

56 Emmanuel Macron, "Statement on Inclusive and Sustainable Artificial Intelligence for People and the Planet.," elysee.fr, Section: Press release, February 11, 2025, accessed July 11, 2025, https://www.elysee.fr/en/emmanuel-macron/2025/02/11/statement-on-inclusive-and-sustainable-artificial-intelligence-for-people-and-the-planet

57 Emmanuel Macron, "Paris Declaration on Maintaining Human Control in AI enabled Weapon Systems.," elysee.fr, Section: Communiqué de presse, February 11, 2025, accessed July 11, 2025, https://www.elysee.fr/emmanuel-macron/2025/02/11/paris-declaration-on-maintaining-human-control-in-ai-enabled-weapon-systems

58 Office for Digital and Emerging Technologies, "News and Resources," July 11, 2025, accessed July 11, 2025, https://www.un.org/digital-emerging-technologies/content/news

59 United Nations, *Governing AI for humanity: final report* (New York, NY: United Nations, September 2024), ISBN: 978-92-1-106787-3, https://www.un.org/sites/un2.un.org/files/governing_ai_for_humanity_final_report_en.pdf

60 "List of SCAI Questions," Singapore Conference on AI, July 11, 2025, accessed July 11, 2025, https://www.scai.gov.sg/2023/list-of-scai-questions/list-of-scai-questions/

61 Singapore Conference on AI, "The Singapore Consensus on Global AI Safety Research Priorities," Singapore Conference on AI, May 8, 2025, accessed July 11, 2025, https://www.scai.gov.sg/2025/scai2025-report/; Infocomm Media Development Authority, *The Singapore Consensus on Global AI Safety Research Priorities: Building a Trustworthy, Reliable and Secure AI Ecosystem* (Singapore: Infocomm Media Development Authority, May 2025), https://file.go.gov.sg/sg-consensus-ai-safety.pdf

62 International Network of AI Safety Institutes, "Improving International Testing of Foundation Models- A Pilot Testing Exercise from the International Network of AI Safety Institutes.pdf," November 20, 2024, accessed July 11, 2025, https://www.nist.gov/system/files/documents/2024/11/21/Improving%20International%20Testing%20of%20Foundation%20Models-%20%20%20A%20Pilot%20Testing%20Exercise%20from%20the%20International%20Network%20of%20AI%20Safety%20Institutes.pdf

63 International Network of AI Safety Institutes, "International Network of AI Safety Institutes Joint Testing Exercise: Improving Methodologies for AI Model Evaluations Across Global Languages," February 11, 2025, accessed July 11, 2025, https://sgaisi.sg/wp-api/wp-content/uploads/2025/03/International-Network-of-AI-Safety-Institutes-Joint-Testing-Exercise-Improving-Methodologies-for-AI-Model-Evaluations-Across-Global-Languages.pdf

64 Singapore AI Safety Institute, "Improving Methodologies for-LLM Evaluations Across Global Languages Evaluation Report," June 2025, accessed July 11, 2025, https://sgaisi.sg/wp-api/wp-content/uploads/2025/06/Improving-Methodologies-for-LLM-Evaluations-Across-Global-Languages-Evaluation-Report-1.pdf

65 Ministry of Public Administration and Artificial Intelligence, Trinidad & Tabago, "Annex to "FOSS for Good" Technical Assistance Package," September 2021, accessed July 11, 2025, https://www.mpa.gov.tt/sites/default/files/file_upload/psacourses/SCPTA/Annex.pdf

66 Singapore Press Centre, "Singapore Prepares Ahead to Leverage Artificial Intelligence for a Better Future," Default, May 29, 2024, accessed July 11, 2025, https://www.sgpc.gov.sg/detail?url=/media_releases/imda/press_release/P-20240529-2&page=/detail&HomePage=home

67 IMDA, "Singapore concludes fruitful chairmanship of the ASEAN Digital Ministers Meeting, delivering concrete outcomes to bring ASEAN towards a brighter Digital Future," October 2023, accessed January 17, 2025, https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/press-releases/2025/5th-asean-digital-ministers-meeting

68 ASEAN, "ASEAN Guide on AI Governance and Ethics," February 2024, accessed July 11, 2025, https://asean.org/wp-content/uploads/2024/02/ASEAN-Guide-on-AI-Governance-and-Ethics_beautified_201223_v2.pdf

69 Ministry of Digital Development and Information, "1st ASEAN Digital Ministers' Meeting approves Singapore led initiatives," Ministry of Digital Development and Information, January 22, 2021, accessed July 11, 2025, https://www.mddi.gov.sg/newsroom/1st-asean-digital-ministers-meeting-approves-singapore-led-initiatives/

70 ASEAN, "ASEAN Guide on AI Governance and Ethics," February 2024, accessed July 11, 2025, https://asean.org/wp-content/uploads/2024/02/ASEAN-Guide-on-AI-Governance-and-Ethics_beautified_201223_v2.pdf

71 UCARE.AI, "AI Intelligent agents and tools for insurers and insurance platforms," 2024, accessed July 19, 2025, https://www.ucare.ai/

72 ASEAN, "Expanded ASEAN Guide on AI Governance and Ethics Generative AI," January 2025, accessed July 11, 2025, https://asean.org/wp-content/uploads/2025/01/Expanded-ASEAN-Guide-on-AI-Governance-and-Ethics-Generative-AI.pdf

73 Ministry of Digital Development and Information, "Singapore Concludes Fruitful Chairmanship of the ASEAN Digital Ministers Meeting," Ministry of Digital Development and Information, January 17, 2025, accessed July 11, 2025, https://www.mddi.gov.sg/newsroom/singapore-concludes-fruitful-chairmanship-of-adgmin/

74 Infocomm Media Development Authority, "Joint Exercise: IMDA and US NIST Mapping," Infocomm Media Development Authority, October 13, 2023, accessed July 11, 2025, https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/press-releases/2023/nist-imda-joint-mapping-exercise

75 "AI RMF and AI Verify Crosswalk," October 12, 2023, accessed July 11, 2025, https://aiverifyfoundation.sg/downloads/AI_RMF_and_AI_Verify_Crosswalk.pdf

76 Infocomm Media Development Authority, "Joint Exercise: IMDA and US NIST Mapping," Infocomm Media Development Authority, October 13, 2023, accessed July 11, 2025, https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/press-releases/2023/nist-imda-joint-mapping-exercise

77 AI Verify Foundation, "Crosswalk: ISO/IEC 42001 and AI Verify," June 2024, accessed July 11, 2025, https://aiverifyfoundation.sg/wp-content/uploads/2024/06/Crosswalk-AIV-and-ISO42001-final.pdf

78 AI Verify Foundation, "Crosswalk between AI Verify testing framework and G7's Code of Conduct," May 2025, accessed July 11, 2025, https://file.go.gov.sg/crosswalk-aivtf-coc.pdf; Ministry of Foreign Affairs Japan, "Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems," July 11, 2025, accessed July 11, 2025, https://www.mofa.go.jp/files/100573473.pdf

79 Infocomm Media Development Authority, "Crosswalk Between NIST AI Risk Management Framework: Generative AI Profile (AI 600-1) and Singapore / IMDA AI Verify Testing Framework," May 28, 2025, accessed July 11, 2025, https://file.go.gov.sg/crosswalk-aivtfxairmf-genaiprofile.pdf

80 Ministry of Digital Development and Information, "New Singapore UK Agreement to Strengthen Global AI Safety and Governance," Ministry of Digital Development and Information, November 6, 2024, accessed July 11, 2025, https://www.mddi.gov.sg/newsroom/new-singapore-uk-agreement-to-strengthen-global-ai-safety-governance/

81 Ministry of Digital Development and Information, "Singapore and the European Union Agree to Strengthen Collaboration on AI Safety," Ministry of Digital Development and Information, November 20, 2024, accessed July 11, 2025, https://www.mddi.gov.sg/newsroom/singapore-european-union-agree-to-strengthen-collaboration-ai-safety/

82 Ministry of Digital Development and Information, "Singapore and China advance cooperation at inaugural Singapore China Digital Policy Dialogue," Ministry of Digital Development and Information, June 27, 2024, accessed July 11, 2025, https://www.mddi.gov.sg/newsroom/sg-china-advance-cooperation-at-inaugural-digital-policy-dialogue/

83 Ministry of Digital Development and Information, "Singapore and Australia Expand Cooperation on AI with New Memorandum of Understanding," Ministry of Digital Development and Information, December 16, 2024, accessed July 11, 2025, https://www.mddi.gov.sg/newsroom/singapore-australia-expand-cooperation-on-ai-with-new-mou/

84 Ministry of Foreign Affairs Singapore, "U.S. - Singapore Digital Economic Cooperation Roadmap," July 31, 2024, accessed July 11, 2025, http://www.mfa.gov.sg/Newsroom/Press-Statements-Transcripts-and-Photos/2024/07/20240731-Blinken-visit

85 SEA-LION.AI, "Our Models," November 5, 2024, accessed July 19, 2025, https://sea-lion.ai/our-models/

86 Yosephine Susanto et al., *SEA-HELM: Southeast Asian Holistic Evaluation of Language Models*, Issue: arXiv:2502.14301, arXiv:2502.14301, June 2, 2025, accessed July 10, 2025, arXiv: 2502.14301[cs], http://arxiv.org/abs/2502.14301

87 Yosephine Susanto et al., *SEA-HELM: Southeast Asian Holistic Evaluation of Language Models*, Issue: arXiv:2502.14301, arXiv:2502.14301, June 2, 2025, accessed July 10, 2025, arXiv: 2502.14301[cs], http://arxiv.org/abs/2502.14301

88 AI Singapore, "SEA-LION Foundation Family," March 19, 2025, accessed July 10, 2025, https://docs.sea-lion.ai/models/sea-lion_adaptations

89 AI Singapore, "MERaLiON/MERaLiON-AudioLLM-Whisper-SEA-LION," Hugging Face, December 2024, accessed July 10, 2025, https://huggingface.co/MERaLiON/MERaLiON-AudioLLM-Whisper-SEA-LION

90 AI Singapore, "MERaLiON/MERaLiON-2-10B-ASR," Hugging Face, May 2025, accessed July 10, 2025, https://huggingface.co/MERaLiON/MERaLiON-2-10B-ASR

91 Yingxu He et al., *MERaLiON-AudioLLM: Bridging Audio and Language with Large Language Models*, Issue: arXiv:2412.09818, arXiv:2412.09818, January 16, 2025, accessed July 10, 2025, arXiv: 2412.09818[cs], http://arxiv.org/abs/2412.09818

92 AI Singapore, "Leaderboard / SeaEval," November 2024, accessed July 10, 2025, https://huggingface.co/spaces/MERaLiON/SeaEval_Leaderboard; AI Singapore, "Leaderboard / AudioBench," August 2024, accessed July 10, 2025, https://huggingface.co/spaces/MERaLiON/AudioBench-Leaderboard

93 AI Singapore, "SEA-Guard API," May 29, 2025, accessed July 10, 2025, https://docs.sea-lion.ai/guides/inferencing/api

94 Department for Science, Innovation & Technology, "Introduction to AI assurance," https://assets.publishing.service.gov.uk/media/65ccf508c96cf3000c6a37a1/Introduction_to_AI_Assurance.pdf

95 Yogasai Gazula, "Demystifying the AI Assurance Landscape," Responsible AI, February 5, 2025, accessed July 10, 2025, http://www.responsible.ai/demystifying-the-ai-assurance-landscape/

96 "Testimonials," AI Verify Foundation, accessed July 10, 2025, https://aiverifyfoundation.sg/foundation-members/what-our-general-members-say/

97 ANNEX –LIST OF PARTICIPANTS IN SANDBOX, "Generative AI Evaluation Sandbox," 2023, accessed July 19, 2025, https://www.imda.gov.sg/-/media/imda/files/news-and-events/media-room/media-releases/2023/10/generative-ai-evaluation-sandbox/annex-a---list-of-participants-in-sandbox.pdf

98 Olivia Minnock, "Alibaba to open first joint research centre in Singapore with Nanyang Technological University," March 2, 2018, https://www.ntu.edu.sg/docs/default-source/corporate-ntu/media-hub/0487-ntu-singapore-and-alibaba-group-launch-joint-research-institute-on-artificial-intelligence-technologies/business-chief_180302_alibaba-to-open-first-joint-research-centre-in-singapore-with-nanyang-technological-university.pdf?sfvrsn=6559f822_2; AI Verify Foundation, "Project Moonshot," AI Verify Foundation, July 11, 2025, accessed July 11, 2025, https://aiverifyfoundation.sg/project-moonshot/

99 Alibaba Cloud Community, "Alibaba's DAMO Academy Unveils LLMs Designed For Southeast Asia," Alibaba Cloud Community, December 12, 2023, accessed July 10, 2025, https://www.alibabacloud.com/blog/alibabas-damo-academy-unveils-llms-designed-for-southeast-asia_600648

100 ANNEX –LIST OF PARTICIPANTS IN SANDBOX, "Generative AI Evaluation Sandbox," 2023, accessed July 19, 2025, https://www.imda.gov.sg/-/media/imda/files/news-and-events/media-room/media-releases/2023/10/generative-ai-evaluation-sandbox/annex-a---list-of-participants-in-sandbox.pdf; Anthropic, "Challenges in Red Teaming AI Systems," June 13, 2024, accessed July 10, 2025, https://www.anthropic.com/news/challenges-in-red-teaming-ai-systems

101 Yijian Finance (易简财经), "From Innovation Works to 01.AI: Kai-Fu Lee's AI Pivot and Global Ambitions (从创新工场到零一万物：李开复的 AI 转型与全球野心)," August 1, 2024, accessed July 10, 2025, https://i.ifeng.com/c/8bhOgshmutd

102 Dan Hendrycks and Mantas Mazeika, *X-Risk Analysis for AI Research*, Issue: arXiv:2206.05862, arXiv:2206.05862, September 20, 2022, accessed July 15, 2025, arXiv: 2206.05862[cs], http://arxiv.org/abs/2206.05862; Dan Hendrycks et al., *Unsolved Problems in ML Safety*, Issue: arXiv:2109.13916, arXiv:2109.13916, June 16, 2022, accessed July 15, 2025, arXiv: 2109.13916[cs], http://arxiv.org/abs/2109.13916; "Introduction to AI Safety, Ethics and Society," July 15, 2025, accessed July 15, 2025, https://www.aisafetybook.com/textbook

103 Yangyang Guo et al., *Technical Report for ICML 2024 TiFA Workshop MLLM Attack Challenge: Suffix Injection and Projected Gradient Descent Can Easily Fool An MLLM*, Issue: arXiv:2412.15614, arXiv:2412.15614, December 20, 2024, accessed July 15, 2025, arXiv: 2412.15614[cs], http://arxiv.org/abs/2412.15614

104 Yash Sinha, Murari Mandal, and Mohan Kankanhalli, *UnStar: Unlearning with Self-Taught Anti-Sample Reasoning for LLMs*, Issue: arXiv:2410.17050, arXiv:2410.17050, October 22, 2024, accessed July 15, 2025, arXiv: 2410.17050[cs], http://arxiv.org/abs/2410.17050

105 Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli, *Hallucination is Inevitable: An Innate Limitation of Large Language Models*, Issue: arXiv:2401.11817, arXiv:2401.11817, February 13, 2025, accessed July 15, 2025, arXiv: 2401.11817[cs], http://arxiv.org/abs/2401.11817

106 Hongzhan Lin et al., *FACT-AUDIT: An Adaptive Multi-Agent Framework for Dynamic Fact-Checking Evaluation of Large Language Models*, Issue: arXiv:2502.17924, arXiv:2502.17924, March 2, 2025, accessed July 15, 2025, arXiv: 2502.17924[cs], http://arxiv.org/abs/2502.17924

107 Chenhang Cui et al., *Safe + Safe = Unsafe? Exploring How Safe Images Can Be Exploited to Jailbreak Large Vision-Language Models*, Issue: arXiv:2411.11496, arXiv:2411.11496, November 28, 2024, accessed July 15, 2025, arXiv: 2411.11496[cs], http://arxiv.org/abs/2411.11496

108 Junfeng Fang et al., *SafeMLRM: Demystifying Safety in Multi-modal Large Reasoning Models*, Issue: arXiv:2504.08813, arXiv:2504.08813, April 9, 2025, accessed July 15, 2025, arXiv: 2504.08813[cs], http://arxiv.org/abs/2504.08813

109 Jingtan Wang et al., *Helpful or Harmful Data? Fine-tuning-free Shapley Attribution for Explaining Language Model Predictions*, arXiv:2406.04606, June 7, 2024, accessed July 19, 2025, arXiv: 2406.04606[cs], http://arxiv.org/abs/2406.04606

110 Zijian Zhou et al., *DETAIL: Task DEmonsTration Attribution for Interpretable In-context Learning*, arXiv:2405.14899, December 15, 2024, accessed July 19, 2025, arXiv: 2405.14899[cs], http://arxiv.org/abs/2405.14899

111 Nhung Bui et al., *On Newton's Method to Unlearn Neural Networks*, arXiv:2406.14507, August 27, 2024, accessed July 19, 2025, arXiv: 2406.14507[cs], http://arxiv.org/abs/2406.14507

112 "KAN Min Yen - NUS Computing," accessed July 19, 2025, https://www.comp.nus.edu.sg/cs/people/kanmy/

113 Esther Gan et al., *Reasoning Robustness of LLMs to Adversarial Typographical Errors*, arXiv:2411.05345, November 8, 2024, accessed July 19, 2025, arXiv: 2411.05345[cs], http://arxiv.org/abs/2411.05345

114 Do Xuan Long et al., *Aligning Large Language Models with Human Opinions through Persona Selection and Value–Belief–Norm Reasoning*, arXiv:2311.08385, December 14, 2024, accessed July 19, 2025, arXiv: 2311.08385[cs], http://arxiv.org/abs/2311.08385

115 Jiaying Wu et al., *Seeing Through Deception: Uncovering Misleading Creator Intent in Multimodal News with Vision-Language Models*, arXiv:2505.15489, May 26, 2025, accessed July 19, 2025, arXiv: 2505.15489[cs], http://arxiv.org/abs/2505.15489

116 Yufei He et al., *Evaluating the Paperclip Maximizer: Are RL-Based Language Models More Likely to Pursue Instrumental Goals?*, Issue: arXiv:2502.12206, arXiv:2502.12206, February 16, 2025, accessed July 15, 2025, arXiv: 2502.12206[cs], http://arxiv.org/abs/2502.12206

117 Wenhao You et al., *MIRAGE: Multimodal Immersive Reasoning and Guided Exploration for Red-Team Jailbreak Attacks*, Issue: arXiv:2503.19134, arXiv:2503.19134, March 24, 2025, accessed July 15, 2025, arXiv: 2503.19134[cs], http://arxiv.org/abs/2503.19134

118 Shuyang Hao et al., *Tit-for-Tat: Safeguarding Large Vision-Language Models Against Jailbreak Attacks via Adversarial Defense*, Issue: arXiv:2503.11619, arXiv:2503.11619, March 14, 2025, accessed July 15, 2025, arXiv: 2503.11619[cs], http://arxiv.org/abs/2503.11619

119 Wenjie Qu et al., *Provably Robust Multi-bit Watermarking for AI-generated Text*, arXiv:2401.16820, January 28, 2025, accessed July 19, 2025, arXiv: 2401.16820[cs], http://arxiv.org/abs/2401.16820

120 "[2504.08104] Geneshift: Impact of different scenario shift on Jailbreaking LLM," accessed July 19, 2025, https://arxiv.org/abs/2504.08104

121 Yue Liu et al., *GuardReasoner: Towards Reasoning-based LLM Safeguards*, arXiv:2501.18492, January 30, 2025, accessed July 19, 2025, arXiv: 2501.18492[cs], http://arxiv.org/abs/2501.18492

122 "Data Science Faculty," College of Computing and Data Science, July 15, 2025, accessed July 15, 2025, https://www.ntu.edu.sg/computing/data-science-at-ntu/data-science-faculty

123 "Cyber Security Lab (CSL)," College of Computing and Data Science, July 15, 2025, accessed July 15, 2025, https://www.ntu.edu.sg/computing/research/institutes-centres/csl

124 "Computational Intelligence Lab (CIL)," Computational Intelligence Lab (CIL), July 15, 2025, accessed July 15, 2025, https://www.ntu.edu.sg/cil/faculty-directory

125 "Multimedia and Interactive Computing Lab (MICL)," College of Computing and Data Science, July 15, 2025, accessed July 15, 2025, https://www.ntu.edu.sg/computing/research/institutes-centres/micl

126 Nanyang Technological University, "Generative AI Lab," July 10, 2025, accessed July 10, 2025, https://www.ntu.edu.sg/computing/research/institutes-centres/grail/about-us

127 "Prof Lam Kwok Yan," Nanyang Technological University, July 15, 2025, accessed July 15, 2025, https://dr.ntu.edu.sg/entities/person/Lam-Kwok-Yan

128 Huanyi Ye et al., "Enhancing AI safety of machine unlearning for ensembled models," *Applied Soft Computing* 174 (April 1, 2025): 113011, ISSN: 1568-4946, accessed July 15, 2025, https://doi.org/10.1016/j.asoc.2025.113011, https://www.sciencedirect.com/science/article/pii/S1568494625003229

129 Ziyao Liu et al., *Threats, Attacks, and Defenses in Machine Unlearning: A Survey*, Issue: arXiv:2403.13682, arXiv:2403.13682, February 17, 2025, accessed July 15, 2025, arXiv: 2403.13682[cs], http://arxiv.org/abs/2403.13682

130 Chen Chen et al., *Trustworthy, Responsible, and Safe AI: A Comprehensive Architectural Framework for AI Safety with Challenges and Mitigations*, Issue: arXiv:2408.12935, arXiv:2408.12935, January 15, 2025, accessed July 15, 2025, arXiv: 2408.12935[cs], http://arxiv.org/abs/2408.12935

131 "Prof Luke Ong （翁之昊）," Nanyang Technological University, July 15, 2025, accessed July 15, 2025, https://dr.ntu.edu.sg/entities/person/Luke-Ong

132 Fazl Barez et al., *Open Problems in Machine Unlearning for AI Safety*, Issue: arXiv:2501.04952, arXiv:2501.04952, January 9, 2025, accessed July 15, 2025, arXiv: 2501.04952[cs], http://arxiv.org/abs/2501.04952

133 Xiaojun Jia et al., *Evolution-based Region Adversarial Prompt Learning for Robustness Enhancement in Vision-Language Models*, Issue: arXiv:2503.12874, arXiv:2503.12874, March 18, 2025, accessed July 15, 2025, arXiv: 2503.12874[cs], http://arxiv.org/abs/2503.12874

134 Qi Zhou et al., *Defending LVLMs Against Vision Attacks through Partial-Perception Supervision*, Issue: arXiv:2412.12722, arXiv:2412.12722, December 17, 2024, accessed July 15, 2025, arXiv: 2412.12722[cs], http://arxiv.org/abs/2412.12722

135 Zhenhong Zhou et al., *CORBA: Contagious Recursive Blocking Attacks on Multi-Agent Systems Based on Large Language Models*, Issue: arXiv:2502.14529, arXiv:2502.14529, February 20, 2025, accessed July 15, 2025, arXiv: 2502.14529[cs], http://arxiv.org/abs/2502.14529

136 Mang Ye et al., *A Survey of Safety on Large Vision-Language Models: Attacks, Defenses and Evaluations*, Issue: arXiv:2502.14881, arXiv:2502.14881, February 14, 2025, accessed July 15, 2025, arXiv: 2502.14881[cs], http://arxiv.org/abs/2502.14881

137 Yibo Wang et al., *Panacea: Mitigating Harmful Fine-tuning for Large Language Models via Post-fine-tuning Perturbation*, Issue: arXiv:2501.18100, arXiv:2501.18100, January 30, 2025, accessed July 15, 2025, arXiv: 2501.18100[cs], http://arxiv.org/abs/2501.18100

138 Huazheng Wang et al., *Erasing Without Remembering: Implicit Knowledge Forgetting in Large Language Models*, Issue: arXiv:2502.19982, arXiv:2502.19982, May 20, 2025, accessed July 15, 2025, arXiv: 2502.19982[cs], http://arxiv.org/abs/2502.19982

139 Zihan Wang et al., *BadLingual: A Novel Lingual-Backdoor Attack against Large Language Models*, Issue: arXiv:2505.03501, arXiv:2505.03501, May 6, 2025, accessed July 15, 2025, arXiv: 2505.03501[cs], http://arxiv.org/abs/2505.03501

140 Guanlin Li et al., *Picky LLMs and Unreliable RMs: An Empirical Study on Safety Alignment after Instruction Tuning*, Issue: arXiv:2502.01116, arXiv:2502.01116, February 3, 2025, accessed July 15, 2025, arXiv: 2502.01116[cs], http://arxiv.org/abs/2502.01116

141 Chenhang Cui et al., *Safe + Safe = Unsafe? Exploring How Safe Images Can Be Exploited to Jailbreak Large Vision-Language Models*, Issue: arXiv:2411.11496, arXiv:2411.11496, November 28, 2024, accessed July 15, 2025, arXiv: 2411.11496[cs], http://arxiv.org/abs/2411.11496

142 Cong-Duy Nguyen et al., *CutPaste&Find: Efficient Multimodal Hallucination Detector with Visual-aid Knowledge Base*, Issue: arXiv:2502.12591, arXiv:2502.12591, February 18, 2025, accessed July 15, 2025, arXiv: 2502.12591[cs], http://arxiv.org/abs/2502.12591

143 Kun Wang et al., *A Comprehensive Survey in LLM(-Agent) Full Stack Safety: Data, Training and Deployment*, Issue: arXiv:2504.15585, arXiv:2504.15585, June 9, 2025, accessed July 15, 2025, arXiv: 2504.15585[cs], http://arxiv.org/abs/2504.15585

144 Shuai Zhao et al., *Unlearning Backdoor Attacks for LLMs with Weak-to-Strong Knowledge Distillation*, Issue: arXiv:2410.14425, arXiv:2410.14425, May 20, 2025, accessed July 15, 2025, arXiv: 2410.14425[cs], http://arxiv.org/abs/2410.14425

145 "Research | School of Computing and Information Systems," Singapore Management University, July 15, 2025, accessed July 15, 2025, https://computing.smu.edu.sg/research

146 Mengdi Zhang et al., *LLMScan: Causal Scan for LLM Misbehavior Detection*, Issue: arXiv:2410.16638, arXiv:2410.16638, May 25, 2025, accessed July 15, 2025, arXiv: 2410.16638[cs], http://arxiv.org/abs/2410.16638

147 Haoyu Wang, Christopher M. Poskitt, and Jun Sun, *AgentSpec: Customizable Runtime Enforcement for Safe and Reliable LLM Agents*, Issue: arXiv:2503.18666, arXiv:2503.18666, April 7, 2025, accessed July 15, 2025, arXiv: 2503.18666[cs], http://arxiv.org/abs/2503.18666

148 Wei Zhao et al., *Defending Large Language Models Against Jailbreak Attacks via Layer-specific Editing*, Issue: arXiv:2405.18166, arXiv:2405.18166, June 14, 2024, accessed July 15, 2025, arXiv: 2405.18166[cs], http://arxiv.org/abs/2405.18166

149 "Design • AI," Singapore University of Technology and Design (SUTD), July 15, 2025, accessed July 15, 2025, https://www.sutd.edu.sg/about/design-ai/

150 Soujanya Poria, "The Blue Sky Research in LLM and Alignment" (Presenters: _:n2188, July 15, 2025), accessed July 15, 2025, https://docs.google.com/presentation/d/17edmwBfMHhduRq4KSKrXz8_pK_aokH0SHawiR4ZNK9U

151 Rima Hazra et al., *Sowing the Wind, Reaping the Whirlwind: The Impact of Editing Language Models*, Issue: arXiv:2401.10647, arXiv:2401.10647, May 16, 2024, accessed July 15, 2025, arXiv: 2401.10647[cs], http://arxiv.org/abs/2401.10647

152 Tej Deep Pala et al., *Ferret: Faster and Effective Automated Red Teaming with Reward-Based Scoring Technique*, Issue: arXiv:2408.10701, arXiv:2408.10701, August 20, 2024, accessed July 15, 2025, arXiv: 2408.10701[cs], http://arxiv.org/abs/2408.10701

153 Prannaya Gupta et al., *WalledEval: A Comprehensive Safety Evaluation Toolkit for Large Language Models*, Issue: arXiv:2408.03837, arXiv:2408.03837, August 19, 2024, accessed July 15, 2025, arXiv: 2408.03837[cs], http://arxiv.org/abs/2408.03837

154 Soujanya Poria, "About me," accessed July 15, 2025, https://soujanyaporia.github.io/

155 "A*STAR Research Overview," A*STAR HQ Corporate Website, July 15, 2025, accessed July 15, 2025, https://www.a-star.edu.sg/Research/overview

156 "A*STAR Research Focus," 00. A*STAR HQ Corporate Website, July 15, 2025, accessed July 15, 2025, https://www.a-star.edu.sg/Research/research-focus

157 "Artificial General Intelligence," 56. Centre for Frontier AI Research (CFAR), July 15, 2025, accessed July 15, 2025, https://www.a-star.edu.sg/cfar/research/research-pillars/artificial-general-intelligence

158 "Institute for Infocomm Research (I2R), Singapore's largest ICT research institute, collaborates with globally competitive companies and invents impactful technologies anchored by scientific research," Asia Research News, May 8, 2013, accessed July 15, 2025, https://www.asiaresearchnews.com/content/institute-infocomm-research-i2r-singapore%E2%80%99s-largest-ict-research-institute-collaborates

159 "MERaLiON is Available for Download from Hugging Face," 14. Institute for Infocomm Research (I2R), December 26, 2024, accessed July 15, 2025, https://www.a-star.edu.sg/i2r/research/I2RTechs/research/i2r-techs-solutions/meralion-is-available-for-download-from-hugging-face

160 "Machine Intellection," 14. Institute for Infocomm Research (I2R), July 15, 2025, accessed July 15, 2025, https://www.a-star.edu.sg/i2r/research-capabilities/machine-intellection

161 "Vision & Mission," 18. Institute of High Performance Computing (IHPC), July 15, 2025, accessed July 15, 2025, https://www.a-star.edu.sg/ihpc/about-us/vision-mission

162 "Our Digital Government efforts," Government Technology Agency (GovTech), July 15, 2025, accessed July 15, 2025, https://www.tech.gov.sg/about-us/what-we-do/our-digital-government-efforts/

163 "Data Science and Artificial Intelligence," Singapore Government Developer Portal, July 15, 2025, accessed July 15, 2025, https://www.developer.tech.gov.sg/products/collections/data-science-and-artificial-intelligence/

164 Gabriel Chua, Shing Yee Chan, and Shaun Khoo, *A Flexible Large Language Models Guardrail Development Methodology Applied to Off-Topic Prompt Detection*, Issue: arXiv:2411.12946, arXiv:2411.12946, April 9, 2025, accessed July 15, 2025, arXiv: 2411.12946[cs], http://arxiv.org/abs/2411.12946

165 Isaac Lim et al., *Safe at the Margins: A General Approach to Safety Alignment in Low-Resource English Languages – A Singlish Case Study*, Issue: arXiv:2502.12485, arXiv:2502.12485, April 8, 2025, accessed July 15, 2025, arXiv: 2502.12485[cs], http://arxiv.org/abs/2502.12485

166 Jessica Foo, Pradyumna Shyama Prasad, and Shaun Khoo, *Know Or Not: a library for evaluating out-of-knowledge base robustness*, Issue: arXiv:2505.13545, arXiv:2505.13545, May 19, 2025, accessed July 15, 2025, arXiv: 2505.13545[cs], http://arxiv.org/abs/2505.13545

167 "AI Safety Institute (AISI)," Digital Trust Centre (DTC), July 15, 2025, accessed July 15, 2025, https://www.ntu.edu.sg/dtc/aisi

168 Nanyang Technological University, "Our People," Digital Trust Centre (DTC), accessed July 19, 2025, https://www.ntu.edu.sg/dtc/our-people

169 "AI Governance Research Grant Call - AI Singapore," April 12, 2023, accessed July 15, 2025, https://aisingapore.org/governance/grant-call2023/

170 "AI Research Grant Calls - AI Singapore," September 5, 2022, accessed July 15, 2025, https://aisingapore.org/research/grant-call/

171 Mengdi Zhang et al., *LLMScan: Causal Scan for LLM Misbehavior Detection*, Issue: arXiv:2410.16638, arXiv:2410.16638, May 25, 2025, accessed July 15, 2025, arXiv: 2410.16638[cs], http://arxiv.org/abs/2410.16638

172 Yufei He et al., *Evaluating the Paperclip Maximizer: Are RL-Based Language Models More Likely to Pursue Instrumental Goals?*, Issue: arXiv:2502.12206, arXiv:2502.12206, February 16, 2025, accessed July 15, 2025, arXiv: 2502.12206[cs], http://arxiv.org/abs/2502.12206

173 Noemi Dreksler et al., "What the public thinks about AI and the implications for governance," Brookings, April 9, 2025, accessed July 15, 2025, https://www.brookings.edu/articles/what-the-public-thinks-about-ai-and-the-implications-for-governance/; Nestor Maslej et al., *The AI Index 2025 Annual Report* (Stanford, CA: Institute for Human-Centered AI (HAI), Stanford University, April 2025), https://hai.stanford.edu/assets/files/hai_ai_index_report_2025.pdf

174 "Ipsos AI Monitor 2024," Ipsos AI Monitor, June 2024, accessed July 15, 2025, https://www.ipsos.com/sites/default/files/ct/news/documents/2024-06/Ipsos-AI-Monitor-2024-final-APAC.pdf

175 Nicole Gillespie et al., *Trust, attitudes and use of artificial intelligence: A global study 2025*, Artwork Size: 4974511 Bytes (The University of Melbourne, 2025), 4974511 Bytes, accessed July 19, 2025, https://doi.org/10.26188/28822919, https://assets.kpmg.com/content/dam/kpmgsites/xx/pdf/2025/05/trust-attitudes-and-use-of-ai-global-report.pdf

176 Karen Wong, "Survey: 48% of HK, SG and ID consumers concerned about loss of human touch in AI," Marketing-Interactive, May 15, 2025, accessed July 19, 2025, https://www.marketing-interactive.com/survey-48-of-hk-sg-id-consumers-concerned-about-loss-of-human-touch-in-ai

177 Nicole Gillespie et al., *Trust, attitudes and use of artificial intelligence: A global study 2025*, Artwork Size: 4974511 Bytes (The University of Melbourne, 2025), 4974511 Bytes, accessed July 19, 2025, https://doi.org/10.26188/28822919, https://assets.kpmg.com/content/dam/kpmgsites/xx/pdf/2025/05/trust-attitudes-and-use-of-ai-global-report.pdf

178 YouGov, "Generative AI in media: Consumer sentiment across Hong Kong, Indonesia and Singapore," 2025, accessed July 19, 2025, https://commercial.yougov.com/rs/464-VHH-988/images/GenAIMediaReport_HKIDSG_PDF_English_YouGov.pdf?version=1