CONCORDIA AI
安远AI

State of AI Safety in Singapore

May 2026

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, Kwan Yee Ng, and Brian Tse.

How to cite this report

Jonathan Lee, Kwan Yee Ng, and Brian Tse, "State of AI Safety in Singapore (2026)," Concordia AI, May 2026.

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
- Yifan Jia, AIDX Tech
- Mohan Kankanhalli, National University of Singapore
- Kwok-Yan Lam, Nanyang Technological University
- Shashvat Shukla, University College London
- Zhi Xuan Tan, National University of Singapore
- Edward Yee, FAR.AI

Table of Contents

About Concordia AI i

Authors i

Acknowledgements ii

Executive Summary 1

Introduction 3

1 Domestic Approach 5
- 1.1 Voluntary Frameworks . . . 7
- 1.2 AI Safety Testing and Assurance . . . 9
- 1.3 Hard Regulations . . . 12
- 1.4 Standards . . . 14

2 International Approach 16
- 2.1 Multilateral Initiatives . . . 17
- 2.2 Regional Initiatives . . . 21
- 2.3 Bilateral Efforts . . . 22

3 Homegrown AI Ecosystem and Industry 24
- 3.1 Homegrown GPAI ecosystem . . . 25
- 3.2 Third-party AI Assurance Suppliers . . . 29
- 3.3 Foreign AI Developers . . . 30

4 Technical Research 32
- 4.1 Universities . . . 34
- 4.2 Public-sector and Research Institutions . . . 45
- 4.3 Overall Trends . . . 48

Conclusion 51

Appendix 52

Notes 55

Executive Summary

Achieving AI's global benefits and managing its risks requires broad international cooperation, yet current global debates concentrate disproportionately on the few nations building frontier AI systems. Singapore shows that smaller, resource-constrained states can influence emerging AI safety norms.

Since our 2025 report, Singapore's AI safety ecosystem has demonstrated a growing emphasis on implementation-focused governance, with the release of governance frameworks such as the *Model AI Governance Framework (MGF) for Agentic AI*. It has also codified testing methodologies and is working to translate them into international standards. This report, current to early May 2026, surveys Singapore's AI safety ecosystem across four domains: its domestic approach to AI governance, its international approach, its homegrown AI ecosystem and industry, and its technical AI safety research.

Domestic Approach

- **Agentic AI governance has become a distinct domestic priority.** Key developments across the different government agencies include the Infocomm Media Development Authority's (IMDA) *MGF for Agentic AI* and planned agentic AI testing guidelines, GovTech's *Agentic Risk and Capability Framework*, and the Cyber Security Agency's (CSA) agentic AI security work.
- **Singapore's domestic governance is becoming more testing-driven and harm-specific.** Safety infrastructure expanded through the Global AI Assurance Sandbox and the *Starter Kit for Testing LLM-Based Applications*, and new initiatives include the AI Safety Red Teaming Challenge in January 2026, which addressed application-layer data leakage risks. At the same time, hard law remains targeted at specific harms: the Online Safety Act 2025 addresses AI-enabled online harms through takedown, access-restriction, and victim-redress mechanisms.

International Approach

- **Singapore's international strategy is becoming more oriented toward practical tools and standards.** Across the United Nations (UN), World Economic Forum (WEF), the International Organization for Standardization (ISO), the global AI summit series, and the International Network for Advanced AI Measurement, Evaluation and Science (Network), Singapore is increasingly contributing practical tools, testing methodologies, evaluation work, and standards proposals. This has included joint agentic AI testing exercises with other AI Safety Institutes and the proposed ISO/IEC 42119-8 international standard on generative AI testing at the SC 42 plenary.
- **Singapore also used its convening role to connect international AI safety debates with regional implementation.** Through the SC 42 plenary, the International Scientific Exchange on AI Safety 2026, the Network, and the ASEAN Working Group on AI Governance (WG-AI), Singapore is working to translate emerging global norms on testing and evaluation into tools and benchmarks applicable to Southeast Asia.

Homegrown AI Ecosystem and Industry

- **Singapore's homegrown general-purpose AI (GPAI) ecosystem advanced in both capabilities and safety tooling.** SEA-LION v4 expanded into multimodal image-text capabilities and lighter edge-deployable models, while MERaLiON-3-preview broadened Singapore's speech-AI capabilities. At the same time, new tools such as SEA-Guard, LionGuard 2, and AI Guardian show that safety is being embedded across the homegrown ecosystem, though the focus remains mainly on multilingual moderation, toxicity prevention, and deployment safeguards rather than broader AI safety risks.
- **Foreign GPAI developers are becoming more integrated into Singapore's AI eco- system, but mostly through partnerships aimed at internal model deployments.** Companies such as Google, Mistral AI, Microsoft, and Alibaba are working with government agencies to develop and test models, multilingual and regional adaptation, and safety benchmarking. However, these models remain tied to specific government or internal use cases, rather than being adopted publicly or across the ecosystem.

Technical AI Safety Research

- **Technical AI safety research capacity is growing, with strengths in safety evaluation and multilingual deployment.** The number of lead AI safety researchers increased significantly from 14 to 21, with additions concentrated at the National University of Singapore (NUS), Nanyang Technological University (NTU), and the Agency for Science, Technology and Research (A*STAR), while a wider community of research fellows, graduate students, and co-authors supported their output. These scholars' research on evaluation benchmarks and multilingual safety aligns with the AI safety ecosystem's broader focus on developing AI models for low-resource regional languages and assessing risks such as toxicity in these models.
- **Research attention is shifting toward agentic and system-level risks, but foundational safety gaps remain.** Papers now cover LLM agents, multi-agent systems, and downstream risks. However, research remains weighted toward empirical attack-and-defense, with less work on theoretical safety, interpretability, formal verification, scalable oversight, and foundational alignment. This suggests that Singapore has strengths in empirical research that are well matched to its role as a testing and assurance hub, but it is less developed in deeper safety-by-design research for advanced AI systems.

Introduction

How has Singapore's AI safety ecosystem evolved since July 2025, and what significance does this hold for global AI governance? This question has become more important as international AI safety discussions increasingly focus not only on broad governance principles, but also on how safety should be assessed and tested in practice. Singapore offers a useful case study of how this is taking shape.

Up to 2024, Singapore's AI governance publications—the *Model AI Governance Framework* (January 2019), its 2020 update, and the *Model AI Governance Framework for Generative AI* (May 2024)—were generally aimed at translating broad responsible AI principles into practical organizational guidance. This period also saw targeted AI-related legislation and the development of technical evaluation toolkits, notably AI Verify and Project Moonshot.

Since mid-2025, Singapore's ecosystem has moved further toward more implementation-focused governance. This is visible in the *Model AI Governance Framework for Agentic AI* (January 2026), which is more directive and operationally granular than its predecessors, as well as publications such as the *Agentic Risk and Capability Framework* (December 2025). That said, both agentic AI documents remain voluntary in nature—there is a shift in specificity and orientation, but no move toward binding regulation.

At the same time, Singapore's AI safety work has become more explicitly focused on codifying testing methodologies: the *Starter Kit for Testing LLM-Based Applications* (January 2026) was finalized, and the Global AI Assurance Sandbox continued following its 2025 Pilot. These developments suggest a maturing ecosystem in which Singapore is not only conducting testing and assurance exercises, but also seeking to formalize how such work should be carried out.

Since the 2025 report, Singapore has remained active across a range of multilateral, regional, and bilateral processes, but with a stronger emphasis on evaluation, measurement, and assurance. This can be seen in its contributions to the International Network for Advanced AI Measurement, Evaluation and Science, its efforts within the Association of Southeast Asian Nations (ASEAN) region to explore safety testing tools and regional benchmarks, and bilateral technical cooperation, such as the work by the Singapore and Korea AI Safety Institutes on data-leakage risks in AI agents. In parallel, developments in Singapore's homegrown AI ecosystem and technical research base indicate a broader layering of safety work across the stack, from multilingual safeguards and deployment tools to a growing body of research on agentic and system-level risks.

This report examines Singapore's AI safety ecosystem across four interconnected areas:

1. **Domestic Approach**: Singapore's governance architecture has continued to develop through voluntary frameworks, practical testing and assurance initiatives, targeted legislation, and standards -setting, with particular emphasis on agentic AI and on codifying testing methodologies.

2. **International Approach**: Singapore has used multilateral, regional, and bilateral channels to shape AI governance discussions through practical contributions in testing, evaluation, and assurance.
3. **Homegrown AI Ecosystem and Industry**: Singapore's local model ecosystem spans both public-sector and industry developments, including homegrown general-purpose AI model development, public-sector safety tooling, third-party assurance suppliers, and the role of foreign AI developers in Singapore.
4. **Technical Research:** This section surveys the current state of AI safety research in Singapore, highlighting areas of comparative strength, emerging trends, and the gaps that remain in the research base.

Scope

This report focuses on the risks posed by advanced general-purpose AI (GPAI) systems. It adopts the terminology of the *International AI Safety Report 2026* and the *Singapore Consensus on Global AI Safety Research Priorities*,[1,2] both of which use "general-purpose AI" to refer to AI systems with broad capabilities across a wide range of tasks. In this report, GPAI includes generative AI systems and, where relevant, agentic AI systems built on such models. The inclusion of agentic systems builds on the 2025 report and to better reflect ongoing developments in the field, including growing policy and technical attention to systems capable of acting with greater autonomy.

We examine three broad categories of risk arising from GPAI systems: malicious use, malfunctions, and systemic risks. Malicious-use risks include, for example, deceptive synthetic content like deepfakes, and cyber or biological misuse. Malfunction risks include reliability failures, bias, security failures, loss of control or oversight in deployment, and failures arising in agentic or multi-agent systems. Systemic risks include broader labor-market, social, geopolitical, and environmental effects.

We do not attempt a comprehensive account of all AI-related regulation or policy in Singapore. Accordingly, the report does not systematically examine sector-specific frameworks governing narrower applications, including in areas such as law, finance or healthcare,[a] and does not treat broader trade-related measures as part of its core scope (except where they bear directly on the governance of GPAI systems).[b]

The report draws on official government publications, standards documents, technical and policy papers, industry materials, and other publicly available sources. Because the analysis is based on public information, it necessarily provides only a partial picture: some initiatives may remain non-public, and in fast-moving areas, the picture may change quickly. Our research cut-off was early May 2026; 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. We note the developments in these sectors such as the Launch of Guide for Using Generative Artificial Intelligence in the Legal Sector in March 2026.[3]

b. This includes advisories like the 2025 Joint Advisory: Export controls on advanced semiconductor and artificial intelligence (AI) technologies, which will not be covered in the report.[4]

Domestic Approach

Key takeaways

- Agentic AI governance emerged as a distinct priority. This shift is visible across IMDA's *Model AI Governance Framework for Agentic AI (MGF-Agentic AI)*, CSA's *Securing Agentic AI: A Discussion Paper* and a draft *Addendum to the Guidelines and Companion Guide on Securing AI Systems*, GovTech's *Agentic Risk & Capability (ARC) Framework*, and planned testing guidelines for agentic AI applications.
- The *MGF-Agentic AI* marks a shift from broad responsible AI principles toward more practical, operationally specific guidance for governing high-autonomy systems, while remaining a voluntary framework.
- Since the last report, Singapore has expanded its AI safety testing infrastructure. It launched the Global AI Assurance Sandbox as a successor to the earlier pilot, hosted the 2026 AI Safety Red Teaming Challenge, and expanded its multilingual, deployment-stage testing suite, AILuminate.
- Singapore's use of hard law has expanded incrementally to target specific categories of AI-enabled online harms and cybersecurity risks. The Online Safety Act 2025 adds takedown powers, access-restriction mechanisms, and victim-redress provisions to hold platforms accountable. Recent CSA action suggests that local Critical Information Infrastructure owners' cybersecurity obligations may become an early focus of formal AI-related oversight, particularly in response to AI-enabled cyber threats.
- Singapore's AI standards regime has expanded through new standards on AI Management System certification bodies (SS 42006) and real-world AI use cases (TR 139). However, these developments primarily support certification infrastructure and AI adoption, rather than introducing new technical safety requirements.

Singapore's AI governance landscape is shaped by a cluster of public bodies spanning policy, regulation, standards, implementation, and research. At the center is the Ministry of Digital Development and Information (MDDI) (called the Ministry of Communications and Information until 2024). MDDI has positioned AI as part of Singapore's broader digital-development agenda, most notably through the National AI Strategy 2.0, launched in 2023 to drive AI development, adoption, and capability-building. MDDI works to situate AI within Singapore's wider national-level objectives, while coordinating the agencies that handle its more specific governance, implementation, and security dimensions. These agencies include the Infocomm Media Development

Figure 1.1: Singapore's public bodies working on AI safety and governance

Authority (IMDA), Personal Data Protection Commission (PDPC), Government Technology Agency of Singapore (GovTech), and Cyber Security Agency of Singapore (CSA).

Among these, IMDA has been centrally involved in AI governance, developing Singapore's *Model AI Governance Frameworks* (MGFs) and local testing and assurance initiatives. IMDA also sits at the center of Singapore's broader AI safety and assurance architecture: it established the AI Verify Foundation (AIVF) in 2023 to develop open source testing tools and assurance approaches. AIVF has worked together with IMDA on initiatives such as the Generative AI Evaluation Sandbox and Global AI Assurance Pilot.

IMDA also maintains close institutional links with the Singapore AI Safety Institute (AISI). The Nanyang Technological University's Digital Trust Centre is designated as the Singapore AISI and drives much of its technical work, but IMDA is responsible for policy development, international engagements, and partnerships with other AISIs and government agencies.[5] This link is also reflected organizationally, with the AISI's Head of Policy sitting within IMDA.

PDPC has contributed guidance on personal data, particularly in the earlier phases of Singapore's AI governance work, such as the first edition of the MGF in 2019. CSA focuses on cybersecurity risks and contributes to the cybersecurity dimension of AI governance. GovTech, for its part, operationalizes these principles across the public sector, for example through publishing practical safety tools and engineering resources (see the *Homegrown AI Ecosystem and Industry* section).

Singapore's AI standards architecture sits alongside this governance structure. Enterprise Singapore (ESG) administers the national standardization program, providing funding, offering secretariat support to the Singapore Standards Council (SSC), and coordinating Singapore's participation in international standards develop-

ment. The SSC oversees standards development across sectors and also coordinates Singapore's participation in international standards work.[6] Within SSC, the Information Technology Standards Committee (ITSC) leads national standardization activities such as the development of local technical standards and adoption guidelines; it is appointed by SSC under ESG's national standardization program and receives technical support from IMDA.[7] The AI Technical Committee (AITC), which sits under ITSC, carries out detailed AI standards work: it recommends adoption of relevant international AI standards, supports the development of new standards where needed, promotes awareness of AI standards, and represents Singapore in *ISO/IEC JTC 1/SC 42 on Artificial Intelligence*.[8]

1.1 Voluntary Frameworks

Singapore's domestic approach to AI governance has continued to rely primarily on voluntary frameworks and guidance rather than binding legislation. Its earliest official publication on AI governance was the 2018 *Discussion Paper on AI and Personal Data*, which initiated consultation with industry and government stakeholders and laid the foundation for Singapore's subsequent MGF. Over time, subsequent updates to the framework expanded its coverage from traditional AI to generative AI. These voluntary frameworks reflect Singapore's efforts to translate broad principles of responsible AI into practical guidance for organizations, first in relation to traditional AI systems and later in response to the distinctive risks posed by generative AI.

Table 1.1: Voluntary AI Governance Guidelines

Voluntary AI Governance Guidelines		
Date	**Publication**	**Released by**
Jun 2018	Discussion Paper on AI and Personal Data	PDPC
Jan 2019	Model AI Governance Framework	IMDA, PDPC
Jan 2020	Model AI Governance Framework Update • Companion to the MGF — Implementation and Self-Assessment Guide for Organizations • Compendium of Use Cases — Practical Illustrations of the MGF	IMDA, PDPC
Jun 2023	Discussion Paper on Generative AI — Implications for Trust and Governance	IMDA[a]
May 2024	Model AI Governance Framework for Generative AI	IMDA, AIVF
Jan 2026	Model AI Governance Framework for Agentic AI	IMDA

a. In collaboration with Aicadium, an AI Centre of Excellence under Temasek, a state-owned investment firm in Singapore.

1.1.1 Model AI Governance Framework for Agentic AI

In 2026, the MGF evolved further to cover agentic AI. On 22 January 2026, IMDA launched the ***Model AI Governance Framework for Agentic AI (MGF-Agentic AI)*** as a guide for organizations seeking to deploy agentic AI, either through in-house development or third-party solutions (See Table 1.1).

The *MGF-Agentic AI* makes clear why agentic AI systems require a distinct governance approach. Unlike generative AI systems that primarily produce outputs, agentic systems can take actions, adapt to new information, and interact with other agents and external systems in order to complete tasks on behalf of users. These expanded capabilities create new risk vectors: because agents may access sensitive data, call tools, and alter their environment, they can generate not only incorrect outputs but also erroneous, unauthorized, or cascading real-world actions. [9] The framework therefore highlights agent-specific risk considerations, such as agents' access to sensitive data, their level of autonomy, and the need to limit risk through design choices, including boundaries on agents' tool use and data access.

The framework's governance approach is structured around four broad areas: assessing and bounding risks upfront, making humans meaningfully accountable, implementing technical controls and processes, and enabling responsible end-user use. As in the earlier *MGF for Generative AI*, human accountability remains a core principle present in the MGFs. At the same time, the framework recognizes that meaningful human control and oversight must be integrated throughout the lifecycle, while also acknowledging that continuous human oversight over all agent workflows becomes impractical at scale. It therefore adopts a full-lifecycle approach to technical and organizational controls, encompassing guardrails at the design stage, safety and security testing before deployment, continuous monitoring and logging after deployment, and measures to ensure that end users are equipped to use these systems responsibly.

In this respect, *MGF-Agentic AI* also reflects a broader shift in Singapore's AI governance approach. Earlier iterations of the MGF were more strongly oriented toward high-level, principle-based guidance: they set organizational expectations based on broad responsible AI principles, but generally remained relatively open-ended in how these expectations should be operationalized.[b] *MGF-Agentic AI*, by contrast, is more specific and implementation-oriented by articulating expectations around risk assessment, technical controls, testing, monitoring, and end-user responsibility.[10]

Another notable difference lies in the publication process. For earlier MGFs, the Singapore government typically published consultation or discussion papers to seek broader feedback before releasing the framework roughly six to nine months later. By contrast, *MGF-Agentic AI* was published after private consultation with government agencies and private-sector organizations, and was described in its press release as a "living document." This framing could suggest a more iterative mode of governance for agentic AI systems, in which the framework may continue to evolve alongside emerging industry practice and technical developments, rather than being shaped primarily through a discrete public consultation phase prior to release.[11]

b. For example, the *MGF for Generative AI states that* "it is important that the industry coalesces around best practices in development and safety evaluation," and that model developers and deployers are "best placed to determine what safety measures to use."

1.1.2 Guidelines to Securing AI Systems

In October 2025, CSA published ***Securing Agentic AI: A Discussion Paper*** together with a draft ***Addendum to the Guidelines and Companion Guide on Securing AI Systems*** for public consultation. This pairing fits a broader Singapore pattern in AI governance Table 1.1: an initial discussion paper is used to surface new risks and frame the problem, before more operational guidance is issued.

The *Discussion Paper* frames the security problem posed by agentic AI. It explains how agentic systems that can plan, act, and use tools with reduced human prompting create new attack surfaces, and why conventional controls remain necessary but insufficient.[12] The draft *Addendum* then turns this analysis into operational guidance. Designed to be read with CSA's earlier *Guidelines* and *Companion Guide*, it helps organizations assess agentic AI risks through a lifecycle approach, including workflow mapping, threat-vector identification, and practical controls across development, deployment, operations, and end-of-life stages.

1.2 AI Safety Testing and Assurance

Rather than concentrating solely on frontier model safety testing, Singapore has built out a range of practical assurance tools, open-source testing resources, and deployment-stage evaluation initiatives covering both base models and real-world applications.

Led by the AI Verify Foundation (AIVF), Singaporean testing and assurance initiatives have included: the AI Verify Toolkit, which allows organizations to conduct self-assessment on traditional AI systems; the open source Project Moonshot platform, which allows developers to benchmark and red-team LLM and applications; the Global AI Assurance Sandbox for real-world testing of generative AI applications with third-party testers; and the co-development of AILuminate benchmark with MLCommons. These earlier initiatives show that while other countries tend to focus on upstream frontier-model evaluations, Singapore's approach has leaned more heavily toward application-level assurance and downstream safety testing.

1.2.1 Global AI Assurance Sandbox

In the 2025 report, we noted the launch of the **Global AI Assurance Sandbox** in July 2025 as a follow-up to the Global AI Assurance Pilot. Although information on the specific testing case studies of the Sandbox itself has not been released, the Pilot gives a useful sense of what this looks like in practice: it paired 17 application deployers with 16 specialist testing firms to test real-world generative AI applications such as a scam and online fact-checker, customer service and public-sector chatbots, AI-enabled candidate screening tools, investment and wealth-management assistants, medical report summarization tools, and multilingual internal knowledge bots.[13]

The Pilot was organized around four baseline categories: hallucination, undesirable content, data disclosure, and vulnerability to adversarial prompts. The Sandbox has since pre-qualified 13 testing firms to be selected as testing partners, and is also open to sector-specific regulators so that they can develop and obtain real-world feedback on AI governance and testing approaches.

1.2.2 AILuminate with MLCommons

In the 2025 report, we noted that AIVF had signed a Memorandum of Intent with MLCommons to collaborate on a common set of safety testing benchmarks for generative AI models. This collaboration centered on **AILuminate**, MLCommons' safety benchmark for general-purpose chat systems, which is designed to test model responses across multiple hazard categories.[c] AILuminate v1.0 was released in December 2024, followed by AILuminate v1.1 French in February 2025. In May 2025, MLCommons and AIVF also announced proof-of-concept Chinese-language benchmarking work, developed in collaboration with National University of Singapore (NUS).

Since then, AILuminate has expanded beyond text-only safety benchmarking. Newer workstreams include multimodal evaluation, which tests safety in text-and-image interactions across languages and cultures, and jailbreak evaluation, which measures resilience to adversarial attacks. The multimodal benchmark was publicly outlined in March 2026, with an initial release planned for Summer 2026, while AILuminate Jailbreak v0.5 was released in October 2025, with v1.0 planned for Q1 2026. AIVF and IMDA have been involved in these newer efforts, particularly in supporting multilingual and culturally grounded evaluation, including work in languages such as Tamil and Malay designed to reflect local contexts and concerns.[14][15]

1.2.3 Testing Frameworks

Beyond conducting practical testing exercises (such as the Global AI Assurance Sandbox) and creating benchmarks, Singapore is also codifying how AI systems should be tested. Two recent examples are the ***Starter Kit for Testing LLM-Based Applications for Safety and Reliability (Starter Kit)***, which provides step-by-step guidance for pre-deployment testing of LLM-based applications against baseline risks, and GovTech's ***Agentic Risk and Capability (ARC) Framework***, which offers a structured way to identify, assess, and mitigate safety and security risks in agentic AI systems.

A major update since the 2025 report was the finalization of the *Starter Kit* in January 2026. Whereas the earlier consultation draft in May 2025 focused on four risks—hallucination, undesirable content, data disclosure, and vulnerability to adversarial prompts—the January 2026 Version 1.0 expanded the risk categories: "hallucination" is broadened to "hallucination and inaccuracy", and there is a new fifth risk category, "bias in decision making."[16] IMDA describes Version 1.0 as a first step toward "codifying standards for AI testing and assurance," and states that the recommended testing methodologies will be progressively made available through Project Moonshot.[17] Project Moonshot is explicitly positioned as a way to help organizations implement the *Starter Kit*, by connecting applications to benchmark tests and evaluators recommended in the guidelines.

When IMDA launched *MGF-Agentic AI* in January 2026, it also stated that it was developing guidelines for testing agentic AI applications.[18] Although these testing guidelines have not yet been released as of the report

c. It includes a safety assessment standard, hazard taxonomy, response evaluation criteria, and a human-generated prompt dataset designed to test hazardous scenarios. The dataset includes public practice prompts, private benchmarking prompts, and demo prompts, with both adversarial and non-adversarial examples.

cutoff date in May 2026, this announcement, and the fact that the *Starter Kit* is named "Version 1.0", suggest that Singapore could be preparing to include the testing of agentic systems in later versions of the *Starter Kit*.

In parallel, GovTech published the *ARC Framework* in December 2025. ARC is best understood as a technical governance framework for identifying, assessing, and mitigating safety and security risks in agentic AI systems. Rather than serving as a benchmark or test suite on its own, it provides a structured way to analyze where risks arise in agentic systems—across their components, design choices, and capabilities—and to link these risks to practical technical controls. The framework is oriented toward potentially harmful agentic behaviors and high-consequence hazards, including risks such as Chemical, Biological, Radiological, Nuclear, and Explosives-related (CBRNE) misuse, compromised systems, and user endangerment.[19]

1.2.4 Singapore AI Safety Red Teaming Challenge

Singapore's safety-testing work broadened this year with a new red-teaming project. Building on a 2024 multilingual red-teaming exercise on cultural bias, the **Singapore AI Safety Red Teaming Challenge 2026**, led by IMDA, shifted to application-layer data leakage risks in generative AI applications. Participants attempted to extract sensitive information from seven simulated generative AI applications—presented in English and regional languages—using a jeopardy-style capture-the-flag format. The exercise drew over 80 participants from 14 Asian countries.[20] Initial observations point to three practical concerns: weakly secured applications can be induced to leak data through relatively simple prompting; application-level protections can behave inconsistently because of generative AI's probabilistic nature; and safety performance can degrade sharply across languages. As an example of the latter, extracting protected data from a particular application took 1.09 hours and 58 techniques in English, but the attack succeeded instantly with a single Khmer prompt. Through this continued emphasis on multilingual, application-layer testing, we can see that Singapore's safety efforts are geared toward concrete downstream harms rather than abstract model-level concerns. A full Challenge report is set to be published later in 2026.

The Challenge is also becoming more institutionalized and regionally connected. Coverage expanded from nine countries in the 2024 edition to 14 in 2026, covering all ASEAN countries in addition to China,[d] India, Japan, and South Korea. The Challenge also drew observers from the ASEAN Working Group on AI Governance (WG-AI) and the International Network for Advanced AI Measurement, Evaluation and Science (Network),[e] indicating that its findings may feed into wider testing efforts by these bodies.

d. Concordia AI participated in this challenge as a partner institute from China.
e. Japan AISI participated in the challenge, while Korea AISI and France AISI were observers.

1.2.5 Singapore AI Safety Institute

The **Singapore AI Safety Institute** (AISI) is also a key player in the safety testing ecosystem, focusing on areas such as testing and evaluation, safe model design and deployment, content assurance, and governance and policy. A key technical workstream for Singapore AISI is leading joint testing efforts with international partners, either bilaterally with other AISIs or through the Network[f]. Updates on its testing initiatives and results will be covered in the *International Approach* section.

Singapore's testing and assurance activity over the past year has begun to extend beyond generative AI applications toward agentic risks—with the announcement of forthcoming testing guidelines for agentic AI applications signalling that the same emphasis on downstream, application-level evaluation is being carried forward into this newer domain.

1.3 Hard Regulations

Singapore does not have a national AI-specific law, nor a comprehensive hard regulatory regime directed at AI systems as such or at frontier AI risks. Instead, its use of hard law remains targeted and harm-specific.[g] In the 2025 report, we noted legislative developments such as the *Penal Code* and the *Elections (Integrity of Online Advertising) Amendment Bill*, both of which addressed deepfakes and other misuse of AI-generated material. The *Penal Code* has since been amended in December 2025 to specify that creation, possession, and distribution of sexually explicit, voyeuristic, or child abuse material that is AI-generated is also an offence.[21] Another key update has been the *Online Safety (Relief and Accountability) Act 2025*. This adds another AI-relevant legal instrument to Singapore's regulatory framework, though it is not a dedicated AI statute.

1.3.1 Online Safety (Relief and Accountability) Act 2025

The ***Online Safety (Relief and Accountability) Act 2025***, passed in November 2025, creates legal mechanisms to address harms caused by synthetic and manipulated content online. The Act notes generative AI as one example of digital means by which content may be altered or generated, and then used to harass, humiliate, impersonate, or otherwise harm individuals. The Act is significant because it establishes a stronger remedial and enforcement framework around these harms. It also creates an Online Safety Commission with powers to issue notices, directions, and orders in relation to harmful online material, including the abuse of inauthentic material (e.g. deepfakes).

The Commission is scheduled to be set up in the first half of 2026, led by the Commissioner of Online Safety.[22] Once established, it will have the powers to direct social media services (and other relevant actors) to remove or disable access to harmful material, restrict Singapore users' access to particular content or online locations, and in some cases support further access-blocking measures against non-compliant online locations.[23] The Act is designed not simply to prohibit harmful conduct, but to enable rapid mitigation and victim relief in the

f. The Network was formerly known as the International Network of AISIs.

g. 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 legislation on generative AI and general-purpose AI as outlined in the scope.

online environment. In practice, the key coercive measures are not AI-specific penalties imposed on model developers, but powers of content takedown, access restriction, corrective publication, and blocking, aimed at limiting the spread and impact of harmful material within Singapore.

The Act reinforces a broader pattern in Singapore's regulatory approach to AI: hard law is used to address concrete and socially salient harms—especially instances where AI-generated content creates risks of deception, humiliation, or abuse—while broader AI governance continues to rely on guidance and other voluntary instruments. There may be signs of tighter oversight emerging in some regulated sectors, such as finance, or in specific risk categories, like chatbots,[h][25] but these remain area-specific and do not constitute a movement towards a comprehensive national AI law.

1.3.2 Cybersecurity and Critical Infrastructure Oversight

On May 6 2026, Commissioner of Cybersecurity David Koh issued a letter to owners of Critical Information Infrastructure (CII)—that is, designated computer systems supporting essential services in sectors such as energy, water, and healthcare[i]— on the cybersecurity implications of frontier AI.[27] The letter warned that frontier AI had "materially shifted the cybersecurity baseline" for CIIs, citing recent models' ability to identify zero-day vulnerabilities and execute complex attacks autonomously.[28]

The letter matters because it is more targeted than a general advisory,[29] even if it does not appear to impose new legal duties. By addressing boards and senior leaders of CII owners directly, CSA is signalling that AI-enabled cyber risks should be treated as a matter of board oversight and management accountability rather than a technical security concern. This may offer an early indication of how AI-related risks could be incorporated into existing cybersecurity oversight for critical infrastructure. In the Parliamentary Sitting on May 5 2026 responding to questions on AI-cyber risks, Senior Minister of State (MDDI) Tan Kiat How said that CSA will review standards and obligations for CII owners to account for faster attack timelines enabled by AI.[30]

These developments suggest that Singapore's hard regulation approach to AI is likely to remain targeted and sectorally embedded. Rather than introducing a broad AI-specific regime, Singapore appears to be addressing AI risks through existing sectoral and cybersecurity frameworks where possible, especially where those risks affect critical infrastructure, national security, or other high-risk environments.

h. Although this report does not cover governance in industry-specific verticals, we note that some sectors have started discussion on stricter controls on AI.[24]

i. The 11 CII sectors are: Energy, Water, Banking and Finance, Healthcare, Transport (which includes Land, Maritime, and Aviation), Infocomm, Media, Security and Emergency Services, and Government.[26]

1.4 Standards

Singapore develops sector-specific standards through standards committees, technical committees, and working groups. Singapore Standards (SS) may be produced either through adoption of international standards or through domestic development, with consensus among representatives from government, industry, professional bodies, and other stakeholders. Where there is more immediate industry demand and no consensus standard yet exists, a Technical Reference (TR) may be issued. Technical References are voluntary, and they typically undergo a three-year testing period. After this period, they are either elevated to Singapore Standards, continue as Technical References, or are withdrawn.

Singapore's AI standards work is led through the Artificial Intelligence Technical Committee (AITC), which was established in July 2019 under the national standards system to support the development of AI standards in Singapore and Singapore's participation in *ISO/IEC JTC 1/SC 42 on Artificial Intelligence*.

In the 2025 report, we identified *TR 99:2021 Artificial Intelligence Security (TR 99)* and *SS ISO/IEC 42001:2024 Information Technology – Artificial Intelligence – Management System (SS 42001)* as the main Singapore standards most relevant to AI governance and safety. TR 99 was particularly significant as Singapore's first AI-specific standard, and it remains the most clearly safety-relevant standard. 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 being taken down). Four case studies were provided to illustrate possible attacks, in social media (content), finance (credit scoring), healthcare (diagnosis), and cybersecurity (malware detection). Although TRs can be superseded by a SS or withdrawn after three years, TR 99 continues to be listed as "Current", [31] and there is no clear public indication that it has been superseded or withdrawn. SS 42001, by contrast, is a broader AI management-system standard, oriented toward governance, organizational controls, and risk management.

1.4.1 New Standards

Since the 2025 report, the following new standards have been released:

- ***SS ISO/IEC 42006:2025*** - *Information technology - Artificial intelligence - Requirements*, for bodies providing audit and certification of AI Management Systems (AIMS); and
- ***TR 139:2025*** - *Artificial intelligence (AI) use cases.*

SS 42006, an identical adoption of ISO/IEC 42006:2025, sets out requirements for assessment bodies that certify organizations against SS 42001—the AI management system standard applicable to organizations that provide, develop, or use AI systems. The SS does not cover nor establish new technical requirements for the safe development or deployment of AI systems.

TR 139 compiles and describes real-world AI applications across 11 sectors in Singapore, and is best read as a tool for ecosystem development and AI diffusion rather than safety governance. While some of its examples relate to the safe deployment of AI, such as the use of LLMs for generative AI testing, these appear

as illustrations of practice rather than as part of a sustained effort to develop safety requirements or risk controls. In this respect, the standard reflects a policy emphasis on encouraging AI uptake and showcasing practical uses.

The new updates suggest that Singapore's AI standards regime is growing in breadth, but not necessarily safety focus.

International Approach

Key takeaways

- Singapore's multilateral AI governance strategy is becoming more technical and testing-driven. It remains active at the United Nations (UN), World Economic Forum (WEF), International Organization for Standardization (ISO), the global AI summit series, and the International Network for Advanced AI Measurement, Evaluation and Science (Network). Across these forums, it is increasingly contributing practical tools, testing methodologies and new ISO/IEC standards.
- Singapore is positioning itself as a convenor for scientific and technical AI safety work. By hosting the SC 42 plenary meeting on AI standard, announcing the International Scientific Exchange on AI Safety 2026 (ISE), and participating in joint testing through the Network, Singapore is using international platforms to gather experts to discuss the technical aspects of AI safety. Such convenings have become a core part of Singapore's international strategy, especially where AI governance needs to be translated into testing standards and safety benchmarks.
- Following the release of the ASEAN Guides on AI Governance and Ethics, the regional agenda is shifting toward testing and evaluation, with Singapore leading these efforts through the ASEAN Working Group on AI Governance (WG-AI). As the only ASEAN member in the Network, it is well placed to translate global developments in AI testing and evaluation into regional tools and benchmarks.
- The Singapore–Korea AI Safety Institutes (AISI) Memorandum of Understanding (MOU) in January 2026 stands out as the most substantive new bilateral development, marking Singapore's first publicly described AISI-to-AISI partnership and linking cooperation directly to technical outputs such as joint testing on data-leakage risks in AI agents.

This section outlines major developments in Singapore's international approach to AI governance since the previous report's cutoff in July 2025. It examines multilateral initiatives, such as Singapore's participation in UN processes, the global AI summit series, the Network, engagement at WEF, and the International Scientific Exchange on AI Safety; regional initiatives, especially through ASEAN; and bilateral agreements and partnerships on AI governance.

2.1 Multilateral Initiatives

2.1.1 United Nations Global Dialogue on AI Governance

Singapore has been involved in UN AI governance across both formal intergovernmental processes and expert-led policy initiatives, including supporting UN General Assembly resolutions on AI governance in 2024 and participating in the UN Secretary-General's High-level Advisory Body on Artificial Intelligence (HLAB-AI).[a] Singapore also hosted HLAB-AI's final meeting in May 2024. HLAB-AI's final report, published in September 2024, called for the creation of an independent scientific panel on AI and a recurring policy dialogue to support progress towards a multilateral AI governance framework.[32] These recommendations led to the launch of the **Global Dialogue on AI Governance**.[33]

In September 2025, Singapore participated in the High-level Multi-stakeholder Informal Meeting in New York, which formally launched the Global Dialogue as a follow-up to the Pact for the Future and the Global Digital Compact. At the meeting, Singapore announced the Singapore Digital Gateway (SGDG), which it presented as a resource for policymakers in other countries and multilateral organizations. The SGDG consolidates more than 30 Singapore resources on digital and AI governance, including strategies, frameworks, playbooks, and implementation tools.[34] These include the MGF publications, Project Moonshot, and the AI Verify Testing Framework, as well as regional resources such as the *ASEAN Guide on AI Governance and Ethics* (2024) and the *Expanded ASEAN Guide on AI Governance and Ethics – Generative AI* (2025). In doing so, Singapore positioned its domestic and regional governance tools as resources that could support broader international capacity-building and policy exchange.

2.1.2 International Network for Advanced AI Measurement, Evaluation and Science

The **International Network for Advanced AI Measurement, Evaluation and Science** (Network), previously known as the International Network of AI Safety Institutes, is a multilateral forum for technical cooperation on AI testing and evaluation, bringing together national AISIs and other equivalent government-linked bodies working on the science of AI measurement. The Network was first launched in November 2024, with Singapore participating from the outset. Singapore's involvement can be seen across a sequence of technical exercises and multilateral convenings, including joint testing work in 2025 and subsequent meetings in December 2025 and February 2026.

In July 2025, the Network released the results of its third joint testing exercise, "Advancing Methodologies for Agentic Evaluations Across Domains," building on insights from two earlier joint testing exercises conducted in November 2024 and February 2025. The exercise brought together representatives from Singapore, Japan, Australia, Canada, the European Commission, France, Kenya, South Korea, and the United Kingdom. It was divided into two strands: one on common risks involving leakage of sensitive information and fraud, led by the Singapore AISI, and another on cybersecurity, led by the UK AISI.[35] Its purpose was to examine

a. These included three UN General Assembly resolutions in 2024, respectively focused on safe, secure and trustworthy AI systems for sustainable development; capacity-building and technical assistance for developing countries; and the applicability of international law across the AI life cycle.

methodological challenges in agentic testing and to develop emerging best practices for future evaluations. Among the key lessons were the need for more realistic tasks and tools, the additional scaffolding required for agentic setups, and the importance of examining agent trajectories rather than only final task outcomes.[36]

The Network reconvened in San Diego on December 4–5 2025, alongside the NeurIPS conference. Singapore participated in the San Diego meeting, where discussions pointed to growing convergence around core evaluation principles, including transparency and reproducibility, stronger validity, embedded quality assurance, and a greater focus on multilingual and multicultural performance.[37]

The Network convened again at the India AI Impact Summit in February 2026. Although no formal public readout of the meeting was issued, the UK AISI indicated that the goal of the meeting was to compare lessons learned, test approaches against real-world use cases, and identify priorities for the next phase of AI measurement and evaluation.[38]

The Summit also included a public panel session on best practices in AI measurement and evaluation, with Head of Policy Wan Sie Lee representing the Singapore AISI.[39] In the panel, Lee noted that the Network had separately conducted a multilingual testing exercise on agentic systems, and highlighted that such multilingual evaluations are more complex than evaluating LLMs alone because they must account not only for dataset translation, but also the additional scaffolding involved in agentic systems, including differences in the languages used for tool-calling.[40]

2.1.3 World Economic Forum

In January 2026, Singapore launched its *Model AI Governance Framework for Agentic AI* (*MGF-Agentic AI*) at the **World Economic Forum's annual meeting in Davos**, using the occasion to introduce its latest governance approach for increasingly autonomous AI systems to an international audience.[41] This was part of a pattern: Singapore launched the first *Model AI Governance Framework* (MGF) at Davos in January 2019, returned in January 2020 to release the framework's second edition together with implementation guidance, and in January 2024 announced the proposed *MGF for Generative AI* for international consultation.[42][43] Across these iterations, Singapore has used Davos as a platform to internationalize successive versions of its AI governance approach as new technological developments emerged.

Across these frameworks, Singapore's approach has consistently been presented as voluntary and practical, with an emphasis on enabling organizational adoption rather than imposing binding rules. The MGFs are directed above all at companies and other deployers seeking operational guidance on how to govern AI in practice. This orientation is also reflected in how the frameworks have been developed and circulated. The 2020 implementation guidance was produced with input from more than 60 organizations, while the 2026 *MGF-Agentic AI* includes an annex explicitly inviting feedback during the public consultation period, and was accompanied by supportive statements from companies such as Amazon Web Services and Google Cloud.[44][45]

As a high-profile cross-sector gathering, Davos gives Singapore's frameworks visibility among the companies most likely to implement them, while also potentially signaling to a global business audience that Singapore remains a serious and welcoming destination for AI investment and development.

2.1.4 India AI Impact Summit

In February 2026, Singapore participated in **India AI Impact Summit**. The summit can be understood as part of the AI summit series launched at Bletchley Park in October 2023 and continued in Seoul in May 2024 and Paris in February 2025. Over time, this series has broadened from a relatively concentrated focus on frontier AI safety at Bletchley towards a wider agenda spanning trusted AI, inclusion, public-interest applications, science, and international cooperation. Singapore has remained engaged throughout: Minister for Digital Development and Information Josephine Teo participated in each summit, and Singapore signed the declarations or outcome statements issued at Bletchley Park, Seoul, and Paris.

At the India AI Impact Summit, Minister Teo's interventions reflected Singapore's current emphasis on science-based governance, practical assurance, and the risks posed by increasingly autonomous systems. In her keynote on monitoring the impacts of agents, she argued that agentic AI requires governments and industry to move beyond static model evaluation toward approaches that can monitor real-world impacts.[46] She also framed this as an institutional challenge, noting that countries vary in their ability to evaluate, assure, and govern advanced AI, which risks widening a "global assurance divide".[47][48]

In the panel on AI safety at the global level, Teo similarly emphasized implementation, arguing that safety science must be translated into usable guardrails, standards, and tools. She also characterized AI as a threat, a target, and a tool: it can be used to attack systems, AI systems themselves may be vulnerable to cyberattacks, and AI can also be used to defend against such threats.[49] This framing was linked especially to agentic and multi-agent systems, where cybersecurity and safety concerns increasingly intersect.

The Summit also served as a platform for Singapore to announce new initiatives. Singapore announced, first, the establishment of a Network of AI-for-Science Institutions with India and Canada, following its role in co-chairing the Summit's Science Working Group.[50] The initiative was framed as a virtual network to support knowledge-sharing, ecosystem-building, and the systematic use of AI to accelerate experimentation, modeling, and discovery across applied scientific domains—though what this would mean in practice remains to be seen. Second, Singapore announced that it would host the second edition of the International Scientific Exchange (ISE) on AI Safety later in 2026, with a particular focus on developments such as agentic AI and the safety and security challenges arising from it.

The Summit concluded with the adoption of the New Delhi Declaration on AI Impact, which was endorsed by 92 countries and international organizations, with Singapore listed among the signatories.[51] The Declaration identified "Secure and Trusted AI" as one of its seven pillars, recognizing the importance of securing AI systems and adopting technical solutions and appropriate policy frameworks that enable innovation while promoting the public interest.

2.1.5 SC 42 Plenary Meeting

In April 2026, Singapore hosted the **SC 42 biannual plenary meeting**.[52] SC 42 is the subcommittee for AI within ISO-IEC Joint Technical Committee 1 (JTC), covering areas such as foundational standards, trustworthiness, use cases, data, and testing. The plenary was the first SC 42 plenary hosted in the ASEAN region and served as the venue for discussion of ISO/IEC 42119-8, a Singapore-proposed standard for testing generative AI systems.[53]

The proposed standard focuses on benchmarking and red-teaming methodologies, aiming to improve the reproducibility and comparability of generative AI testing results.[54] According to the press release, ISO/IEC 42119-8 draws on the IMDA's AI Verify Toolkit, the *Starter Kit for Testing of LLM-Based Applications*, and the Global AI Assurance Sandbox. It also forms part of Singapore's broader standards strategy, alongside Enterprise Singapore's (ESG) ISO/IEC 42001 accreditation program and Singapore's contributions of real-world cases to support ISO/IEC TR 24030's documentation of AI applications in practice.

On the sidelines of the plenary, IMDA and ESG hosted several capacity-building initiatives: a Broad-Based Foundational Training Workshop with the American National Standards Institute (ANSI) to strengthen AI standards capability among ASEAN member states; an ISO-organized AI Standards and Policy Workshop for national standards bodies and AI policymakers from 15 countries; and an AI Assurance public sharing session.

2.1.6 International Scientific Exchange on AI Safety 2026

In February 2025, Singapore announced that it would host the **International Scientific Exchange on AI Safety (ISE)** in 2026. ISE 2026 builds on an existing track record of convening international expert discussions on AI governance and safety. The inaugural Singapore Conference on AI for the Global Good (SCAI) in December 2023 brought together international experts to develop the SCAI Questions, reflecting a broad range of concerns with the global implications of advanced AI. That foundation was carried forward in SCAI: International Scientific Exchange on AI Safety in April 2025, which retained the SCAI branding while placing a clearer emphasis on AI safety. The 2025 meeting convened more than 100 participants from 11 countries and produced the ***Singapore Consensus on Global AI Safety Research Priorities***, framed as a living document intended to facilitate global conversations, improve understanding of risk management, and spur international research collaboration around technical AI safety. The forthcoming ISE 2026 continues this trajectory, while also making the shift in emphasis more explicit through a title focused entirely on AI safety.

As announced at the India AI Impact Summit, ISE 2026 will build on this momentum by revisiting safety research priorities and exploring new developments in the AI landscape, including the safety challenges posed by increasingly autonomous systems such as agentic AI.[55] The focus on agentic AI is notable given how rapidly the topic has risen up the international safety agenda, and reflects a broader convergence between Singapore's domestic governance priorities (see *Domestic Approach* section) and the concerns driving multilateral safety dialogue.

Across these forums, a consistent pattern is visible in Singapore's international approach to AI governance: it participates as a member, contributes its domestic tools and frameworks for wider adoption, and, where possible, takes on a convening or co-chairing role. This pattern has continued to shape Singapore's international engagements since our 2025 report, including its convenings and participation in the second half of 2025 and in 2026.

2.2 Regional Initiatives

2.2.1 Association of Southeast Asian Nations (ASEAN)

At the regional level, Singapore has continued to play a leading role through the **ASEAN Working Group on AI Governance (WG-AI)**, including by hosting the 5th WG-AI meeting in Singapore in September 2025**.** Established by Singapore in March 2024, the WG-AI was designed to coordinate ASEAN's AI governance work and to serve as the focal point for the region's AI cooperation with Dialogue Partners including the US, China, Japan, Korea, and India. Its initial focus in 2024 and early 2025 was on developing and operationalizing the *ASEAN Guide on AI Governance and Ethics*, and subsequently expanding this work to address generative AI through the *Expanded ASEAN Guide on AI Governance and Ethics*. The *Expanded Guide* drew on Singapore's *MGF for Generative AI*, while adapting its recommendations to ASEAN's regional context.[56]

In January 2026, this regional safety-testing agenda was reflected in IMDA's **Singapore AI Safety Red Teaming Challenge 2026** (see *Domestic Approach* section). While Singapore led the initiative, the Challenge involved the wider region: participation extended to all ASEAN member states, alongside China, India, Japan, and South Korea, and observers included representatives from the ASEAN WG-AI and the Network. This suggests that Singapore is using practical exercises such as red-teaming challenges not only to test AI applications, but also to surface common vulnerabilities and build shared technical knowledge around multilingual and application-layer AI safety risks across the region.

Later that month, the 6th ASEAN Digital Ministers' Meeting in Hanoi acknowledged the broader progress of the WG-AI. This included its coordination role on AI matters, engagement with Dialogue and Development Partners and industry, and its work building on the 2024 and 2025 ASEAN Guides. The meeting did not specifically reference the WG-AI's proposed work on testing tools or regional safety benchmarks, but affirmed its continuing role in advancing ASEAN's wider AI governance agenda.

In February 2026, Minister Josephine Teo reiterated that Singapore was "working within ASEAN to explore practical tools for AI safety testing" and aimed to "collectively develop a set of AI safety benchmarks that reflect our region's concerns."[57] This statement gave a clearer policy direction to the more practical testing agenda already visible through the Challenge and the WG-AI's ongoing work.

The WG-AI initiatives and the Safety Challenge suggest that Singapore is steering ASEAN towards more practical mechanisms for AI safety testing, benchmarking, and implementation that are aligned with international developments while also responding to the specific needs, deployment contexts, and concerns of Southeast Asia. Singapore is currently the only ASEAN member in the Network, and is therefore well placed to trans-

late emerging international lessons on measurement and evaluation into regional governance efforts. This is especially relevant for multilingual testing. The Network's work has already included evaluations in English, Mandarin Chinese, and major Indian languages such as Hindi and Telugu. It has also highlighted methodological challenges relating to language variation, tool translation, and agent behavior across different linguistic contexts, and these lessons are likely to be relevant for ASEAN's own multilingual and lower-resource language settings.

2.3 Bilateral Efforts

2.3.1 Interoperability of Governance and Testing Frameworks

Singapore has sought to improve interoperability between its AI governance tools and external frameworks through a series of mapping exercises or "crosswalks." Earlier work included mappings between the *AI Verify Testing Framework* and the U.S. National Institute of Standards and Technology (NIST) *AI Risk Management Framework*, ISO/IEC 42001, the G7 *Hiroshima AI Process International Code of Conduct*, and the NIST *AI Risk Management Framework Generative AI Profile*. These initiatives were aimed at showing points of compatibility across frameworks and reducing compliance friction for firms operating across multiple governance environments.[58]

No major new crosswalks or bilateral interoperability initiatives have been publicly announced since the 2025 report. Further interoperability work may yet emerge for newer publications such as the *Starter Kit for Testing LLM-Based Applications*, but no such development has been publicly announced.

2.3.2 Bilateral Cooperation in AI Governance

Singapore's bilateral approach to AI governance has combined formal cooperation agreements, policy dialogues, and digital economy arrangements to promote AI governance frameworks, information-sharing, safety research, and testing collaboration. Earlier examples included the November 2024 Singapore–United Kingdom (UK) Memorandum of Cooperation on AI safety, the European Union (EU) –Singapore Administrative Arrangement on collaboration on the safety of AI, and bilateral dialogue with China through the Singapore–China Digital Policy Dialogue.[59][60] More broadly, Singapore has also embedded AI governance provisions in a wider set of digital economy and digital partnership agreements with partners such as Australia, the United Kingdom, Korea, the European Union, New Zealand, and Chile.

The clearest new development since the 2025 report is Singapore's cooperation with the Republic of Korea through their respective AISIs. During Prime Minister Lawrence Wong's visit to Seoul in November 2025, both sides stated that they would implement a **Memorandum of Understanding (MOU) between the Korea AI Safety Institute and IMDA** and explore further cooperation on AI safety and governance through their national AISIs.[61] In January 2026, the Singapore AISI and Korea AISI then announced that they had signed the MOU and released joint testing work on data-leakage risks in AI agents.[62][63] This appears to

be the first public agreement directly between Singapore's AISI and a foreign AISI counterpart, and it was accompanied by concrete technical collaboration rather than being a policy commitment alone.

The Singapore–Korea bilateral testing exercise evaluated whether AI agents can execute realistic multi-step tasks without leaking sensitive data. It focused on common enterprise and productivity scenarios, and was motivated by the concern that agents may mishandle confidential information even in non-malicious settings, for example because they misunderstand contextual privacy norms, hallucinate, or fail to distinguish between appropriate and inappropriate disclosure. The findings showed that data leakage remained an issue even in benign tasks, even though the agents were given task-specific data handling guidelines and explicit instructions not to leak data.

Singapore and the EU continued to work on implementing their existing Administrative Arrangement. At the **Second EU–Singapore Digital Partnership Council** in December 2025, both sides reaffirmed their intention to deepen cooperation on safe, trustworthy, and human-centric AI, including through sharing on the evaluation of language models between the EU's Alliance for Language Technologies European Digital Infrastructure Consortium (ALT-EDIC)[b] initiative and AI Singapore (SEA-LION family).[65]

Singapore's engagement with China likewise appears to have continued along broader digital-policy lines. At the second **Singapore–China Digital Policy Dialogue** in September 2025, discussion centered on National AI Strategy 2.0 implementation, talent development, AI Centres of Excellence, trusted data flows, and industrial applications. The AI Governance Working Group announced at the first Digital Policy Dialogue in June 2024 was not mentioned in the official readout for the 2025 Digital Policy Dialogue, though this does not necessarily mean that working-level cooperation ceased.

The Singapore–Korea MOU stood out as the most substantive new bilateral development in this period, notable both as Singapore's first public AISI-to-AISI agreement and because it was accompanied by concrete technical output on data-leakage risks in AI agents. More broadly, technical cooperation and knowledge-sharing on AI safety appears to increasingly take place in multilateral and regional settings, including the ASEAN WG-AI and the Network.

b. ALT-EDIC (the Alliance for Language Technologies) is an EU consortium, established in 2024 to build shared European data infrastructure and services for language technologies. Its core role is to federate multilingual and multimodal data across participating states to support the development of more capable European language models.[64]

Homegrown AI Ecosystem and Industry

Key takeaways

- Both homegrown SEA-LION and MERaLiON model families released new updates, with SEA-LION v4 expanding into multimodal image-text capabilities and lighter edge-deployable models, while MERaLiON-3-preview broadened Singapore's speech-AI capabilities. These updates remain capability-based, leaving safety fine-tuning to downstream developers.
- Singapore's homegrown general-purpose AI (GPAI) ecosystem now has a broader set of tools across the safety stack, with SEA-Guard, LionGuard 2, and AI Guardian. The focus of these tools remains on preventing toxicity and on multilingual moderation, rather than addressing the broader spectrum of AI safety risks.
- The third-party assurance ecosystem broadened, with newer entrants joining initiatives such as the Global AI Assurance Sandbox. While the ecosystem remains concentrated at the model and application layers, the inclusion of compliance- and governance-oriented providers suggests that AI assurance may gradually broaden from technical testing toward regulatory and organizational assurance.
- Foreign AI developers have become more visibly engaged in Singapore's AI ecosystem, particularly through controlled experimentation in public-sector settings, multilingual and regional adaptation, and domain-specific safety benchmarking. However, most of these collaborations remain tied to specific government or internal use cases, rather than supporting public or ecosystem-wide adoption.

This section explores Singapore's AI ecosystem and industry across the following areas: the homegrown GPAI ecosystem that includes locally developed GPAI models focused on Southeast Asian languages and accompanying safety tools; the AI assurance sector in Singapore; and foreign GPAI companies based in Singapore.

3.1 Homegrown GPAI ecosystem

3.1.1 SEA-LION and MERaLiON Homegrown Model Families

Singapore's National Multimodal LLM Programme, driven by AI Singapore, the Agency for Science, Technology and Research (A*STAR), and IMDA, has developed two open source model families, SEA-LION and MERaLiON, that focus primarily on regional languages rather than frontier capabilities. Singapore does not have other homegrown public GPAI models.[a]

The SEA-LION (Southeast Asian Languages in One Network) family was first launched in late 2023 by AI Singapore as part of Singapore's effort to develop open models better suited to Southeast Asia's languages, cultures, and contexts. Its value proposition was to provide Singapore with a homegrown open model that was better able to account for underrepresented Southeast Asian languages; over time, this regional language focus expanded outward through adjacent country-specific models, such as Indonesia's Sahabat-AI, co-developed with AI Singapore.[67]

A major update is the release of **SEA-LION v4** in August 2025. SEA-LION v4 is the project's first collection of multimodal models, employing Apertus, Gemma, and Qwen-based models,[b] with different parameters, context length and multimodal capabilities to suit different users. In contrast to the earlier text-based SEA-LION models, v4's capabilities include image-and-text inputs while retaining a focus on Southeast Asian languages and cultural context (Figure 3.1). In addition, lightweight models (4 billion parameters) were developed to cater for mobile and edge devices.

Building on the SEA-LION family, MERaLiON—Multimodal Empathetic Reasoning and Learning in One Network—extends Singapore's national model development efforts, with capabilities in speech, audio, and multimodal understanding. The first release, MERaLiON-AudioLLM, was introduced in December 2024 as an audio-text model built on SEA-LION v3's 10B-parameter architecture. It was designed to process spoken regional language forms, including Singlish, alongside English and Mandarin, thereby addressing speech and code-switching patterns common in Singapore and Southeast Asia. In May 2025, A*STAR released MERaLiON-2, which expanded the model family's language coverage to include Malay, Tamil, Indonesian, Thai, and Vietnamese, while adding stronger capabilities in multimodal speech understanding, code-switching, and paralinguistic tasks such as emotion recognition.

In March 2026, this was followed by **MERaLiON-3-10B-preview**, a preview version of the next-generation MERaLiON-3 speech-text model.[68] The "preview" designation reflects that the release is still research- and evaluation-oriented rather than a fully finalized production model; the model card provides a web demonstration and notes that MERaLiON-3 will be released in the future. The Southeast Asian benchmark used for evaluation will also be released separately as part of a paper. MERaLiON-3-preview expands the family's emphasis from multilingual speech recognition and transcription toward broader speech and audio under-

a. There are LLMs developed for internal government use, such as Phoenix by Home Team Science and Technology Agency, which will not be discussed in this report.[66]

b. SEA-LION v3.5 was only built on Llama.

Name	Version	Parameters	Quantised Versions	Model Type	Context length
SEA-LION (Apertus) SEA-LION's flagship model built on Swiss AI's open-source models	v4 (IT)	8B	N/A	Instruct	65K
SEA-LION (Gemma) SEA-LION's flagship model built on Google's open-source models	v4 (IT) v4 (VL) v4 (VL)	27B 27B 4B	8-bit, 4-bit, mlx-4bit N/A N/A	Instruct, Vision Instruct, Vision Instruct, Vision	128K
SEA-LION (Qwen) SEA-LION's flagship model built on Qwen's open-source models	v4 (IT) v4 (VL) v4 (VL)	32B 8B 4B	8-bit, 4-bit N/A N/A	Instruct Instruct, Vision Instruct, Vision	32K 256K 256K
SEA-LION (Llama) Vision-language model for image + text comprehension	v3.5 (R) v3.5 (R)	70B 8B	16-bit, 8-bit, 4-bit, mlx-4bit 16-bit	Reasoning Reasoning	128k 128k

Figure 3.1: SEA-LION v4 models

standing capabilities, including recognizing a speaker's age and gender, answering spoken questions, contextual paralinguistic question answering[c], audio captioning, and summarizing spoken dialogue. Compared with MERaLiON-2, the 3-preview shows stronger performance on several speech and paralinguistic benchmarks, with higher scores in the above capabilities.

However, similar to SEA-LION v4, the model release card for MERaLiON-3 states that it is not specifically safety-aligned; both model families place responsibility on developers and deployers to conduct safety fine-tuning, validation, and security measures before deployment.

c. This refers to the evaluation of a system's ability to understand both verbal and non-verbal cues.

3.1.2 Safety Tools and Evaluations

Several safety-related tools and benchmarks have emerged since mid-2025. AI Singapore's **SEA-Guard** is a safety guard model, added to the SEA-LION family to screen potentially harmful content. AI Singapore provided an early prototype in May 2025, before formally launching it in October that year. SEA-Guard provides a simple "safe/unsafe" classification for user prompts and model responses.[69] Trained on Southeast Asia-specific cultural knowledge, it is intended to detect, moderate, and handle content according to regional cultural norms and safety standards. This gives it a better chance of identifying Southeast Asia-specific slurs, taboos, and sensitive topics that more generic safety models may overlook. AI Singapore reports that SEA-Guard performs more strongly and consistently than open source state-of-the-art baselines in both English and Southeast Asian languages.[70] At the same time, its current scope remains relatively limited[d]: SEA-Guard evaluates only a single user prompt or response at a time and does not yet support system prompts or multi-turn conversations, which means it remains focused on single-turn harmful-content filtering.

Whereas SEA-Guard is positioned as a regional safety layer for the SEA-LION ecosystem, GovTech's LionGuard series is designed more specifically for Singapore government use cases. In late July 2025, GovTech launched **LionGuard 2**, an upgraded open source guardrail building on the original LionGuard, released in 2024.[e] LionGuard 2 is a multilingual content-moderation classifier tailored to the Singapore context with relatively modest compute and training requirements. Already deployed within the Singapore Government, it extends the earlier model's English and Singlish focus by adding support for Chinese and Malay, while also introducing a more refined taxonomy of harms covering insults, sexual content, violence, self-harm, and misconduct.[72]

We also see evidence of other domestic agencies creating safety benchmarks for their LLMs, though these appear to be designed primarily for internal use. In November 2025, the Home Team Science and Technology Agency of Singapore (HTX) announced an expansion of its partnership with Mistral AI. The announcement noted that HTX and Mistral AI had already been co-developing HT-Lexicon, an AI safety benchmark tailored for Singapore Home Team's internal use.[f73] This points to a broader pattern in which domestic agencies are not only adopting external safety tools, but also developing specialized internal benchmarks aligned with their own operational requirements.

These safety-oriented models and guardrails are complemented by region-specific evaluation tools, which help assess whether such systems can reliably identify harmful or sensitive content across Southeast Asian languages and cultural settings. One such tool is **SEA-HELM (Southeast Asian Holistic Evaluation of Language Models)**, a benchmarking suite designed to assess large language models on tasks such as proficiency in Southeast Asian chat, instruction-following in Southeast Asian languages, and a suite of English tasks. Results from the benchmark are published on a public leaderboard.[74]

d. This limitation was also observed in the early release, as covered in the 2025 report.

e. This release was after the cut-off date (July 2025) of the 2025 State of AI Safety Singapore report.[71]

f. Home Team refers to the departments and statutory boards under the Singapore Ministry of Home Affairs (MHA) responsible for Singapore's domestic security. This includes the police, prison and civil defense forces, etc.

A more recent development is **SEA-SafeguardBench**, a multilingual safety benchmark for large language models, released by AI Singapore in December 2025. The benchmark evaluates responses to harmful-content prompts across eight Southeast Asian languages and contains 21,640 human-verified prompts. Its results showed that state-of-the-art models and safeguards generally perform less well on Southeast Asian languages and culturally specific harms than on English. The paper also called for developers to create more region-specific safety tooling, rather than relying on English-centric benchmarks alone.[75]

Despite these developments, the current evaluation landscape remains limited in scope. Although safety is listed as one of the five pillars of SEA-HELM, their leaderboard is more useful for demonstrating the multilingual capabilities of Singapore's homegrown models in comparison with other GPAI systems than for providing a comprehensive view of safety performance. Its safety component remains relatively narrow: official documentation indicates that it focuses mainly on toxicity detection and covers only Indonesian, Thai, Vietnamese, and Filipino. SEA-SafeguardBench expands coverage to eight languages, but it still concentrates primarily on harmful-content risks such as toxicity and misinformation, rather than the broader range of concerns relevant to AI safety. There is also no evidence that SEA-SafeguardBench is integrated into the SEA-HELM leaderboard; as of now, it functions as a standalone evaluation. This limitation has already been acknowledged in the SEA-HELM literature, which notes the need to broaden language coverage, task scope, and benchmark design over time.[7677]

The developments above are complemented by public-sector tools that aim to operationalize safety testing and runtime safeguards in real-world deployment. Beyond localized guardrails such as LionGuard 2, GovTech has continued to build a broader public-sector AI safety stack. In May 2025, it launched **AI Guardian**, which comprises **Litmus** for pre-deployment testing and **Sentinel** for runtime guardrails. Litmus supports the testing stage before deployment by automating checks to assess whether AI applications meet required safety and reliability standards. Sentinel, by contrast, is used during live operation to protect deployed systems against risks such as prompt injection, toxic content, and personal-data exposure.[78] Since its launch, around one-third of Singapore government agencies have reportedly used Litmus and Sentinel in their day-to-day workflows.[g]

The *Starter Kit* published in January 2026 further illustrates how this tooling is being operationalized. It uses Litmus as a case study and notes that the tool includes 1,600 prompts across four categories—Security, Specialised Advice, Undesirable Content, and Political Content—with a default threshold of requiring 95% safe responses for government chatbots.[80]

These developments suggest that Singapore's homegrown GPAI ecosystem is expanding its focus to include more safety layers across the stack. In the 2025 report, we noted that Singapore's homegrown GPAI efforts were centered primarily on improving multilingual and multimodal performance.[81] The emergence of SEA-Guard, LionGuard 2, and SEA-SafeguardBench, as well as the public-sector deployment of AI Guardian, indicate that safety is no longer being treated only as a downstream responsibility, but is increasingly being built into models, guardrails, benchmarks, and deployment practices. At the same time, Singapore's home-

g. The proportion was as of November 2025, since then the number could have increased.[79]

grown models and evaluation tools remain primarily oriented towards performing well in regional languages and cultural contexts, rather than matching the capabilities of frontier models. These models therefore may pose fewer of the extreme risks commonly associated with the more advanced GPAI systems. Even so, there remains room for further development; in particular, safety evaluations could be broadened beyond toxicity and harmful content to capture a wider range of AI risks.

3.2 Third-party AI Assurance Suppliers

Third-party assurance providers continue to play an important role in Singapore's broader AI safety ecosystem. Singapore has a wide range of third-party AI assurance providers offering products and services to AI developers and downstream users. Their customers include organizations that may choose not to build full in-house assurance capabilities because AI governance is not core to their business, as well as those that require specialist advice in niche areas, or independent external validation of internal assessments. The types of products and services that these providers offer vary in breadth and focus: some offer technical testing tools or AI governance platforms, while others provide end-to-end assurance advisory services to support a firm's AI governance and risk management practices. Beyond supplying commercial products and services to public- and private-sector clients, these firms also participate in Singapore's growing AI testing ecosystem through co-developing testing tools with the AI Verify Foundation (AIVF) such as Project Moonshot and participating in its initiatives.

We have not observed a major expansion in the number of third-party AI assurance providers in Singapore since the 2025 report. The Global AI Assurance Sandbox continued to bring in assurance providers to test existing risks from the Global AI Assurance Pilot and new risks such as data leakage and prompt-injection vulnerabilities (see *Domestic Approach* section).[82] Importantly, the Sandbox also aims to have its findings eventually feed into future technical testing standards.In addition, more than 10 assurance providers also contributed feedback and real-world testing case studies to the Starter Kit published in January 2026. Besides expanding participation and the range of risk categories tested, the Sandbox also gives AIVF and assurance providers practical testing experience to inform future guidance, sector-specific expectations, and more standardized practices.

Given the large number of third-party AI assurance providers in the ecosystem, we do not aim to provide an exhaustive list. Instead, we provide a snapshot of providers active in Singapore's assurance ecosystem in 2025–2026, drawn from lists of participants in AIVF initiatives or public records of collaboration with Singaporean companies or government agencies. Table 3.1 divides them using three assurance levels—model, system/application, and organization.[hi]

h. 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, across different organizational functions such as operations, documentation, and reporting disclosures. 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.

i. This list is non-exhaustive due to limitations of public research, companies may have changed services, and new entrants may have entered the market since the publication of this report.

Table 3.1: Third-party AI assurance providers

Third-party AI Assurance Providers		
Assurance level	**Typical assurance mechanisms**	**Examples of providers**
Model	Red-teaming, algorithmic bias audits, model testing and evaluation, formal verification	Citadel AI, Collinear, Guardrails AI, AIDX Tech, Knovel, Advai, Resaro.AI, Vulcan, Reversec, PRISM Eval, Parasoft, Verify AI
System / Application	Risk assessment, impact assessment	AIDX Tech, Asenion, Reversec, AIQURIS, Fairly, LatticeFlow AI, QuantPi, Verify AI
Organization	Governance training, advice on organizational policies, compliance audits	BSI, PWC, SoftServe

It is natural to see the strongest concentration of providers at the model and application layers in the table above, given that many of the listed firms are drawn from initiatives such as the Global AI Assurance Pilot and Sandbox, both of which focus on testing deployed AI applications rather than organization-level governance. A notable development is that newer entrants such as Asenion[j] and BSI also provide assurance services that extend beyond technical testing into compliance, governance, and certification-oriented work. Although there is currently no public information indicating whether these firms are providing such services within the Sandbox itself, the Sandbox's expanded scope now includes findings such as breaches of industry-specific regulatory requirements and internal compliance requirements.[83] This may suggest that the government is beginning to consider how AI assurance exercises could also surface compliance and regulatory issues, in addition to technical safety and security risks. Over time, this could help support more formalized testing standards and assurance practices.

3.3 Foreign AI Developers

Many large technology companies and frontier AI labs from the US, China, and other countries are active in Singapore through a combination of local operations, public-sector partnerships, research activity, and the deployment of their GPAI systems. These companies include Google (Gemini), Meta (Llama), Alibaba (Qwen), ByteDance (Seed), Tencent (Hunyuan), Baidu (ERNIE), OpenAI (GPT), Anthropic (Claude), and Mistral AI. This section focuses on concrete, publicly documented initiatives through which foreign AI developers contribute to Singapore's AI ecosystem, particularly in relation to local capacity-building and model development, multilingual adaptation, safety, and assurance. It does not aim to cover all foreign AI firms active in Singapore. In some cases, for firms such as ByteDance, Anthropic, and OpenAI, there are indications of safety-related activity through hiring, research presence, or public announcements of partnerships, but public information remains too limited to support a fuller assessment.

j. Asenion was launched in June 2025 through Fairly AI acquisition of anch.AI, Fairly AI was previously covered in our 2025 report.

First, several foreign developers are supporting controlled experimentation with advanced models in public-sector and security-sensitive environments. Google, for example, announced in August 2025 that it is working with Singapore government agencies on an AI agents sandbox, and it enabled access to Gemini through Google Distributed Cloud in air-gapped settings.[84] Other examples are similarly bounded and use-case-specific. Mistral is co-developing HTX's Phoenix models for Home Team use cases;[85] it has also worked with Singapore's Ministry of Defence, the Defence Science and Technology Agency, and DSO National Laboratories to develop generative AI models for the Singapore Armed Forces' mission planning.[86] Microsoft is also working with HTX to fine-tune Phi-4-multimodal for Home Team applications.[87] These examples suggest that foreign developers are being integrated into Singapore's ecosystem primarily through use-case-specific deployment and adaptation.

Second, the clearest area of substantive technical alignment appears to be multilingual and regional adaptation. This is where foreign developers' technical capabilities most visibly intersect with a distinctive Singapore priority: improving model performance on Southeast Asian languages and multilingual usage patterns that are often underrepresented in mainstream GPAI systems. In November 2025, Google DeepMind opened new research lab in Singapore focused on understanding regional languages.[8889] Google had also expanded its collaboration with AI Singapore to support SEA-LION v4, including one of the project's first multimodal models built on Gemma 3.

Alibaba Cloud was also involved in the development of SEA-LION v4, which was built on Qwen3.[90] Alibaba Cloud and AI Singapore presented this as a collaboration that pairs Qwen's multilingual and reasoning capabilities with Singapore's regional expertise. Qwen3 was pre-trained on a large multilingual corpus covering 119 languages and dialects and around 36 trillion tokens, giving it wider linguistic coverage than many other GPAI models, including Southeast Asian languages often underrepresented in mainstream models. Its post-training also placed emphasis on multilingual usage patterns especially relevant to Southeast Asia, including code-switching, informal chat, and mixed use of English and local languages. These examples suggest that multilingual adaptation is one of the clearest areas where foreign partnerships are supporting capacity-building, locally relevant model capabilities, and related data infrastructure.

Third, some partnerships touch more directly on applied safety and assurance. One example is HTX and Mistral AI's development of HT-Lexicon, a safety benchmark tailored for the Home Team internal use.[91] Similarly, programs such as Google's public-sector sandboxing work aim to test AI models safely before adoption. However, these efforts remain focused on controlled deployment within government, rather than the public release of models, benchmarks, or safety tools for wider use.

Overall, these partnerships suggest that Singapore's AI ecosystem is developing not in isolation, but in active engagement with major foreign developers. At present, the clearest focus appears to be on adapting and testing models for controlled or internal uses, rather than releasing them publicly. Even so, these collaborations may still have wider significance if the lessons from internal deployment, sandboxing, and benchmarking feed into future guidance or broader adoption practices.

Technical Research

Key takeaways

- Singapore's technical AI safety research capacity grew 50% in the past year, from 14 to 21 profiled lead researchers, with new inclusions concentrated at NUS, NTU, and A*STAR. Complementing this growth in lead researchers is a wider research community behind the selected papers, including research fellows and graduate students.
- AI safety research is still concentrated in universities, but is now expanding into government institutions. We have profiled two researchers from A*STAR's Centre for Frontier AI Research and included GovTech's formal applied safety research. In contrast, industry-based technical AI safety research remains sparse.
- Singapore shows a comparative strength in safety evaluation and multilingual deployment research. Its work on evaluation benchmarks such as RabakBench, and on multilingual safety aligns with a broader focus on developing AI models for low-resource regional languages and assessing risks such as toxicity in these models.
- Research attention is beginning to shift toward agentic and system-level risks. Research papers on agentic AI safety topics have more than doubled since the 2025 report. The trend appears across multiple institutions, suggesting growing interest in the safety of LLM-based agents and downstream systems.
- Notable gaps remain. Researchers tend to focus on attack-and-defense, but there is comparatively less work on theoretical safety, interpretability, formal verification, and foundational alignment. An update to the *Singapore Consensus on Global AI Safety Research Priorities* in May 2026 would be timely and could help direct attention toward less-developed priorities.

Researchers engaged in AI safety research in Singapore are spread across leading universities and government research centers. Leading universities with 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). Government agencies with research output include the Agency for Science, Technology and Research (A*STAR), the Government Technology Agency (GovTech), and the Singapore AI Safety Institute (AISI).

This section reviews technical AI safety research relating to general-purpose AI (GPAI) systems.[a] Our scope is informed by the technical research categories identified in the *Singapore Consensus on Global AI Safety Research Priorities*, and the risk taxonomy from the *International AI Safety Report*.[92]

Methodology

We examine four commonly recognized areas of AI safety research:

- **Alignment** research seeks to ensure AI systems are controllable and to make them less hazardous by focusing on dangers such as power-seeking tendencies, dishonesty, or hazardous goals.[b]
- **Robustness** enables models to withstand hazards.[c]
- **Monitoring** research aims to reveal and prevent harmful behavior by making AI systems more transparent. This includes understanding models' internal representations, monitoring anomalies, and evaluating hazardous capabilities. Within monitoring, **interpretability** involves work on explainability, saliency maps, mechanistic interpretability, and representation engineering; **evaluations** include benchmarks to evaluate dangerous capabilities and propensity to cause harm.
- **Systemic safety** research seeks to understand and mitigate the broader societal risks associated with AI deployment, beyond the capabilities of individual models.[d]

In this report, we profile the above institutions and highlight active researchers within them who have produced at least five AI safety papers since January 2025, using this as a quantitative, output-based indicator of sustained engagement with the field and for consistency with the previous report.[e] We also include prominent academics in Singapore who hold leadership roles in major AI safety convenings, regardless of publication count.[f]

a. This refers to AI models and systems that can perform a variety of tasks, rather than being specialized for one specific function or domain.

b. This includes Reinforcement Learning from Human Feedback (RLHF) for question refusal, representation control and unlearning specific capabilities, value alignment, machine ethics, etc. It does not include RLHF for capabilities improvement.

c. Robustness research includes adversarial attacks, data poisoning, Trojan attacks, extraction of model weights and training data, etc.

d. This includes societal harms such as AI-generated deepfakes, misinformation, cybersecurity, and CBRN risks.

e. In the previous report, we surveyed papers published between January 2024 and July 2025. For this report, we count papers published between January 2025 and May 2026.

f. This includes technical research conferences like the Singapore Conference on AI, as well as AI academic conferences such as NeurIPS, ICLR and ICML.

This five-paper threshold has known limitations. It is purely output-based and does not account for publication venue quality, citation impact, or methodological contribution—a researcher with five "work-in-progress" workshop papers satisfies the criterion just as well as a researcher with five conference (e.g. NeurIPS) papers. Readers should treat publication counts as a minimum indicator of activity, not a measure of impact. Further, our profiles below focus on anchor or lead authors. Each of our selected papers was written by multiple co-authors or corresponding authors, including research fellows, graduate students, and other contributors. Their work is also part of the Singapore AI safety ecosystem, even where they are not individually profiled.

Our methodology excludes some important contributors: those with high-impact but lower-volume output, those working in adjacent areas not fully captured by our scope, those contributing primarily through industry or policy channels, and established scholars whose recent output falls below our threshold. The list aims to provide a representative snapshot rather than an exhaustive account of Singapore's technical AI safety ecosystem as of May 2026.

Finally, the report's institution-by-institution structure tends to under-emphasize cross-institutional collaboration. Co-authored papers appear under a single institutional profile even where the work reflects cross-institutional partnerships.

4.1 Universities

4.1.1 National University of Singapore (NUS)

AI research at NUS is conducted across multiple schools and faculties, with a major concentration in the Department of Computer Science within the School of Computing. The School of Computing also houses the NUS AI Institute (NAII), which was launched in March 2024 to consolidate AI research across the university, bringing together academics from more than 30 departments across 13 faculties and university-level institutes.

NAII organizes its work into three broad research areas: AI + X, AI Governance and Policy, and Foundational AI. Within Foundational AI, the Responsible and Safe AI domain explicitly addresses AI safety issues, including fairness, robustness, explainability, alignment with human values, trust, security, privacy-preserving technologies, and bias mitigation.

CHANG Ee-Chien: Associate Professor, School of Computing

CHANG Ee-Chien is an Associate Professor at the NUS School of Computing and Lead Principal Investigator of the National Cybersecurity R&D Laboratory. His research broadly spans information security, cloud security, applied cryptography, and data privacy. His AI safety-related work includes defenses against adversarial attacks and image watermarking.

Selected papers:

"SnapGuard: Lightweight Prompt Injection Detection for Screenshot-Based Web Agents"[93]	Apr 2026	Du, M., Fang, H., Ma, H., …, Yin, Q., Chang, E.C.
"ResGuard: Enhancing Robustness Against Known Original Attacks in Deep Watermarking"[94]	Apr 2026	Wang, H., Fang, H., Qiu, Y., Wang, S., Chang, E.C.
"TrapSuffix: Proactive Defense Against Adversarial Suffixes in Jailbreaking"[95]	Feb 2026	Du, M., Fang, H., Ma, H., …, Ji, S., Chang, E.C.

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

CHUA Tat Seng is the KITHCT Chair Professor at the NUS School of Computing and an AI and Domain Lead for Responsible and Safe AI at the NUS Artificial Intelligence Institute. He is also co-Director of NExT, a joint research center of NUS and Tsinghua University. His AI safety research includes work on jailbreaks, safety evaluation, hallucination, agent safety, and LLM alignment.

Selected papers:

"Risky-Bench: Probing Agentic Safety Risks under Real-World Deployment"[96]	Feb 2026	Zheng, J., Luo, Y., Xu, J., …, Zhang, A., Chua, T.S.
"Controllable Value Alignment in Large Language Models through Neuron-Level Editing"[97]	Feb 2026	Yang, Y., Li, J., Liu, J., …, Hong, R., Chua, T.S.
"Lingua-SafetyBench: A Benchmark for Safety Evaluation of Multilingual Vision-Language Models"[98]	Jan 2026	Shi, E., Shao, P., Zhang, Y., …, Shen, F., Chua, T.S.

DONG Jin Song: Professor, School of Computing

DONG Jin Song is a Professor at the NUS School of Computing. His research spans formal methods with AI agents, safety and security systems, trusted AI, probabilistic reasoning, sports analytics, and verified AI code synthesis. Within AI safety, his recent work includes secure and trustworthy agents, prompt injection and agent-compromise risks, and safety and privacy risks in multimodal AI systems.

Selected papers:

"Zombie Agents: Persistent Control of Self-Evolving LLM Agents via Self-Reinforcing Injections"[99]	Mar 2026	Yang, X., He, Y., Ji, S., Hooi, B., Dong, J.S.
"Enhancing Model Defense Against Jailbreaks with Proactive Safety Reasoning"[100]	Jan 2026	Yang, X., Deng, G., Shi, J., Zhang, T., Dong, J.S.
"DRIP: Defending Prompt Injection via Token-wise Representation Editing and Residual Instruction Fusion"[101]	Nov 2025	Liu, R., Lin, Y., Huang, Z., Dong, J.S.

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

Bryan Hooi is an Assistant Professor at the NUS School of Computing, and is affiliated with the NUS Institute of Data Science. His research aims to make machine learning systems more reliable and applicable to a wider variety of real-world contexts. His recent AI safety work includes prompt injection defense, jailbreak-related work, unlearning, and improving the reliability of vision-language models.

Selected papers:

"LookAhead Tuning: Safer Language Models via Partial Answer Previews"[102]	Feb 2026	Liu, K., Wang, M., Luo, Y., …, Hooi, B., Deng, S.
"Automating Steering for Safe Multimodal Large Language Models"[103]	Sep 2025	Wu, L., Wang, M., Xu, Z., …, Hooi, B., Deng, S.
"Robustness via Referencing: Defending against Prompt Injection Attacks by Referencing the Executed Instruction"[104]	Apr 2025	Chen, Y., Li, H., Sui, Y., …, Song, Y., Hooi, B.

KAN Min-Yen: Associate Professor, School of Computing

KAN Min-Yen is an Associate Professor and Vice Dean of Undergraduate Studies at NUS. He leads the Web, Information Retrieval / Natural Language Processing Group (WING) at the School of Computing. His research spans digital libraries, natural language processing, and information retrieval, with recent work focusing on safety, ethics, multicultural issues in foundation models, retrieval-augmented generation, and misinformation-related research.

Selected papers:

"The Facade of Truth: Uncovering and Mitigating LLM Susceptibility to Deceptive Evidence"[105]	Jan 2026	Wan, H., Wu, J., Luo, M., Li, F., Zeng, Z., Kan, M.Y.
"What's Left Unsaid? Detecting and Correcting Misleading Omissions in Multimodal News Previews"[106]	Jan 2026	Li, F., Wu, J., Fu, T., …, Zhou, W., Kan, M.Y.
"LLMs Are Biased Towards Output Formats! Systematically Evaluating and Mitigating Output Format Bias of LLMs"[107]	Apr 2025	Do, X., Nguyen, N., Sim, T., …, Chen, N., Kan, M.Y.

Mohan KANKANHALLI: Professor of Computer Science and Director of NAII

Mohan KANKANHALLI is Provost's Chair Professor of Computer Science at NUS, Director of the NUS Artificial Intelligence Institute (NAII), and Deputy Executive Chairman (Talent) of AI Singapore, Singapore's national AI R&D program. His broader research spans multimodal computing, computer vision, and trustworthy AI, and recent AI safety work includes research on hallucinations, jailbreaking, machine unlearning and value alignment.

Selected papers:

"Do Prompts Guarantee Safety? Mitigating Toxicity from LLM Generations through Subspace Intervention"[108]	Feb 2026	Singh, H., Xu, Z., Subramanyam, A., Kankanhalli, M.
"LLMs Can Unlearn Refusal with Only 1,000 Benign Samples"[109]	Jan 2026	Guo, Y., Xu, Z., Liu, S., Zheng, Z., Kankanhalli, M.
"Strong Preferences Affect the Robustness of Preference Models and Value Alignment"[110]	Mar 2025	Xu, Z., Kankanhalli, M.

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

Bryan LOW is an Associate Professor of Computer Science at NUS, the Director of AI Research at AI Singapore, and the Deputy Director of the NUS AI Institute. He also leads GLOW.AI, whose research includes AI for science, data-centric AI, and AI applications. His recent safety-related work includes inference-time safety steering, interpretable in-context learning, and evaluation metrics for machine unlearning.

Selected papers:

"BARRIERSTEER: LLM Safety via Learning Barrier Steering"[111]	Feb 2026	Tran, T., Verma, A., Wong, K., Low, B.K.H., Rus, D., Xiao, W.
"Position: Balance Human Agency & AI Assistance in the Tussle for the Right to"[112]	Feb 2026	Khoo, Z., Ryan, Y., Oo, N., …, Hahn, J., Low, B.K.H.
"WaterDrum: Watermarking for Data-centric Unlearning Metric"[113]	Feb 2026	Lu, X., Niu, X., Lau, G., …, Ng, S., Low, B.K.H.

TAN Zhi Xuan: Assistant Professor and A*STAR Research Scientist

TAN Zhi Xuan is an Assistant Professor in the NUS Department of Computer Science with a joint appointment at the A*STAR Institute of High Performance Computing (IHPC). They also lead the Cooperative Intelligence & Systems (CoSI) Lab, which focuses on scaling cooperative intelligence via rational, model-based AI engineering. Tan's research spans probabilistic programming, model-based planning, Bayesian inference, AI alignment, and computational cognitive science, with a focus on building reliable, coherent, and human-like cooperative systems.

Selected papers:

"Resource Rational Contractualism Should Guide AI Alignment"[114]	Mar 2026	Levine, S., Franklin, M., Zhi-Xuan, T., …, Lazar, S., Gabriel, I.
"Full-Stack Alignment: Co-Aligning AI and Institutions with Thick Models of Value"[115]	Dec 2025	Edelman, J., Zhi-Xuan, T., Lowe, R., …, Vendrov, I., Wilken-Smith, J.
"Beyond Preferences in AI Alignment"[116]	Jul 2025	Zhi-Xuan, T., Carroll, M., Franklin, M.

ZHANG Jiaheng: Assistant Professor, School of Computing

ZHANG Jiaheng is an Assistant Professor in the NUS School of Computing. He is also a Presidential Young Professor at NUS. His broader research spans AI safety, cryptography, privacy-preserving machine learning, and blockchain systems. His recent AI safety work includes research on LLM robustness and safeguards, including jailbreak attacks and defenses.

Selected papers:

"DiffuGuard: How Intrinsic Safety is Lost and Found in Diffusion Large Language Models"[117]	Mar 2026	Li, Z., Nie, Z., Zhou, Z., …, Guo, Y., Zhang, J.
"ExtendAttack: Attacking Servers of LRMs via Extending Reasoning"[118]	Mar 2026	Zhu, Z., Liu, Y., Xu, Z., …, Zhu, X., Zhang, J.
"DMark: Order-Agnostic Watermarking for Diffusion Large Language Models"[119]	Oct 2025	Wu, L., Zhong, L., Qu, W., …, Shen, C., Zhang, J.

4.1.2 Nanyang Technological University (NTU)

AI research at NTU is supported both by university-level coordination and by faculty-based labs. At the university level, the AI Research Institute (AI.R) serves as a coordinating platform for AI research across NTU and as an interface between NTU's AI research ecosystem and government and industry partners.[120] This university-level structure is complemented by a broader set of research institutes, centres, laboratories, and faculty-led research groups across NTU.

More technical AI safety-relevant work appears to be concentrated within the College of Computing and Data Science (CCDS), where faculty research groups and labs such as the Generative AI Lab (GrAIL), Cyber Security Lab (CSL), Computational Intelligence Laboratory (CIL), and Multimedia and Interactive Computing Lab (MICL) publish work on AI safety.

GrAIL focuses on generative models and explicitly includes trustworthy AI among its research areas; CSL works on the security of high-assurance systems and emphasizes techniques such as formal methods, program analysis, machine learning, and AI for system correctness and attack detection; CIL is a center for both teaching and research, spanning classical AI, machine learning, adaptive systems, and computational intelligence; MICL brings together work in computer vision, language, graphics, and interactive computing, with research on intelligent processing and synthesis of images, audio, and video.

Bo AN: Professor and Co-Director of AI.R

Bo AN is a President's Chair Professor and Head of the Division of Artificial Intelligence in the College of Computing and Data Science at NTU. He is also Co-Director of AI.R. His research spans artificial intelligence, multi-agent systems, reinforcement learning, computational game theory, and optimization, with recent work also engaging questions of agentic AI safety, alignment, and trustworthy decision-making.

Selected papers:[g]

"Empirical Study on Robustness and Resilience in Cooperative Multi-Agent Reinforcement Learning"[121]	Oct 2025	Li, S., Mao, Z., Li, H., …, An, B., Yang, Y., Lv, W., Liu, X.
"Vulnerable Agent Identification in Large-Scale Multi-Agent Reinforcement Learning"[122]	Sep 2025	Li, S., Zheng, Y., Mao, Z., …, An, B., Yang, Y., Lv, W., Liu, X.
"A Survey on Trustworthy LLM Agents: Threats and Countermeasures"[123]	Aug 2025	Yu, M., Meng, F., Zhou, X., …, An, B., Wen, Q.

LAM Kwok-Yan: Professor and Singapore AISI Executive Co-Director

LAM Kwok-Yan 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.[124] His technical AI safety work centers on machine unlearning, LLM threat modeling, and agentic safety.

Selected papers:

"Plato's Form: Toward Backdoor Defense-as-a-Service for LLMs with Prototype Representations"[125]	Feb 2026	Chen, C., Sun, Y., Gao, J., …, Wang, Q., Lam, K.Y.
"RedVisor: Reasoning-Aware Prompt Injection Defense via Zero-Copy KV Cache Reuse"[126]	Feb 2026	Liu, M., Zhang, S., Long, C., Lam, K.Y.
"The Shadow Self: Intrinsic Value Misalignment in Large Language Model Agents"[127]	Jan 2026	Chen, C., Kim, Y., Yang, Y., …, Zheng, Y., Lam, K.Y.

LIU Yang: Professor and Executive Director of Cyber Security Research Centre

LIU Yang is a Professor at NTU College of Computing & Data Science, SCDT Research Group, and CSL. He is Executive Director of the Cyber Security Research Centre (CYSREN) at NTU, and Executive

g. These papers are a collaboration between five academic and research institutions, with Bo AN and his team representing NTU.

Director of the CyberSG R&D Programme Office (CRPO). His recent AI safety-related work includes AI security, trustworthy AI, and attacks on LLM safeguards, including jailbreak research.

Selected papers:

"Semantic-Aligned Adversarial Evolution Triangle for High-Transferability Vision-Language Attack"[128]	Jun 2025	Jia, X., Gao, S., Guo, Q., …, Liu, Y., Cao, X.
"Evolution-based Region Adversarial Prompt Learning for Robustness Enhancement in Vision-Language Models"[129]	Mar 2025	Jia, X., Gao, S., Qin, S., …, Liu, Y., Cao, X.
"CORBA: Contagious Recursive Blocking Attacks on Multi-Agent Systems Based on Large Language Models"[130]	Feb 2025	Zhou, Z., Li, Z., Zhang, J., …, Liu, Y., Guo, Q.

LUU Anh Tuan: Associate Professor, CCDS

LUU Anh Tuan is an Associate Professor at NTU College of Computing & Data Science, AI Research Group, and GrAIL. Luu was formerly a research scientist at the Institute for Infocomm Research (I2R), A*STAR. His research lies at the intersection of AI, deep learning, and natural language processing, with a focus on large language models, graph neural networks, trustworthy AI, and natural language processing (NLP) applications. His recent AI safety work focuses especially on AI security, including backdoor attacks and defenses.

Selected papers:

"UniFLE: Uniform Fusion of Multiple LoRA Experts for Backdoor Defense in Large Language Models"[131]	Feb 2026	Zhao, S., Lin, Q., Jia, Y., …, Li, Y., Luu, A.T.
"P2P: A Poison-to-Poison Remedy for Reliable Backdoor Defense in LLMs"[132]	Oct 2025	Zhao, S., Wu, X., Zhao, S., …, Jia, Y., Luu, A.T.
"Rethinking Reasoning: A Survey on Reasoning-based Backdoors in LLM"[133]	Oct 2025	Hu, M., Wu, X., Suo, Z., …, Luu, A.T., Zhao, S.

Luke ONG: Distinguished University Professor and Vice President

Luke is Distinguished University Professor at NTU, where he serves as Vice President (AI and Digital Economy) and Founding Dean of the College of Computing and Data Science. He is also Deputy Executive Chairman (Applied and Translational) and the Chief Scientist at AI Singapore. His broader research spans semantics of computation, programming languages, verification, logic, and algorithms. His recent AI safety-related work includes machine unlearning and broader work on AI safety frameworks and verification-oriented approaches to trustworthy AI.

Selected papers:

"SB-TRPO: Towards Safe Reinforcement Learning with Hard Constraints"[134]	Jan 2026	Kanwar, A., Wagner, D., Ong, L.
"Reinforcement Learning with ω-Regular Objectives and Constraints"[135]	Nov 2025	Wagner, D., Witzman, L., Ong, L.
"Open Problems in Machine Unlearning for AI Safety"[136]	Jan 2025	Barez, F., Fu, T., Prabhu, A., …, Ong, L., …, Geva, M., Gal, Y.

Soujanya PORIA: Associate Professor, Electrical & Electronic Engineering

Soujanya PORIA is an Associate Professor at NTU's School of Electrical and Electronic Engineering. Before that, he was an Associate Professor at SUTD and a senior scientist at A*STAR. His broader research spans multimodal AI and LLMs, with recent work in AI safety focusing on alignment, and on robustness in LLM-based multi-agent systems.

Selected papers:

"LLM Alignment should go beyond Harmlessness–Helpfulness and incorporate Human Agency"[137]	Mar 2026	Naseem, U., Chakraborty, T., Chang, K., …, Peng, N., Poria, S.
"OffTopicEval: When Large Language Models Enter the Wrong Chat, Almost Always!"[138]	Mar 2026	Lei, J., Gumma, V., Bhardwaj, R., …, Zadeh, A., Poria, S.
"Measuring and Enhancing Trustworthiness of LLMs in RAG through Grounded Attributions and Learning to Refuse"[139]	Apr 2025	Song, M., Sim, S., Bhardwaj, R., …, Majumder, N., Poria, S.

TAO Dacheng: Distinguished University Professor, CCDS

TAO Dacheng is a Distinguished University Professor at NTU College of Computing & Data Science and the CVL Research Group. He also leads the Generative AI Lab (GrAIL), which conducts research on computer vision and text, video, and audio generative models. His recent AI safety work has centered on multimodal foundation models. This includes studying jailbreak vulnerabilities and defenses against backdoor attacks in multimodal LLM and vision language model (VLM) systems.

Selected papers:

"SRD: Reinforcement-Learned Semantic Perturbation for Backdoor Defense in VLMs"[140]	Mar 2026	Xu, S., Liang, S., Zheng, H., …, Rutkowski, L., Tao, D.
"Odysseus: Jailbreaking Commercial Multimodal LLM-integrated Systems via Dual Steganography"[141]	Dec 2025	Li, S., Cheng, J., Li, Y., Jia, X., Tao, D.
"SafeBench: A Safety Evaluation Framework for Multimodal Large Language Models"[142]	Oct 2025	Ying, Z., Liu, A., Liang, S., …, Liu, X., Tao, D.

ZHANG Tianwei: Associate Professor, CCDS

ZHANG Tianwei is an Associate Professor in the College of Computing and Data Science at and Deputy Director of CYSREN at NTU. His broader research spans computer systems security, trusted and trustworthy machine learning, software and hardware security, privacy, and cloud computing. His recent AI safety-related work includes security and robustness research on large language and reasoning models.

Selected papers:

"On the Adversarial Robustness of Large Vision-Language Models under Visual Token Compression"[143]	Jan 2026	Zhang, X., Liu, H., Bai, L., …, Zhang, T., Hu, H.
"SafeRedir: Prompt Embedding Redirection for Robust Unlearning in Image Generation Models"[144]	Jan 2026	Liu, R., Chen, K., Qiu, H., …, Zhang, T., Ng, S.
"Picky LLMs and Unreliable RMs: An Empirical Study on Safety Alignment after Instruction Tuning"[145]	Feb 2025	Li, G., Chen, K., Guo, S., …, Zhang, T., Li, J.

4.1.3 Singapore Management University (SMU)

AI-related research at SMU is centered in the School of Computing and Information Systems (SCIS). SCIS has an interdisciplinary structure, comprising three core research areas and four integrative research areas. Among these, the integrative area on Safety, Security and Fairness is particularly relevant to AI safety, covering four directions: Security and Governance of Software/AI Systems, Trustworthiness of Digital Platforms and Devices, Misinformation and Disinformation, and Privacy-Preserving Data Sharing and Analytics.[146]

SUN Jun: Professor and Lead Principal Investigator, SCIS

SUN Jun is a Professor of Computer Science at the SMU School of Computing and Information Systems, and the Co-Director of the Centre for Research for Intelligent Software Engineering at SMU. His research applies formal methods to enhance the safety and security of complex systems, with a recent emphasis on frontier AI systems. He also studies how AI is transforming software engineering, including the extent to which AI systems may substitute for human programmers. His recent AI safety work includes work on agent safety, LLM robustness and security, backdoor and data-protection risks, and alignment-related vulnerabilities.

Selected papers:

"ProbGuard: Probabilistic Runtime Monitoring for LLM Agent Safety"[147]	Mar 2026	Wang, H., Poskitt, C., Wei, J., Sun, J.
"AIR: Improving Agent Safety through Incident Response"[148]	Feb 2026	Xiao, Z., Sun, J., Chen, J.
"AgentSpec: Customizable Runtime Enforcement for Safe and Reliable LLM Agents"[149]	Jul 2025	Wang, H., Poskitt, C., Sun, J.

4.1.4 Singapore University of Technology and Design (SUTD)

AI-centric research at SUTD is anchored largely in the Information Systems Technology and Design pillar, one of the university's focus areas, which emphasizes computing, systems, and intelligence.[h] However, SUTD does not appear to frame AI safety as a standalone institutional research focus. Instead, its interdisciplinary model means that relevant work is distributed across broader areas such as responsible AI, trustworthy language models, and the societal deployment of AI systems. We noticed that SUTD's strengths are in the social and multilingual dimensions of AI, with a few researchers working on multilingual LLM capability and alignment.

h. 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)

Roy LEE Ka-Wei: Assistant Professor

Roy LEE is an Assistant Professor in the Information Systems Technology and Design pillar at SUTD, a faculty member of the Design and Artificial Intelligence program, and lead of the Social AI Studio. His research interests span responsible AI, social computing, and machine learning. His recent work on AI safety focuses especially on multilingual safety in Singapore, including the development of localized benchmarks and safeguards for low-resource and code-mixed settings such as Singlish, Malay, and Tamil. Beyond this, his work also covers agentic AI governance, persuasion, hallucination, and misinformation.

Selected papers:

"Lost in Localization: Building RabakBench with Human-in-the-Loop Validation to Measure Multilingual Safety Gaps"[150]	Feb 2026	Chua, G., Tan, L., Ge, Z., Lee, R.K.W.
"Persuasion Dynamics in LLMs: Investigating Robustness and Adaptability in Knowledge and Safety with DuET-PD"[151]	Nov 2025	Tan, B., Chin, D., Liu, Z., Chen, N., Lee, R.K.W.
"Toxicity Red-Teaming: Benchmarking LLM Safety in Singapore's Low-Resource Languages"[152]	Nov 2025	Hu, Y., Hee, M., Nakov, P., Lee, R.K.W.

4.2 Public-sector and Research Institutions

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

A*STAR is Singapore's leading public-sector R&D agency, with a mission to advance scientific discovery and develop innovative technology for economic impact. Its research ecosystem spans biomedical sciences, physical sciences, and engineering, through a network of research institutes, national centers, and programs. Within this landscape, AI safety-relevant work is distributed across several institutes, most notably the Institute for Infocomm Research (I²R), the Institute of High Performance Computing (IHPC), and the Centre for Frontier AI Research (CFAR).[153]

While I²R and IHPC contribute through work on interpretability, robustness, efficient AI, and multimodal systems, CFAR is the most prominent focal point for frontier AI research with direct safety relevance. Officially opened in 2022, CFAR develops next-generation AI technologies through five research pillars, including Resilient and Safe AI, which focuses on robust and secure AI systems. CFAR now has two researchers who meet our criteria.

Joey Tianyi ZHOU: CFAR Deputy Director

Joey ZHOU is the Deputy Director at CFAR. He also holds joint appointments with the Centre for Advanced Technologies in Online Safety (CATOS) as Principal Scientist, and as Adjunct Faculty at NUS.

His research interests span robust and efficient machine learning, trustworthy AI, and foundation model security. His recent AI safety work covers LLM and LLM-agent full-stack safety, social bias attacks, backdoor vulnerabilities, and benchmarks for detecting AI-generated video and deepfake content.

Selected papers:

"AEGIS: Authenticity Evaluation Benchmark for AI-Generated Video Sequences"[154]	Oct 2025	Li, J., Zhang, X., Zhou, J.T.
"TokenSwap: Backdoor Attack on the Compositional Understanding of Large Vision-Language Models"[155]	Sep 2025	Zhang, Z., Tao, Q., Lv, J., …, Feng, L., Zhou, J.T.
"Understanding Large Language Model Vulnerabilities to Social Bias Attacks"[156]	Jul 2025	Zhao, J., Fang, M., Ye, F., …, Zhou, J.T., Pechenizkiy, M.

ZHANG Jie: CFAR Innovation Lead

ZHANG Jie currently serves as the Innovation Lead at CFAR. His recent work on AI safety includes mechanistic analysis of emergent misalignment in LLMs, detection of prompt-based attacks, robust unlearning, and content safety control for image-generation models. He has also worked on agent-related risk control in scientific AI systems, including the use of agent-based guardrails in chemical science.

Selected papers:

"Character as a Latent Variable in Large Language Models: A Mechanistic Account of Emergent Misalignment and Conditional Safety Failures"[157]	Jan 2026	Su, Y., Zhou, W., Zhang, T., …, Yu, N., Zhang, J.
"Controlling risks of AI in chemical science with agents"[158]	Sep 2025	He, J., Guan, H., Feng, W., …, Zhang, J., Zhou, W.
"JailGuard: A Universal Detection Framework for Prompt-based Attacks on LLM Systems"[159]	Dec 2025	Zhang, X., Zhang, C., Li, T., …, Zhang, J., …, Ma, S., Shen, C.

4.2.2 Singapore AISI

The Digital Trust Centre (DTC)—a national research center at NTU focused on trust technologies in cybersecurity, privacy-preserving technologies, and trustworthy AI—was designated as the Singapore AI Safety Institute (AISI) in May 2024. The AISI brings together researchers across Singapore to conduct research on AI safety evaluation and testing, with work organized around four broad areas: (1) Testing and Evaluation; (2) Safe Model Design, Development and Deployment; (3) Content Assurance; and (4) Governance and Policy.[160]

Since the 2025 report, their research profile has broadened to include LLM-based agent safety and security. We have profiled Singapore AISI Co-Executive Director LAM Kwok-Yan above under NTU.

4.2.3 Government Technology Agency

The Government Technology Agency (GovTech) was launched to harness technology and drive the Smart Nation initiative, after the restructuring of the Infocomm Development Authority (IDA) in 2016. It 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.[161] Within GovTech, the Data Science & AI Division develops AI products and capabilities that can be deployed across government agencies, allowing teams to build on shared infrastructure and tools rather than starting from scratch.[162]

While GovTech focuses on creating tools rather than academic research, its AI work increasingly touches on questions of responsible deployment, evaluation, and governance of real-world AI systems. In particular, GovTech's Responsible AI team has produced research relevant to AI safety, especially in areas such as multilingual safety alignment, evaluation of application-level safety risks, and governance frameworks for agentic AI systems. Although none of GovTech's engineers or researchers individually met the threshold for inclusion as technical AI safety researchers under this report's criteria, papers by the responsible AI team nonetheless indicate a growing internal capability and interest in applied AI safety research.

Responsible AI Team

Selected papers:

"With Great Capabilities Come Great Responsibilities: Introducing the Agentic Risk & Capability Framework for Governing Agentic AI Systems"[163]	Dec 2025	Khoo, S., Foo, J., Lee, R.K.W.
"Measuring What Matters: A Framework for Evaluating Safety Risks in Real-World LLM Applications"[164]	Jul 2025	Goh, J., Khoo, S., Iskandar, N., Chua, G., Tan, L., Foo, J.
"Safe at the Margins: A General Approach to Safety Alignment in Low-Resource English Languages—A Singlish Case Study"[165]	Apr 2025	Lim, I., Khoo, S., Lee, R.K.W., Chua, W., Goh, J., Foo, J.

4.3 Overall Trends

Research Output

Singapore's AI safety research community has grown over the past year. The number of researchers meeting our criteria has risen from 14 to 21, with new inclusions concentrated at NUS (Dong Jin Song, Tan Zhi Xuan, Chang Ee Chien) and NTU (Bo An, Soujanya Poria). SMU and SUTD remain stable by headcount, though with some movement: Roy Lee is a new SUTD inclusion, while Soujanya Poria moved from SUTD to NTU in July 2025.

This growth is taking place against a backdrop of increased national investment in frontier and responsible AI research. In January 2026, the Singapore government announced an additional investment of over S$1 billion over five years, from 2025 to 2030, under the National AI Research and Development (NAIRD) Plan, more than doubling the earlier tranche of over S$500 million committed from 2019 to 2023.[166][167] The plan will launch new Research Centres of Excellence in areas such as responsible AI, resource-efficient AI, emerging AI methodologies, and general-purpose AI, signaling stronger institutional support for AI safety-relevant research within Singapore's national AI R&D agenda.[168]

Two notable institutional developments are worth highlighting. First, A*STAR's CFAR is emerging as a visible node of frontier AI safety research. CFAR was established in 2022 to advance use-inspired basic research in AI within A*STAR. It now has two researchers who meet our threshold—Joey Zhou and Jie Zhang—making it the clearest concentration of frontier AI safety research outside the universities. Second, NTU's AI.R was established to coordinate AI research within the university. Whereas NUS's NAII has pillars that explicitly reflect responsible AI and the societal implications of AI, AI.R is better understood as a broader university-wide coordination platform for AI research, within which safety-relevant work may increasingly be concentrated.

By contrast, industry-based technical AI safety research remains limited. While there are signs of some relevant output from companies—including researchers at ByteDance with work touching on safety-related topics—company-based research in Singapore still appears relatively sparse and is not yet a major driver of the local technical AI safety ecosystem.[169] For now, the center of gravity remains in the universities and public research institutes.

Research Focus Areas

In terms of research focus, the broad picture remains similar to the previous report, but with some shifts in emphasis. The strongest and most consistent themes remain LLM robustness and jailbreaking, backdoor attacks, multilingual safety, hallucinations, and model unlearning. Within this overall profile, two comparative strengths are worth highlighting. First, Singapore researchers are producing a notable amount of safety evaluation infrastructure. Benchmarks and evaluation frameworks such as RiskyBench, AgentNoiseBench, SafeBench, AEGIS, and RabakBench appear among the selected papers, representing a meaningful contribution that goes beyond patching individual failure modes to provide shared tools for the broader safety community. Combined with multilingual safety work from SUTD and GovTech—including localized benchmarks, moderation systems, and analyses of safety gaps in low-resource and/or code-mixed settings such as Singlish, Malay, and Tamil—this gives Singapore's ecosystem a genuine advantage in safety evaluation and deployment-oriented safety research.

Second, there is evidence that agentic AI safety is becoming more prominent. We identified 19 papers on agentic safety topics published by listed researchers from July 2025 to May 2026, compared with seven in the January 2024 to July 2025 period. Adjusting for the different time windows, this represents an increase from roughly 0.4 to 1.9 papers per month, though this comparison should be read cautiously, as the researcher pool also expanded over this period and topic classifications involve subjective judgment.[i] Even so, the trend appears consistent across institutions rather than being driven by any single research group, and is reflected in the wider range of agent-related topics now being covered: LLM-agent full-stack safety, trust and failure modes in multi-agent systems, agent monitoring and incident response, and agent-specific attack and defense dynamics. This suggests that local research attention is beginning to extend beyond model-level vulnerabilities toward system-level risks in downstream applications, especially in more autonomous, tool-using, and interactive agentic systems.

Notable Gaps and Future Directions

We note two gaps in our observations. First, the research profile of the ecosystem is methodologically narrow. Roughly 60–70% of identified safety papers fall within the empirical attack-and-defense paradigm such as identifying vulnerabilities and proposing defenses through experimental evaluation. There is comparatively less work on theoretical safety, formal verification for frontier AI systems, or interpretability-first approaches. While there are important exceptions in adjacent areas of safety-by-design — including Dong Jin Song's and

i. This figure is derived from counting agentic AI safety and related multi-agent robustness papers authored by the researchers listed in this report. It is intended as an indicator of changing emphasis within our sample, rather than a comprehensive count of all Singapore-based research on these topics. Additional relevant papers may exist outside the set of researchers included here or may not have been captured in our scan.

Sun Jun's work on formal methods for AI safety, and Tan Zhi Xuan's and Mohan Kankanhalli's work on alignment — the overall ecosystem still appears stronger at identifying and patching failure modes than at explaining the deeper causes of model failures or building safety by design.

Second, alignment research remains shallower than the labels sometimes suggest. Several researchers are described as working on "alignment" based on our methodology, but the underlying papers often address narrower properties such as harmlessness constraints, toxicity reduction, or value elicitation in bounded settings. More foundational agendas—such as scalable oversight, reward modeling, or other deeper approaches to aligning advanced systems with human values and institutions—remain relatively underrepresented.

At the same time, this snapshot is timely. It comes roughly one year after the release of the *Singapore Consensus on Global AI Safety Research Priorities* at SCAI 2025, which organized technical AI safety priorities around development, assessment, and control challenges for trustworthy, reliable, and secure AI systems. As Singapore moves toward the next phase of this agenda—including likely follow-on discussions around the Consensus in May 2026—the current distribution of research strengths and gaps may help indicate where future work could be encouraged, especially in underexplored priority areas such as control, dangerous capabilities, and higher-stakes frontier-risk evaluation.

Conclusion

Since 2025, Singapore has transitioned from a primarily principles-based governance ecosystem to a maturing, implementation- and testing-focused AI safety ecosystem. The *Model AI Governance Framework for Agentic AI* and the *Agentic Risk and Capability Framework* reflect a shift toward addressing increasingly autonomous systems with specific lifecycle-based technical controls. This domestic orientation is reinforced by a robust testing infrastructure, including the Global AI Assurance Sandbox and the *Starter Kit for Testing LLM-Based Applications*, which seek to codify testing methodologies so that testers need not rely on ad-hoc evaluations.

Internationally, Singapore has successfully leveraged its domestic progress to position itself as a practical convenor and technical contributor. Through the International Network for Advanced AI Measurement, Evaluation and Science and bilateral partnerships like the Singapore–Korea AISI joint testing, Singapore is actively shaping global norms around safety evaluation and measurement. Its leadership within ASEAN further suggests a strategic role in translating these international scientific standards into regional benchmarks and tools that reflect Southeast Asian linguistic and cultural contexts.

The homegrown AI ecosystem and technical research base provide the necessary technical depth to support these policy objectives. The expansion of the SEA-LION and MERaLiON model families, alongside localized safety tools like SEA-Guard and LionGuard 2, demonstrates that safety is increasingly being embedded directly into the homegrown stack. This is mirrored in the expansion of Singapore's technical AI safety research community, which shows a distinct comparative strength in safety evaluation and multilingual deployment research. However, significant gaps remain in foundational alignment and theoretical safety research, suggesting that while Singapore excels at applied safety and patching vulnerabilities, more focus on safety-by-design is required.

Ultimately, Singapore demonstrates how a smaller state can influence global AI governance by focusing on pragmatic, science-based tools and cross-border technical collaboration. As the ecosystem moves toward the International Scientific Exchange on AI Safety 2026 and subsequent updates to national research priorities, the challenge will be to bridge existing methodological gaps and ensure that governance frameworks remain resilient to the rapid emergence of agentic and systemic risks.

Appendix

Table of common Acronyms and Abbreviations

Abbreviation/Acronym	Full Name
A*STAR	Agency for Science, Technology and Research
AI.R	Nanyang Technological University AI Research Institute
AISI	AI Safety/Security Institute
AITC	AI Technical Committee
AIVF	AI Verify Foundation
ALT-EDIC	EU Alliance for Language Technologies European Digital Infrastructure Consortium
ANSI	American National Standards Institute
ARC	Agentic Risk & Capability Framework
ASEAN	Association of Southeast Asian Nations
CBRNE	Chemical, Biological, Radiological, Nuclear, and Explosives
CFAR	Centre for Frontier AI Research
CII	Critical Information Infrastructure
CSA	Cyber Security Agency
DTC	Digital Trust Centre
ESG	Enterprise Singapore
EU	European Union
GovTech	Government Technology Agency
GPAI	General-purpose AI
HLAB-AI	UN High-level Advisory Body on Artificial Intelligence
HTX	Home Team Science and Technology Agency

Continues on next page...

Abbreviation/Acronym	Full Name
I2R	Institute for Infocomm Research
IEC	International Electrotechnical Commission
IHPC	Institute of High Performance Computing
IMDA	Infocomm Media Development Authority
ISO	International Organization for Standardization
ITSC	Information Technology Standards Committee
JTC	ISO/IEC Joint Technical Committee
LLM	Large language models
MDDI	Ministry of Digital Development and Information
MERaLiON	Multimodal Empathetic Reasoning and Learning in One Network
MGF	Model AI Governance Framework
MOU	Memorandum of Understanding
NAII	National University of Singapore AI Institute
NAIRD	National AI Research and Development
Network	International Network for Advanced AI Measurement, Evaluation and Science
NIST	U.S. National Institute of Standards and Technology
NLP	Natural language processing
NTU	Nanyang Technological University
NUS	National University of Singapore
PDPC	Personal Data Protection Commission
RLHF	Reinforcement Learning from Human Feedback
SCAI	Singapore Conference on AI
SEA-LION	Southeast Asian Languages in One Network
SGDG	Singapore Digital Gateway
SMU	Singapore Management University

Continues on next page...

Abbreviation/Acronym	Full Name
SS	Singapore Standard
SSC	Singapore Standards Council
SUTD	Singapore University of Technology and Design
TR	Technical Reference
UN	United Nations
WEF	World Economic Forum
WG-AI	ASEAN Working Group on AI Governance

Notes

1 Yoshua Bengio et al., *International AI Safety Report*, technical report DSIT 2026/001 (Department for Science, Innovation and Technology, 2026), accessed May 6, 2026, https://internationalaisafetyreport.org/

2 Singapore Conference on AI, *The Singapore Consensus on Global AI Safety Research Priorities*, 2025, accessed May 8, 2026, https://file.go.gov.sg/sg-consensus-ai-safety.pdf

3 Ministry of Law, *Launch of Guide for Using Generative Artificial Intelligence in the Legal Sector*, March 2026, accessed May 6, 2026, https://www.mlaw.gov.sg/launch-of-guide-for-using-generative-artificial-intelligence-in-the-legal-sector/

4 Ministry of Trade and Industry, *Joint Advisory: Export controls on advanced semiconductor and artificial intelligence (AI) technologies*, April 2025, accessed May 6, 2026, https://www.mti.gov.sg/newsroom/joint-advisory--export-controls-on-advanced-semiconductor-and-artificial-intelligence--ai--technologies/

5 Singapore AI Safety Institute, *AISI - The Singapore AI Safety Institute*, February 2025, accessed May 6, 2026, https://sgaisi.sg/about-us/

6 Enterprise Singapore, *Singapore Standards Council (SSC)*, accessed May 6, 2026, https://www.enterprisesg.gov.sg/grow-your-business/boost-capabilities/quality-and-standards/singapore-standards-council

7 Infocomm Media Development Authority, *IT Standards Committee (ITSC)*, accessed May 6, 2026, https://www.imda.gov.sg/regulations-and-licensing-listing/ict-standards-and-quality-of-service/industry-committees-and-working-groups/it-standards-committee

8 Infocomm Media Development Authority, *Technical Committees and Working Groups*, 2026, https://www.imda.gov.sg/-/media/imda/files/regulations-and-licensing/licensing/ict-standards-and-quality-of-service/industry-committees-and-working-groups/it-standards-committee/technical-committees-and-working-groups.pdf

9 Infocomm Media Development Authority, *Model AI Governance Framework for Agentic AI*, January 2026, accessed May 6, 2026, https://www.imda.gov.sg/-/media/imda/files/about/emerging-tech-and-research/artificial-intelligence/mgf-for-agentic-ai.pdf

10 Infocomm Media Development Authority, *Model AI Governance Framework for Agentic AI*, January 2026, accessed May 6, 2026, https://www.imda.gov.sg/-/media/imda/files/about/emerging-tech-and-research/artificial-intelligence/mgf-for-agentic-ai.pdf

11 Ministry of Digital Development and Information, *Singapore Launches New Model AI Governance Framework for Agentic AI*, January 2026, accessed May 6, 2026, https://www.mddi.gov.sg/newsroom/singapore-launches-new-model-ai-governance-framework-for-agentic-ai--/

12 Cyber Security Agency of Singapore, *Securing Agentic AI: A Discussion Paper*, October 2025, accessed May 6, 2026, https://www.csa.gov.sg/resources/publications/securing-agentic-ai-a-discussion-paper/

13 AI Verify Foundation, *Global AI Assurance Pilot Case Studies*, May 2025, accessed May 6, 2026, https://assurance.aiverifyfoundation.sg/case-studies/

14 MLCommons, *AILuminate Multimodal*, accessed May 6, 2026, https://mlcommons.org/ailuminate/ailuminate-multimodal/

15 MLCommons, *AILuminate Security Introducing v0.5 of the Jailbreak Benchmark from MLCommons*, technical report (MLCommons, 2025), accessed May 8, 2026, https://mlcommons.org/wp-content/uploads/2025/10/MLCommons___Security___Jailbreak_0_5_Paper-5.pdf

16 Infocomm Media Development Authority and AI Verify Foundation, *Starter Kit for Testing LLM-Based Applications for Safety and Reliability*, January 2026, accessed May 6, 2026, https://www.imda.gov.sg/-/media/imda/files/about/emerging-tech-and-research/artificial-intelligence/starter-kit-for-testing-llm-based-applications-for-safety-and-reliability.pdf

17 Infocomm Media Development Authority and AI Verify Foundation, *Starter Kit for Testing LLM-Based Applications for Safety and Reliability*, January 2026, accessed May 6, 2026, https://www.imda.gov.sg/-/media/imda/files/about/emerging-tech-and-research/artificial-intelligence/starter-kit-for-testing-llm-based-applications-for-safety-and-reliability.pdf

18 Infocomm Media Development Authority, *Singapore Launches New Model AI Governance Framework for Agentic AI*, January 2026, accessed May 8, 2026, https://www.sgpc.gov.sg/detail?url=/media_releases/imda/press_release/P-20260122-2&page=/detail&HomePage=home

19 GovTech, *Risks of agentic systems - Agentic Risk & Capability Framework*, accessed May 6, 2026, https://govtech-responsibleai.github.io/agentic-risk-capability-framework/arc_framework/risks/

20 Infocomm Media Development Authority, *Singapore AI Safety Red Teaming Challenge 2026 - Key Observations*, March 2026, accessed May 6, 2026, https://www.imda.gov.sg/-/media/imda/files/about/emerging-tech-and-research/artificial-intelligence/sg-ai-safety-red-teaming-challenge-2026-key-observations.pdf

21 Singapore Statutes Online, *Criminal Law (Miscellaneous Amendments) Act 2025 - Singapore Statutes Online*, December 2025, accessed May 6, 2026, https://sso.agc.gov.sg/Acts-Supp/21-2025/Published/20251203?DocDate=20251203&ProvIds=P110-

22 gov.sg, *What is the Online Safety (Relief and Accountability) Bill (OSRA)?*, November 2025, accessed May 6, 2026, https://www.gov.sg/explainers/parliament-nov2025/

23 Singapore Statutes Online, *Online Safety (Relief and Accountability) Bill - Singapore Statutes Online*, October 2025, accessed May 6, 2026, https://sso.agc.gov.sg/Bills-Supp/18-2025/Published/20251015?DocDate=20251015&ProvIds=P15-

24 Monetary Authority of Singapore, *Consultation Paper on Proposed Guidelines on Artificial Intelligence Risk Management for Financial Institutions*, November 2025, accessed May 6, 2026, https://www.mas.gov.sg/publications/consultations/2025/consultation-paper-on-guidelines-on-artificial-intelligence-risk-management

25 Ministry of Digital Development and Information, *Speech by MOS Rahayu Mahzam at the Committee of Supply Debate 2026*, March 2026, accessed May 6, 2026, https://www.mddi.gov.sg/newsroom/-speech-by-mos-rahayu-mahzam-at-the-committee-of-supply-debate-2026/

26 Cyber Security Agency of Singapore, "Cybersecurity Act," accessed May 11, 2026, https://www.csa.gov.sg/legislation/cybersecurity-act/

27 Matthew Mohan, "CSA tasks critical information infrastructure leaders to review cyber risks due to AI-enabled threats," May 5, 2026, accessed May 11, 2026, https://www.channelnewsasia.com/singapore/csa-cyberattacks-ai-mythos-cii-owners-parliament-6100061

28 David Koh, *Cybersecurity Implications of Frontier AI*, Letter addressed to CII Owners and Board Chairmen, Cyber Security Agency of Singapore, May 2026

29 Cyber Security Agency of Singapore, "Advisory on Risks associated with Frontier AI Models," 2026, accessed May 11, 2026, https://www.csa.gov.sg/alerts-and-advisories/advisories/ad-2026-004/

30 Ministry of Digital Development and Information, "MDDI's Response to PQ on AI-enabled Cybersecurity Risks," accessed May 11, 2026, https://www.mddi.gov.sg/newsroom/mddi-s-response-to-pq-on-ai-enabled-cybersecurity-risks/

31 Enterprise Singapore, *Singapore Standards*, accessed May 6, 2026, https://www.singaporestandardseshop.sg/Product/SSPdtDetail/553b4562-ef85-4ebb-bfeb-2b35beb1d29c

32 Laura Caroli and Matt Mande, *What the UN Global Dialogue on AI Governance Reveals About Global Power Shifts*, October 2025, accessed May 6, 2026, https://www.csis.org/analysis/what-un-global-dialogue-ai-governance-reveals-about-global-power-shifts

33 Global Dialogue on AI Governance, *FAQ - Global Dialogue on AI Governance*, accessed May 6, 2026, https://www.un.org/global-dialogue-ai-governance/en/faq

34 Ministry of Digital Development and Information, *Artificial Intelligence*, March 2026, accessed May 6, 2026, https://www.digitalgateway.gov.sg/our-resources/artificial-intelligence/

35 Singapore AI Safety Institute, *Advancing Methodologies for Agentic Evaluations Across Domains | AISI*, July 2025, accessed May 6, 2026, https://sgaisi.sg/resources/international-joint-testing-exercise-agentic-testing/

36 International Network of AI Safety Institutes, *International-Joint-Testing-Exercise Evaluation Report*, July 2025, accessed May 6, 2026, https://sgaisi.sg/wp-api/wp-content/uploads/2025/07/International-Joint-Testing-Exercise_3JT-Eval-Report-v2.pdf

37 AI Security Institute, *International consensus and open questions in AI evaluations | AISI Work*, February 2026, accessed May 6, 2026, https://www.aisi.gov.uk/blog/international-ai-network-consensus-and-open-questions

38 AI Security Institute, *International consensus and open questions in AI evaluations | AISI Work*, February 2026, accessed May 6, 2026, https://www.aisi.gov.uk/blog/international-ai-network-consensus-and-open-questions

39 *India AI Impact Summit 2026*, February 2026, accessed May 6, 2026, https://impact.indiaai.gov.in/sessions?s=International+Network

40 Diya TV, *Best practices from the International Network for Advanced AI Measurement, Evaluation and Science.*, February 2026, accessed May 6, 2026, https://www.youtube.com/watch?v=sHc_7QNNPOw

41 Infocomm Media Development Authority, *Singapore Launches New Model AI Governance Framework for Agentic AI*, January 2026, accessed May 6, 2026, https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/press-releases/2026/new-model-ai-governance-framework-for-agentic-ai

42 Infocomm Media Development Authority, *Model AI Governance Framework*, January 2020, accessed May 6, 2026, https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/resource-for-organisation/ai/sgmodelaigovframework2.pdf

43 Drew Network Asia, *Singapore | New Governance Framework for Generative AI: IMDA Seeking Feedback Internationally*, 2024, accessed May 8, 2026, https://www.drewnetworkasia.com/newsroom/singapore-new-governance-framework-for-generative-ai-singapore-s-imda-seeking-feedback-internationally/

44 Infocomm Media Development Authority, *Model AI Governance Framework for Agentic AI*, January 2026, accessed May 6, 2026, https://www.imda.gov.sg/-/media/imda/files/about/emerging-tech-and-research/artificial-intelligence/mgf-for-agentic-ai.pdf

45 Infocomm Media Development Authority, *Factsheet - Model AI Governance Framework for Agentic AI*, January 2026, accessed May 6, 2026, https://www.imda.gov.sg/-/media/imda/files/news-and-events/media-room/media-releases/2026/01/factsheet-model-ai-governance-framework-for-agentic-ai.pdf

46 Ministry of Digital Development and Information, *Opening Keynote by Minister Josephine Teo at Preparing to Monitor the Impacts of Agents: Closing the Global Assurance Divide for Safe and Trusted AI*, February 2026, accessed May 6, 2026, https://www.mddi.gov.sg/newsroom/opening-keynote-by-minister-josephine-teo-at-preparing-to-monitor-the-impacts-of-agents--closing-the-global-assurance-divide-for-safe-and-trusted-ai/

47 Ministry of Digital Development and Information, *Closing Remarks by Minister Josephine Teo at the United Nations Office for Digital and Emerging Technologies (ODET): The Role of Science in International AI Governance Panel Session*, February 2026, accessed May 6, 2026, https://www.mddi.gov.sg/newsroom/closing-remarks-by-minister-josephine-teo-at-the-united-nations-office-for-digital-and-emerging-technologies--odet---the-role-of-science-in-international-ai-governance-panel-session/

48 Ministry of Digital Development and Information, *Opening Keynote by Minister Josephine Teo at Preparing to Monitor the Impacts of Agents: Closing the Global Assurance Divide for Safe and Trusted AI*, February 2026, accessed May 6, 2026, https://www.mddi.gov.sg/newsroom/opening-keynote-by-minister-josephine-teo-at-preparing-to-monitor-the-impacts-of-agents--closing-the-global-assurance-divide-for-safe-and-trusted-ai/

49 Ministry of Digital Development and Information, *Remarks by Minister Josephine Teo at AI Safety at the Global Level: Insights from Digital Ministers & Officials Panel*, February 2026, accessed May 6, 2026, https://www.mddi.gov.sg/newsroom/remarks-by-minister-josephine-teo-at-ai-safety-at-the-global-level--insights-from-digital-ministers---officials-panel/

50 Ministry of Digital Development and Information, *Minister Josephine Teo's Visit to India for India AI Impact Summit (19 Feb –20 Feb 2026) (Factsheet)*, February 2026, accessed May 6, 2026, https://www.mddi.gov.sg/newsroom/minister-josephine-teo-s-visit-to-india-for-india-ai-impact-summit--19-feb---20-feb-2026---factsheet--/

51 Ministry of External Affairs, India, *AI Impact Summit Declaration, New Delhi (February 18 - 19, 2026)*, 2026, https://www.mea.gov.in/bilateral-documents.htm?dtl/40809

52 ISO, *ISO/IEC JTC 1/SC 42 - Artificial intelligence*, September 2023, accessed May 6, 2026, https://www.iso.org/committee/6794475.html

53 Infocomm Media Development Authority, "Singapore Champions New Global AI Testing Standardisation Efforts on Benchmarking and Red Teaming," April 20, 2026, accessed May 11, 2026, https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/press-releases/2026/singapore-champions-new-global-ai-testing-standardisation-efforts

54 Infocomm Media Development Authority, "Singapore Champions New Global AI Testing Standardisation Efforts on Benchmarking and Red Teaming," April 20, 2026, accessed May 11, 2026, https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/press-releases/2026/singapore-champions-new-global-ai-testing-standardisation-efforts

55 Ministry of Digital Development and Information, *Minister Josephine Teo's Visit to India for India AI Impact Summit (19 Feb –20 Feb 2026) (Factsheet)*, February 2026, accessed May 6, 2026, https://www.mddi.gov.sg/newsroom/minister-josephine-teo-s-visit-to-india-for-india-ai-impact-summit--19-feb---20-feb-2026---factsheet--/

56 Ministry of Digital Development and Information, *Singapore Concludes Fruitful Chairmanship of the ASEAN Digital Ministers Meeting*, January 2025, accessed May 6, 2026, https://www.mddi.gov.sg/newsroom/singapore-concludes-fruitful-chairmanship-of-adgmin/

57 Ministry of Digital Development and Information, *Closing Remarks by Minister Josephine Teo at the United Nations Office for Digital and Emerging Technologies (ODET): The Role of Science in International AI Governance Panel Session*, February 2026, accessed May 6, 2026,

https://www.mddi.gov.sg/newsroom/closing-remarks-by-minister-josephine-teo-at-the-united-nations-office-for-digital-and-emerging-technologies--odet---the-role-of-science-in-international-ai-governance-panel-session/

58 Jonathan Lee et al., *State of AI Safety in Singapore (2025)*, technical report (Concordia AI, 2025), accessed May 8, 2026, https://concordia-ai.com/wp-content/uploads/2025/07/State-of-AI-Safety-in-Singapore-2025.pdf

59 Ministry of Digital Development and Information, *New Singapore UK Agreement to Strengthen Global AI Safety and Governance*, November 2024, accessed May 6, 2026, https://www.mddi.gov.sg/newsroom/new-singapore-uk-agreement-to-strengthen-global-ai-safety-governance/

60 Ministry of Digital Development and Information, *Singapore and the European Union Agree to Strengthen Collaboration on AI Safety*, November 2024, accessed May 6, 2026, https://www.mddi.gov.sg/newsroom/singapore-european-union-agree-to-strengthen-collaboration-ai-safety/

61 Ministry of Foreign Affairs, *Official Visit by Prime Minister and Minister for Finance Lawrence Wong to the Republic of Korea (ROK), 2 November 2025*, November 2025, accessed May 6, 2026, https://www.mfa.gov.sg/newsroom/press-statements-transcripts-and-photos/official-visit-by-prime-minister-and-minister-for-finance-lawrence-wong-to-the-republic-of-korea--rok---2-november-2025/

62 Jiyeon Cho, *LinkedIn*, January 2026, accessed May 6, 2026, https://www.linkedin.com/posts/jycho-koreaaisi_kr-sg-aisi-bilateral-testing-result-activity-7419231473037352961-u8wC/

63 Singapore AI Safety Institute, *LinkedIn*, January 2026, accessed May 6, 2026, https://www.linkedin.com/posts/singapore-ai-safety-institute_testing-ai-agents-for-data-leakage-risks-activity-7419402109282021376-VTej/

64 European Commission, *ALT-EDIC - European Language Data Space - European Commission*, accessed May 6, 2026, https://language-data-space.ec.europa.eu/related-initiatives/alt-edic_en

65 European Commission, *Joint Statement of the second meeting of the EU - Singapore Digital Partnership Council | Shaping Europe's digital future*, February 2026, accessed May 6, 2026, https://digital-strategy.ec.europa.eu/en/library/joint-statement-second-meeting-eu-singapore-digital-partnership-council

66 HTxAI, "Pre-Training Phoenix: How we built a globally competitive LLM training dataset," December 12, 2025, accessed May 10, 2026, https://medium.com/htx-ai/pre-training-phoenix-how-we-built-a-globally-competitive-llm-training-dataset-0ac2a7a21d73

67 SEA-LION, "Sahabat AI," November 20, 2025, accessed May 11, 2026, https://sea-lion.ai/case-study/sahabat-ai/

68 I2R, A*STAR, *MERaLiON/MERaLiON-3-10B-preview*, January 2025, accessed May 6, 2026, https://huggingface.co/MERaLiON/MERaLiON-3-10B-preview

69 AI Singapore, *SEA-LION API | SEA-LION Documentation*, October 2025, accessed May 6, 2026, https://docs.sea-lion.ai/guides/inferencing/api

70 AISG | AI Products, *Introducing SEA-Guard (Safety Collection): Safety-First AI, Built for Southeast Asia | SEA-LION*, Section: Announcements, February 2026, accessed May 6, 2026, https://sea-lion.ai/blog/introducing-sea-guard-safety-collection-safety-first-ai-built-for-southeast-asia/

71 Jonathan Lee et al., *State of AI Safety in Singapore (2025)*, technical report (Concordia AI, 2025), accessed May 8, 2026, https://concordia-ai.com/wp-content/uploads/2025/07/State-of-AI-Safety-in-Singapore-2025.pdf; Gabriel Chua, *Introducing LionGuard 2:*

*Multilingual LLM Guardrail for Singapor*e, July 2025, accessed May 6, 2026, https://medium.com/dsaid-govtech/lionguard-2-8066d4e20d16

72 Leanne Tan et al., "LionGuard 2: Building Lightweight, Data-Efficient & Localised Multilingual Content Moderators," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, ed. Ivan Habernal, Peter Schulam, and Jörg Tiedemann (Suzhou, China: Association for Computational Linguistics, November 2025), 264–285, accessed May 6, 2026, https://doi.org/10.18653/v1/2025.emnlp-demos.20, https://aclanthology.org/2025.emnlp-demos.20/

73 Alvin Lim, *HTX expands partnership with Mistral AI*, November 2025, accessed May 6, 2026, https://www.htx.gov.sg/whats-happening/all-news---events/all-news/2025/htx-expands-partnership-with-mistral-ai

74 AI Products Team, AI Singapore, *SEA-HELM Leaderboard*, accessed May 6, 2026, https://leaderboard.sea-lion.ai/

75 Panuthep Tasawong et al., *SEA-SafeguardBench: Evaluating AI Safety in SEA Languages and Cultures*, arXiv:2512.05501 [cs] version: 1, December 2025, accessed May 6, 2026, https://doi.org/10.48550/arXiv.2512.05501, http://arxiv.org/abs/2512.05501

76 AI Singapore, *SEA-HELM | SEA-LION Documentation*, June 2025, accessed May 6, 2026, https://docs.sea-lion.ai/benchmarks/sea-helm

77 Yosephine Susanto et al., *SEA-HELM: Southeast Asian Holistic Evaluation of Language Models*, arXiv:2502.14301 [cs], June 2025, accessed May 6, 2026, https://doi.org/10.48550/arXiv.2502.14301, http://arxiv.org/abs/2502.14301

78 GovTech, *AI Guardian: Litmus and Sentinel*, accessed May 6, 2026, https://isomer-user-content.by.gov.sg/22/6c4f97dc-3b8b-4701-8592-cd12d72012dd/20250916_Litmus%20and%20Sentinel%20one%20pager.pdf

79 Benjamin Goh, *How we built the AI Guardian team at GovTech Singapor*e, December 2025, accessed May 6, 2026, https://medium.com/aiguardian-govtech/how-we-built-the-ai-guardian-team-at-govtech-singapore-3758cf21004d

80 Infocomm Media Development Authority and AI Verify Foundation, *Starter Kit for Testing LLM-Based Applications for Safety and Reliability*, January 2026, accessed May 6, 2026, https://www.imda.gov.sg/-/media/imda/files/about/emerging-tech-and-research/artificial-intelligence/starter-kit-for-testing-llm-based-applications-for-safety-and-reliability.pdf

81 Jonathan Lee et al., *State of AI Safety in Singapore - Concordia AI*, July 2025, accessed May 6, 2026, https://concordia-ai.com/research/state-of-ai-safety-in-singapore/

82 AI Verify Foundation, *Global AI Assurance Sandbox*, 2026, accessed May 8, 2026, https://aiverifyfoundation.sg/ai-assurance/

83 AI Verify Foundation, *Global AI Assurance Sandbox*, 2026, accessed May 8, 2026, https://aiverifyfoundation.sg/ai-assurance/

84 Economic Development Board, *Google Cloud makes 'Gemini Everywhere' vision a reality, doubles down on Enterprise AI commitment to Singapor*e, August 2025, accessed May 6, 2026, https://www.edb.gov.sg/en/about-edb/media-releases-publications/google-cloud-expands-gemini-ai-strengthens-singapore-focus.html

85 Alvin Lim, *HTX expands partnership with Mistral AI*, November 2025, accessed May 6, 2026, https://www.htx.gov.sg/whats-happening/all-news---events/all-news/2025/htx-expands-partnership-with-mistral-ai

86 DSO National Laboratories, *MINDEF, DSTA, DSO Partner Mistral AI on Defence AI | DSO*, March 2025, accessed May 6, 2026, https://www.dso.org.sg/media-article/mindef-dsta-and-dso-partner-mistral-ai-to-advance-generative-ai-for-defence-applications

87 HTX Singapore, *HTX inks contract with Mistral AI and Microsoft to boost AI model development for Home Team*, May 2025, accessed May 6, 2026, https://www.htx.gov.sg/whats-happening/all-news---events/all-news/2025/media-release-htx-inks-contract-with-mistral-ai-and-microsoft-to-boost-ai-model-development-for-home-team

88 Economic Development Board, *Google DeepMind expands presence in Singapore, opening new research lab to advance AI in the Asia Pacific region*, November 2025, accessed May 6, 2026, https://www.edb.gov.sg/en/about-edb/media-releases-publications/google-deepmind-opens-new-ai-research-lab-in-singapore.html

89 "Google DeepMind sets up AI lab in S'pore, seeks to boost model's understanding of local tongues," *The Straits Times* (Singapore), November 2025, ISSN: 0585-3923, accessed May 6, 2026, https://www.straitstimes.com/tech/google-deepmind-sets-up-ai-lab-in-spore-boosts-models-understanding-of-local-tongues

90 Alibaba Cloud, *Alibaba's Qwen Powers AI Singapore's Latest LLM to Strengthen Multilingual Performance in Southeast Asia*, November 2025, accessed May 8, 2026, https://www.alibabacloud.com/en/press-room/alibaba-qwen-powers-ai-singapore-latest-llm

91 Alvin Lim, *HTX expands partnership with Mistral AI*, November 2025, accessed May 6, 2026, https://www.htx.gov.sg/whats-happening/all-news---events/all-news/2025/htx-expands-partnership-with-mistral-ai

92 Yoshua Bengio et al., *International AI Safety Report*, technical report DSIT 2026/001 (Department for Science, Innovation and Technology, 2026), accessed May 6, 2026, https://internationalaisafetyreport.org/

93 Mengyao Du et al., *SnapGuard: Lightweight prompt injection detection for screenshot-based web agents*, April 28, 2026, accessed May 11, 2026, arXiv: 2604.25562 [cs.CR], http://arxiv.org/abs/2604.25562

94 Hanyi Wang et al., *ResGuard: Enhancing robustness against known original attacks in deep watermarking*, April 4, 2026, accessed May 11, 2026, arXiv: 2604.03693 [cs.CV], http://arxiv.org/abs/2604.03693

95 Mengyao Du et al., *TrapSuffix: Proactive defense against adversarial suffixes in jailbreaking*, February 6, 2026, accessed May 11, 2026, arXiv: 2602.06630 [cs.CR], http://arxiv.org/abs/2602.06630

96 Jingnan Zheng et al., *Risky-Bench: Probing Agentic Safety Risks under Real-World Deployment*, arXiv:2602.03100 [cs], February 2026, accessed May 6, 2026, https://doi.org/10.48550/arXiv.2602.03100, http://arxiv.org/abs/2602.03100

97 Enyi Shi et al., *Lingua-SafetyBench: A Benchmark for Safety Evaluation of Multilingual Vision-Language Models*, arXiv:2601.22737 [cs], April 2026, accessed May 6, 2026, https://doi.org/10.48550/arXiv.2601.22737, http://arxiv.org/abs/2601.22737

98 Yonghui Yang et al., *Controllable Value Alignment in Large Language Models through Neuron-Level Editing*, arXiv:2602.07356 [cs], February 2026, accessed May 6, 2026, https://doi.org/10.48550/arXiv.2602.07356, http://arxiv.org/abs/2602.07356

99 Xianglin Yang et al., *Zombie Agents: Persistent Control of Self-Evolving LLM Agents via Self-Reinforcing Injections*, arXiv:2602.15654 [cs], March 2026, accessed May 6, 2026, https://doi.org/10.48550/arXiv.2602.15654, http://arxiv.org/abs/2602.15654

100 Xianglin Yang et al., *Enhancing Model Defense Against Jailbreaks with Proactive Safety Reasoning*, arXiv:2501.19180 [cs], January 2026, accessed May 6, 2026, https://doi.org/10.48550/arXiv.2501.19180, http://arxiv.org/abs/2501.19180

101 Ruofan Liu et al., *DRIP: Defending Prompt Injection via Token-wise Representation Editing and Residual Instruction Fusion*, arXiv:2511.00447 [cs], November 2025, accessed May 6, 2026, https://doi.org/10.48550/arXiv.2511.00447, http://arxiv.org/abs/2511.00447

102 Kangwei Liu et al., "LookAhead Tuning: Safer Language Models via Partial Answer Previews," in *Proceedings of the Nineteenth ACM International Conference on Web Search and Data Mining*, WSDM '26 (New York, NY, USA: Association for Computing Machinery, February 2026), 1190–1194, accessed May 6, 2026, https://doi.org/10.1145/3773966.3779370, https://dl.acm.org/doi/10.1145/3773966.3779370

103 Lyucheng Wu et al., "Automating Steering for Safe Multimodal Large Language Models," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, ed. Christos Christodoulopoulos et al. (Suzhou, China: Association for Computational Linguistics, November 2025), 792–814, accessed May 6, 2026, https://doi.org/10.18653/v1/2025.emnlp-main.41, https://aclanthology.org/2025.emnlp-main.41/

104 Yulin Chen et al., *Robustness via Referencing: Defending against Prompt Injection Attacks by Referencing the Executed Instruction*, arXiv:2504.20472 [cs], April 2026, accessed May 6, 2026, https://doi.org/10.48550/arXiv.2504.20472, http://arxiv.org/abs/2504.20472

105 Herun Wan et al., *The Facade of Truth: Uncovering and Mitigating LLM Susceptibility to Deceptive Evidence*, arXiv:2601.05478 [cs], January 2026, accessed May 6, 2026, https://doi.org/10.48550/arXiv.2601.05478, http://arxiv.org/abs/2601.05478

106 Fanxiao Li et al., *What's Left Unsaid? Detecting and Correcting Misleading Omissions in Multimodal News Previews*, arXiv:2601.05563 [cs], April 2026, accessed May 6, 2026, https://doi.org/10.48550/arXiv.2601.05563, http://arxiv.org/abs/2601.05563

107 Do Xuan Long et al., "LLMs Are Biased Towards Output Formats! Systematically Evaluating and Mitigating Output Format Bias of LLMs," in *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, ed. Luis Chiruzzo, Alan Ritter, and Lu Wang (Albuquerque, New Mexico: Association for Computational Linguistics, April 2025), 299–330, accessed May 6, 2026, https://doi.org/10.18653/v1/2025.naacl-long.15, https://aclanthology.org/2025.naacl-long.15/

108 Himanshu Singh et al., *Do Prompts Guarantee Safety? Mitigating Toxicity from LLM Generations through Subspace Intervention*, arXiv:2602.06623 [cs], February 2026, accessed May 6, 2026, https://doi.org/10.48550/arXiv.2602.06623, http://arxiv.org/abs/2602.06623

109 Yangyang Guo et al., *LLMs Can Unlearn Refusal with Only 1,000 Benign Samples*, arXiv:2601.19231 [cs], January 2026, accessed May 6, 2026, https://doi.org/10.48550/arXiv.2601.19231, http://arxiv.org/abs/2601.19231

110 Ziwei Xu and Mohan Kankanhalli, *Strong preferences affect the robustness of preference models and value alignment*, October 3, 2024, accessed May 11, 2026, arXiv: 2410.02451 [cs.AI], http://arxiv.org/abs/2410.02451

111 Thanh Q. Tran et al., *BarrierSteer: LLM Safety via Learning Barrier Steering*, arXiv:2602.20102 [cs], February 2026, accessed May 6, 2026, https://doi.org/10.48550/arXiv.2602.20102, http://arxiv.org/abs/2602.20102

112 Zhi Xuan Khoo et al., *Position: Balance Human Agency & AI Assistance in the Tussle for the Right to Determine*, February 2026, accessed May 8, 2026, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6185399

113 Xinyang Lu et al., *WaterDrum: Watermarking for Data-centric Unlearning Metric*, arXiv:2505.05064 [cs], February 2026, accessed May 6, 2026, https://doi.org/10.48550/arXiv.2505.05064, http://arxiv.org/abs/2505.05064

114 Sydney Levine et al., *Resource Rational Contractualism Should Guide AI Alignment*, arXiv:2506.17434 [cs], March 2026, accessed May 6, 2026, https://doi.org/10.48550/arXiv.2506.17434, http://arxiv.org/abs/2506.17434

115 Joe Edelman et al., *Full-Stack Alignment: Co-Aligning AI and Institutions with Thick Models of Value*, arXiv:2512.03399 [cs], December 2025, accessed May 6, 2026, https://doi.org/10.48550/arXiv.2512.03399, http://arxiv.org/abs/2512.03399

116 Tan Zhi-Xuan et al., "Beyond Preferences in AI Alignment," *Philosophical Studies* 182, no. 7 (July 2025): 1813–1863, ISSN: 1573-0883, accessed May 6, 2026, https://doi.org/10.1007/s11098-024-02249-w, https://doi.org/10.1007/s11098-024-02249-w

117 Zherui Li et al., *DiffuGuard: How Intrinsic Safety is Lost and Found in Diffusion Large Language Models*, arXiv:2509.24296 [cs], March 2026, accessed May 6, 2026, https://doi.org/10.48550/arXiv.2509.24296, http://arxiv.org/abs/2509.24296

118 Zhenhao Zhu et al., "ExtendAttack: Attacking Servers of LRMs via Extending Reasoning," *Proceedings of the AAAI Conference on Artificial Intelligence* 40, no. 41 (March 2026): 35257–35265, ISSN: 2374-3468, accessed May 6, 2026, https://doi.org/10.1609/aaai.v40i41.40833, https://ojs.aaai.org/index.php/AAAI/article/view/40833

119 Linyu Wu et al., *DMark: Order-Agnostic Watermarking for Diffusion Large Language Models*, arXiv:2510.02902 [cs], October 2025, accessed May 6, 2026, https://doi.org/10.48550/arXiv.2510.02902, http://arxiv.org/abs/2510.02902

120 Nanyang Technological University, "About the Cluster," accessed May 11, 2026, https://www.ntu.edu.sg/research/research-focus/research-cluster-1-artificalandaugumentedintelligence/about-the-cluster

121 Simin Li et al., *Empirical Study on Robustness and Resilience in Cooperative Multi-Agent Reinforcement Learning*, arXiv:2510.11824 [cs], October 2025, accessed May 6, 2026, https://doi.org/10.48550/arXiv.2510.11824, http://arxiv.org/abs/2510.11824

122 Simin Li et al., *Vulnerable Agent Identification in Large-Scale Multi-Agent Reinforcement Learning*, arXiv:2509.15103 [cs], September 2025, accessed May 6, 2026, https://doi.org/10.48550/arXiv.2509.15103, http://arxiv.org/abs/2509.15103

123 Miao Yu et al., "A Survey on Trustworthy LLM Agents: Threats and Countermeasures," in *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25 (New York, NY, USA: Association for Computing Machinery, August 2025), 6216–6226, accessed May 6, 2026, https://doi.org/10.1145/3711896.3736561, https://dl.acm.org/doi/10.1145/3711896.3736561

124 Kwok-Yan Lam, *Prof Lam Kwok Yan*, Nanyang Technological University, 2026, accessed May 8, 2026, https://dr.ntu.edu.sg/cris/rp/rp00321

125 Chen Chen et al., *Plato's Form: Toward Backdoor Defense-as-a-Service for LLMs with Prototype Representations*, arXiv:2602.06887 [cs], February 2026, accessed May 6, 2026, https://doi.org/10.48550/arXiv.2602.06887, http://arxiv.org/abs/2602.06887

126 Mingrui Liu et al., *RedVisor: Reasoning-Aware Prompt Injection Defense via Zero-Copy KV Cache Reuse*, arXiv:2602.01795 [cs], February 2026, accessed May 6, 2026, https://doi.org/10.48550/arXiv.2602.01795, http://arxiv.org/abs/2602.01795

127 Chen Chen et al., *The Shadow Self: Intrinsic Value Misalignment in Large Language Model Agents*, arXiv:2601.17344 [cs], January 2026, accessed May 6, 2026, https://doi.org/10.48550/arXiv.2601.17344, http://arxiv.org/abs/2601.17344

128 Xiaojun Jia et al., "Semantic-Aligned Adversarial Evolution Triangle for High-Transferability Vision-Language Attack," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 47, no. 10 (October 2025): 8489–8505, ISSN: 1939-3539, accessed May 6, 2026, https://doi.org/10.1109/TPAMI.2025.3581476, https://ieeexplore.ieee.org/abstract/document/11045302

129 Xiaojun Jia et al., *Evolution-based Region Adversarial Prompt Learning for Robustness Enhancement in Vision-Language Models*, arXiv:2503.12874 [cs], March 2025, accessed May 6, 2026, https://doi.org/10.48550/arXiv.2503.12874, http://arxiv.org/abs/2503.12874

130 Zhenhong Zhou et al., *CORBA: Contagious Recursive Blocking Attacks on Multi-Agent Systems Based on Large Language Models*, arXiv:2502.14529 [cs], February 2025, accessed May 6, 2026, https://doi.org/10.48550/arXiv.2502.14529, http://arxiv.org/abs/2502.14529

131 Shuai Zhao et al., "UniFLE: Uniform Fusion of Multiple LoRA Experts for Backdoor Defense in Large Language Models," *IEEE Transactions on Dependable and Secure Computing*, February 2026, 1–15, ISSN: 1941-0018, accessed May 6, 2026, https://doi.org/10.1109/TDSC.2026.3662097, https://ieeexplore.ieee.org/abstract/document/11407462

132 Shuai Zhao et al., *P2P: A Poison-to-Poison Remedy for Reliable Backdoor Defense in LLMs*, arXiv:2510.04503 [cs], October 2025, accessed May 6, 2026, https://doi.org/10.48550/arXiv.2510.04503, http://arxiv.org/abs/2510.04503

133 Man Hu et al., *Rethinking Reasoning: A Survey on Reasoning-based Backdoors in LLMs*, arXiv:2510.07697 [cs], October 2025, accessed May 6, 2026, https://doi.org/10.48550/arXiv.2510.07697, http://arxiv.org/abs/2510.07697

134 Ankit Kanwar, Dominik Wagner, and Luke Ong, *SB-TRPO: Towards Safe Reinforcement Learning with Hard Constraints*, arXiv:2512.23770 [cs], January 2026, accessed May 6, 2026, https://doi.org/10.48550/arXiv.2512.23770, http://arxiv.org/abs/2512.23770

135 Dominik Wagner, Leon Witzman, and Luke Ong, *Reinforcement Learning with ω-Regular Objectives and Constraints*, arXiv:2511.19849 [cs], November 2025, accessed May 6, 2026, https://doi.org/10.48550/arXiv.2511.19849, http://arxiv.org/abs/2511.19849

136 Fazl Barez et al., *Open Problems in Machine Unlearning for AI Safety*, arXiv:2501.04952 [cs], January 2025, accessed May 6, 2026, https://doi.org/10.48550/arXiv.2501.04952, http://arxiv.org/abs/2501.04952

137 Usman Naseem et al., "LLM Alignment should go beyond Harmlessness−Helpfulness and incorporate Human Agency," *Cognitive Computation* 18, no. 1 (March 2026): 26, ISSN: 1866-9964, accessed May 6, 2026, https://doi.org/10.1007/s12559-026-10568-9, https://doi.org/10.1007/s12559-026-10568-9

138 Jingdi Lei et al., *OffTopicEval: When Large Language Models Enter the Wrong Chat, Almost Always!*, arXiv:2509.26495 [cs], March 2026, accessed May 6, 2026, https://doi.org/10.48550/arXiv.2509.26495, http://arxiv.org/abs/2509.26495

139 Maojia Song et al., *Measuring and Enhancing Trustworthiness of LLMs in RAG through Grounded Attributions and Learning to Refuse*, arXiv:2409.11242 [cs], April 2025, accessed May 6, 2026, https://doi.org/10.48550/arXiv.2409.11242, http://arxiv.org/abs/2409.11242

140 Shuhan Xu et al., "SRD: Reinforcement-Learned Semantic Perturbation for Backdoor Defense in VLMs," *Proceedings of the AAAI Conference on Artificial Intelligence* 40, no. 14 (March 2026): 11397–11405, ISSN: 2374-3468, accessed May 6, 2026, https://doi.org/10.1609/aaai.v40i14.38121, https://ojs.aaai.org/index.php/AAAI/article/view/38121

141 Songze Li et al., *Odysseus: Jailbreaking Commercial Multimodal LLM-integrated Systems via Dual Steganography*, arXiv:2512.20168 [cs], December 2025, accessed May 6, 2026, https://doi.org/10.48550/arXiv.2512.20168, http://arxiv.org/abs/2512.20168

142 Zonghao Ying et al., *SafeBench: A Safety Evaluation Framework for Multimodal Large Language Models*, arXiv:2410.18927 [cs], October 2024, accessed May 6, 2026, https://doi.org/10.48550/arXiv.2410.18927, http://arxiv.org/abs/2410.18927

143 Xinwei Zhang et al., *On the Adversarial Robustness of Large Vision-Language Models under Visual Token Compression*, arXiv:2601.21531 [cs], January 2026, accessed May 6, 2026, https://doi.org/10.48550/arXiv.2601.21531, http://arxiv.org/abs/2601.21531

144 Renyang Liu et al., *SafeRedir: Prompt Embedding Redirection for Robust Unlearning in Image Generation Models*, arXiv:2601.08623 [cs], January 2026, accessed May 6, 2026, https://doi.org/10.48550/arXiv.2601.08623, http://arxiv.org/abs/2601.08623

145 Guanlin Li et al., *Picky LLMs and Unreliable RMs: An Empirical Study on Safety Alignment after Instruction Tuning*, arXiv:2502.01116 [cs], February 2025, accessed May 6, 2026, https://doi.org/10.48550/arXiv.2502.01116, http://arxiv.org/abs/2502.01116

146 Singapore Management University, "Creating our Digital Future," accessed May 11, 2026, https://computing.smu.edu.sg/research

147 Haoyu Wang et al., *ProbGuard: Probabilistic Runtime Monitoring for LLM Agent Safety*, arXiv:2508.00500 [cs], March 2026, accessed May 6, 2026, https://doi.org/10.48550/arXiv.2508.00500, http://arxiv.org/abs/2508.00500

148 Zibo Xiao, Jun Sun, and Junjie Chen, *AIR: Improving Agent Safety through Incident Response*, arXiv:2602.11749 [cs], February 2026, accessed May 6, 2026, https://doi.org/10.48550/arXiv.2602.11749, http://arxiv.org/abs/2602.11749

149 Haoyu Wang, Christopher M. Poskitt, and Jun Sun, *AgentSpec: Customizable Runtime Enforcement for Safe and Reliable LLM Agents*, arXiv:2503.18666 [cs], July 2025, accessed May 6, 2026, https://doi.org/10.48550/arXiv.2503.18666, http://arxiv.org/abs/2503.18666

150 Gabriel Chua et al., *Lost in Localization: Building RabakBench with Human-in-the-Loop Validation to Measure Multilingual Safety Gaps*, arXiv:2507.05980 [cs], February 2026, accessed May 6, 2026, https://doi.org/10.48550/arXiv.2507.05980, http://arxiv.org/abs/2507.05980

151 Bryan Chen Zhengyu Tan et al., "Persuasion Dynamics in LLMs: Investigating Robustness and Adaptability in Knowledge and Safety with DuET-PD," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, ed. Christos Christodoulopoulos et al. (Suzhou, China: Association for Computational Linguistics, November 2025), 1550–1575, accessed May 6, 2026, https://doi.org/10.18653/v1/2025.emnlp-main.81, https://aclanthology.org/2025.emnlp-main.81/

152 Yujia Hu et al., "Toxicity Red-Teaming: Benchmarking LLM Safety in Singapore's Low-Resource Languages," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, ed. Christos Christodoulopoulos et al. (Suzhou, China: Association for Computational Linguistics, November 2025), 12183–12201, accessed May 6, 2026, https://doi.org/10.18653/v1/2025.emnlp-main.612, https://aclanthology.org/2025.emnlp-main.612/

153 Centre for Frontier AI Research (CFAR), "Research Pillars," accessed May 11, 2026, https://www.a-star.edu.sg/cfar/research/research-pillars

154 Jieyu Li, Xin Zhang, and Joey Tianyi Zhou, "AEGIS: Authenticity Evaluation Benchmark for AI-Generated Video Sequences," in *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25 (New York, NY, USA: Association for Computing Machinery, October 2025), 13346–13353, accessed May 6, 2026, https://doi.org/10.1145/3746027.3758295, https://dl.acm.org/doi/10.1145/3746027.3758295

155 Zhifang Zhang et al., *TokenSwap: Backdoor Attack on the Compositional Understanding of Large Vision-Language Models*, arXiv:2509.24566 [cs], September 2025, accessed May 6, 2026, https://doi.org/10.48550/arXiv.2509.24566, http://arxiv.org/abs/2509.24566

156 Jiaxu Zhao et al., "Understanding Large Language Model Vulnerabilities to Social Bias Attacks," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ed. Wanxiang Che et al. (Vienna, Austria: Association for Computational Linguistics, July 2025), 17620–17636, accessed May 6, 2026, https://doi.org/10.18653/v1/2025.acl-long.862, https://aclanthology.org/2025.acl-long.862/

157 Yanghao Su et al., *Character as a Latent Variable in Large Language Models: A Mechanistic Account of Emergent Misalignment and Conditional Safety Failures*, arXiv:2601.23081 [cs], January 2026, accessed May 6, 2026, https://doi.org/10.48550/arXiv.2601.23081, http://arxiv.org/abs/2601.23081

158 Jiyan He et al., "Controlling risks of AI in chemical science with agents," *AI for Science* 1 (1 2025): 015002, ISSN: 3050-287X, accessed May 9, 2026, https://doi.org/10.1088/3050-287x/adfee5, http://dx.doi.org/10.1088/3050-287X/adfee5

159 Xiaoyu Zhang et al., "JailGuard: A Universal Detection Framework for Prompt-based Attacks on LLM Systems," *ACM Trans. Softw. Eng. Methodol.* 35, no. 1 (December 2025): 8:1–8:40, ISSN: 1049-331X, accessed May 6, 2026, https://doi.org/10.1145/3724393, https://dl.acm.org/doi/10.1145/3724393

160 AI Safety Institute (AISI), *AI Safety Institute (AISI)*, accessed May 6, 2026, https://www.ntu.edu.sg/dtc/aisi

161 GovTech Singapore, "Our Digital Government efforts," accessed May 11, 2026, https://www.tech.gov.sg/about-us/what-we-do/our-digital-government-efforts/

162 AI Safety Institute (AISI), *AI Safety Institute (AISI)*, accessed May 6, 2026, https://www.ntu.edu.sg/dtc/aisi

163 Shaun Khoo, Jessica Foo, and Roy Ka-Wei Lee, *With Great Capabilities Come Great Responsibilities: Introducing the Agentic Risk & Capability Framework for Governing Agentic AI Systems*, arXiv:2512.22211 [cs], December 2025, accessed May 6, 2026, https://doi.org/10.48550/arXiv.2512.22211, http://arxiv.org/abs/2512.22211

164 Jia Yi Goh et al., *Measuring What Matters: A Framework for Evaluating Safety Risks in Real-World LLM Applications*, arXiv:2507.09820 [cs], July 2025, accessed May 6, 2026, https://doi.org/10.48550/arXiv.2507.09820, http://arxiv.org/abs/2507.09820

165 Isaac Lim et al., *Safe at the Margins: A General Approach to Safety Alignment in Low-Resource English Languages – A Singlish Case Study*, arXiv:2502.12485 [cs], April 2025, accessed May 6, 2026, https://doi.org/10.48550/arXiv.2502.12485, http://arxiv.org/abs/2502.12485

166 Ministry of Digital Development and Information, "Singapore Invests Over S$1 billion in National AI Research and Development Plan to Strengthen AI Research Capabilities and Our Position as Global AI Hub," accessed May 11, 2026, https://www.mddi.gov.sg/newsroom/singapore-invests-over-s-1-billion-in-national-ai-research-and-development-plan-to-strengthen-ai-research-capabilities-and-our-position-as-global-ai-hub/

167 EDB Singapore, "Singapore to invest S$1 billion over five years to boost AI public research," accessed May 11, 2026, https://www.edb.gov.sg/en/business-insights/insights/singapore-to-invest-s1-billion-over-five-years-to-boost-ai-public-research.html

168 Ministry of Digital Development and Information, *Press Release - Singapore AI Research Week (24 Jan).pdf*, accessed May 6, 2026, https://isomer-user-content.by.gov.sg/38/67dfcc81-25bb-466c-bf80-1cbb680aaff8/Annex%20A.pdf

169 Leheng Sheng et al., *Reinforcing Chain-of-Thought Reasoning with Self-Evolving Rubrics*, arXiv:2602.10885 [cs], February 2026, accessed May 6, 2026, https://doi.org/10.48550/arXiv.2602.10885, http://arxiv.org/abs/2602.10885