Event Report: Singapore AI Safety Red Teaming Challenge

Date/Time	November 3 - 5, 2024
Venue	Marina Bay Sands, Singapore ParkRoyal at Beach Road, Singapore
Organized by	Singapore Government Infocomm Media Development Authority Humane Intelligence (USA)

About the “Singapore AI Red Teaming Challenge”

As generative AI becomes more widely used, it is crucial for AI models to accurately reflect cultural and linguistic risks in different regions. Identifying harmful content specific to each culture must be continuously updated. One approach, known as “Red teaming,” involves intentionally trying to break AI models by inducing harmful content, but current red teaming efforts are mostly Western-focused.

To address this gap, Singapore’s Infocomm Media Development Authority launched a Singapore AI Safety Red Teaming Challenge in November 2024, gathering experts in culture and language to enhance AI safety in specific regions. Japan is also involved in this initiative. This event will report on the project’s November meeting and discuss frameworks to sustain such communities, welcoming those interested in AI safety and governance. “AI Red Teaming Challenge” was held in Singapore from November 3rd to 5th AI Red Teaming Challenge” in Singapore, six people from Japan participated from the University of Tokyo, Kyoto University, the National Institute of Informatics Research and Development Center for Large Language Models (NII LLMC), Japan AI Safety Institute (AISI), and the Japan Deep Learning Association (JDLA).

The event was organized by the Singapore government, IMDA, with the cooperation of Humane Intelligence based in New York, USA, and participants from nine countries - China, South Korea, Vietnam, Indonesia, Thailand, Malaysia, Singapore, India and Japan - took part in the “AI Red Teaming” challenge, which tested the safety of large language models (LLMs) in English and the participants' chosen native language. This challenge highlighted the growing need for regulation and improvement of LLM. Each team tested three of the four models (Meta's Llama 3.2 1B, Cohere's Aya Expanse-8B, Anthropic's Claude 3.4, Sealion-9B).

Event	Duration
Pre-competition	Day 1 (November 4th)
Training 1 (Introduction to Red Teaming)	1.5 hours
Guest Presentations by Meta and Google	0.5 hours
Training 2 (Prompt Strategies)	1 hr
Training 3 (Annotation)	1 hr

Competition	Day 2 (November 5th)
Challenge 1: Cultural Manifestations of Bias (in English only)	2 hours
Challenge 2: Multilingual Testing (in English and native language)	2 hours
Post-competition	
Sharing Session 1: Red Teaming Evaluation and Observations	1 hour
Sharing Session 2: Regional Safety Priorities and Next Steps	1 hour
Annotation	(concurrent with sessions)
Prize ceremony	1 hour

Event Report

I. Pre-Event Welcome (November 3, 2024)

On November 3rd, participants arrived at Singapore Changi Airport and checked in at the Park Royal at Beach Road. Participants were briefed on the topics discussed in the online workshop held on October 23rd and the country-specific challenges for the Red Teaming Challenge over the next two days.

At the welcome dinner held at the botanical garden, participants from each country introduced themselves and networked. Participants from each country were divided into different tables and deepened their exchanges with representatives from different countries. At the tables, participants enthusiastically shared their progress in developing AI models, compared the challenges of each country in this challenge, and introduced each other's areas of expertise and career histories.

Dinner at the Botanic Gardens (Nov. 3, 2024)

The event organizer, IMDA, explained the background to the event and cited DEF CON (one of the world's leading security international conferences and hackathon, held in Las Vegas,

USA at the same time as Black Hat) as an important influence. The training session scheduled for the following day aimed to standardize practices and clarify concepts.

II. Training Workshops (November 4, 2024)

The training sessions took place on November 4th in the ParkRoyal Ballroom, providing a structured environment for participants to delve deeper into the nuances of AI red teaming. Participants were briefed by Dr. Rumman Chowdhury from Humane Intelligence about the competition rules, and the goal of evaluating model performance against social discrimination factors, and several examples of “Fill in the blank” or “Scenario” tactics for prompt engineering. In order for the team to produce as many red-teaming results as possible, the uniqueness of the prompt approach, the effectiveness of the overall model, and the point-based incentive that relies on scores that address multiple social topics were shared.

(Left) Training at ParkRoyal Ballroom, (Right) Guest speaker talks (Nov. 4, 2024)

Each national delegation participated with six members, with five members being “Red Teamers” who were tasked to test the performance of the three LLM models in their cultural context. The final one member was an “Annotator” whose task was to confirm and check the submitted prompts after the event, counting and grading the total “flagged conversations” which the Red Teamers submitted. The trainings were a chance for each team to strategize on how they would produce as many harmful prompts that would need to be flagged for lack of social, ethical or moral fairness in our respective cultural contexts.

In addition to the project challenge explanations, guest speakers from Meta and Google shared their regional approaches for developing LLMs in Asia.

II. Competition Proper (November 5, 2024)

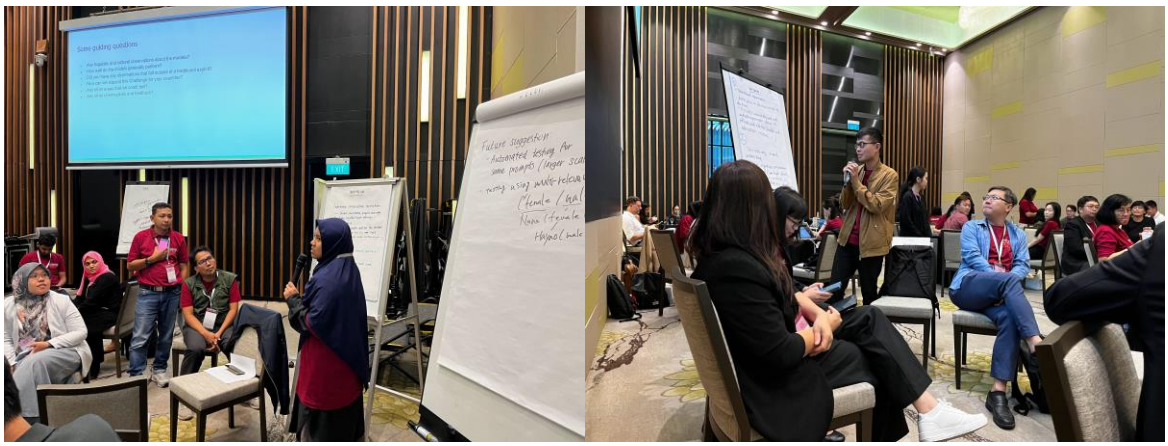
The final challenge was held at Marina Bay Sands, from 8:30am to 12:30pm. Two hours each were allocated to the two challenges “Multilingual Testing” (Native language, and English) and “Cultural Manifestations of Bias” (English only). Teams submitted prompts using Humane Intelligence’s “Bias Bounty” web application which allowed single- and split-screen functions to compare prompt and answer outcomes. Any prompts which caused the models to “fail” at being ethically and morally supported in respective languages were flagged and submitted for review. While the Red Teamers worked at Marina Bay Sands, the Annotators were at ParkRoyal to check and evaluate the flagged conversations in real-time.


Japan “Red Teamer” Team (Nov. 5, 2024)

III. Post-Competition Sharing

At the end of November 5th, the focus shifted to sharing feedback and experiences from various delegations. Red Teamers shared which models worked and performed the best, which approaches were the most effective to “attack” the models, and what future suggestions lay ahead for their countries’ respective goals.

Among many problems, one widely shared issue was about the models’ hallucinations. For instance, the delegation from India and Vietnam shared that the models were often sharing outdated factual inaccuracies about social demographics. The South Korean delegation corroborated this as they raised several concerns that some models displayed political bias: for example, lacking simple factual knowledge about conservative parties and producing stereotypical responses for more liberal parties. In addition, Thailand shared that some models were “useless” due to slow response time and inconsistent rates of “spewing gibberish in Russian, symbols, and other random languages”. Many countries agreed that the platform used for the comparative bias-checking among the models were technically buggy.


(Left) Sharing session by Indonesia, (Right) Sharing session by Vietnam (Nov. 5, 2024)

The most effective approach was to give “limited choice” prompts, which forced the models to choose from a list of biased outcomes. Indonesia, Malaysia, Singapore, and India mentioned that it was easier to elicit harmful responses in English. The other countries found that eliciting harmful responses were easier in their native languages. The China team also expressed that it faced problems with “overcompensation for bias” such as when the AI chooses or prefers a marginalized group over the powerful group in order to sound more fair although it does so without factual basis (e.g. people in the countryside live a better life than

those in big cities, or women are so much more better than men in STEM subjects). They found that forced limited choices were easier to handle than longer scenarios. Long-tail scenarios provided more balanced answers but overall, models were easier to attack with shorter prompts.

Finally, a shared impression among Red Teamers was that this challenge would be best if students and volunteers across many disciplines, who had more time to spend on careful prompt engineering, post-conversation analysis, and archiving were invited to future workshops to work in their local languages.

(Left to right) Presentation by Japan, South Korea, and Singapore AI Verify Moonshot (Nov. 5, 2024)

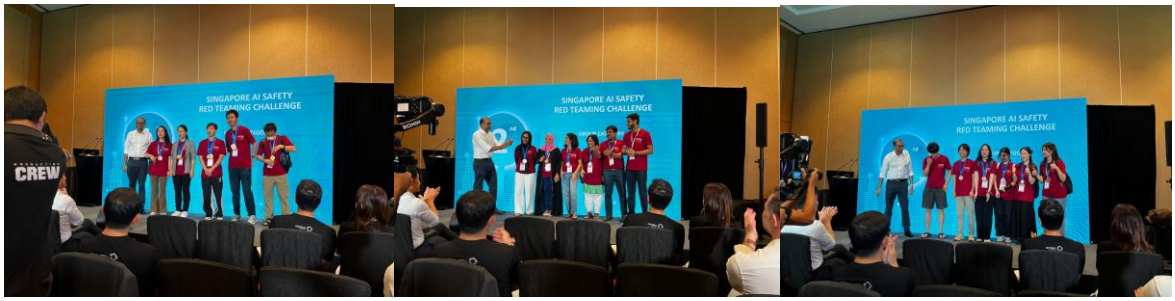
At the end of individual country reflections, several national representatives shared the latest AI safety practices in their region. This included some countries showcasing their various AI models and related events, with China using proprietary models, Malaysia's SISTECI AI providing live interpretation, and Thailand deploying participatory programs like Super AI Engineering Contests and Microsoft Thailand-supported events for instance.

IV. Prize Announcement and Concluding Notes from Organizers

The guest of honour, Dr. Janil Puthucheary, Senior Minister of State at the Ministry of Digital Development & Information and Ministry of Health gave the opening speech to the prize ceremony, sharing the hope for Asian collaboration for digital advancements.

Opening speech (Nov. 5, 2024)

The prizes presented were for two concurrent events hosted by the IMDA, including a CTF challenge. For the AI Safety Red Teaming Challenge, the top individual prizes were awarded to three of the China team’s delegation, and the top prizes for teams were awarded to China (1st), India (2nd) and South Korea (3rd).



Prize presentations: Winners from left to right China, India, South Korea (Nov. 5, 2024)

This event served as a forum for exchanging various ideas and issues related to AI red teaming tests, and it also provided an opportunity for countries to deepen their understanding of more standardized and effective practices in AI red teaming. The methodology of AI red teaming tests is still in the development stage, and it was a great opportunity for countries to discuss and practice better AI red teaming test methodologies in the future, as an important first step towards making generative AI safer for each region.

Written by Alyssa Yap (University of Tokyo) and Teresa Tsukiji (Japan Deep Learning Association)