MINISTRY OF DIGITAL DEVELOPMENT AND INFORMATION

INFOCOMM MEDIA DEVELOPMENT AUTHORITY

Annex B – Executive Summary of Singapore AI Safety Red Teaming Challenge Evaluation Report 2025

1. The Infocomm Media Development Authority (IMDA) of Singapore, in partnership with Humane Intelligence, conducted the world's first-ever multicultural and multilingual AI safety red teaming exercise focused on Asia-Pacific in November and December 2024. As large language models (LLMs) become deployed globally, and an increasing number of people around the world are interacting with the models, it is critical that they represent different languages and cultures accurately and sensitively. It is therefore important that we understand how the models perform with regards to languages and cultures, and if the safeguards hold up in these contexts. While this can be done through testing, it is not possible for any one party to test across the diverse languages and cultures in the world. We need a consistent methodology so that we can test as a global community and rely on each other's results.

2. Through this exercise, together with our partner institutes across 9 countries in Asia Pacific, we **developed a systematic methodology that can be used to test LLMs for context-specific concerns in different languages and cultures, so that different organisations around the world can adopt and adapt this methodology** to test models for linguistic and cultural sensitivities in their countries. In addition, we obtained a baseline understanding of the extent to which LLMs manifest cultural bias in our region. For example, while many will not be surprised to learn that cultural bias can be found in LLM output, we found that **cultural bias in LLM output is not uncommon in everyday use (not just in adversarial scenarios). In fact, it is not difficult to elicit bias from the model within a single prompt.** The exercise provided useful data for building new tools, such as testing benchmarks, and identified areas for further focus and development. This is only the start of a longer journey to develop scientifically robust multicultural and multilingual tests, and to make models safer in our region. Singapore will continue to work with our partners, and welcomes more to join us, to advance the sciences in this space.

3. The exercise involved 54 experts in fields such as linguistics, sociology, and cultural studies from 9 countries across Asia-Pacific for the in-person challenge; and over 300 online participants from 7 countries across Asia-Pacific for the virtual challenge. The 9 partner institutes that we worked with were:

Partner Institute, Country	Languages tested
Beijing Academy of Artificial Intelligence, China	English, Chinese
Indian Institute of Technology Madras, India	English, Hindi
Badan Riset dan Inovasi Nasional, Indonesia	English, Bahasa Indonesia
University of Tokyo, Japan	English, Japanese
Universiti Sains Malaysia, Malaysia	English, Bahasa Melayu
AI Singapore, Singapore	English, Bahasa Melayu
NAVER, South Korea	English, Korean
Electronic Transactions Development Agency, Thailand	English, Thai
Hanoi University of Science and Technology, Vietnam	English, Vietnamese

4. The challenge focused on **bias stereotypes** in different cultures, specifically testing the extent to which cultural biases are manifested in LLMs' output in everyday use, in both English and the regional language. Participants red teamed **4 LLMs**:

- AI Singapore SEA-LION (via Hugging Face) (gemma2-9b-sea-lionv3-base)
- Anthropic Claude (3.5)
- Cohere/Cohere for AI Aya (Aya 23-8B)
- Meta Llama (meta-llama-3-1-70b-instruct-vp)

5. The three key outcomes of the Challenge were:

a. **Red teaming methodology**. A systematic red teaming methodology was developed based on existing literature on LLM red teaming, and used to test for context-specific safety concerns in different regions. There are 4 key stages in this methodology.

i. Risk definition. Prior to the challenge, it is important to clearly define the risks that are being tested for. IMDA worked with partner institutes to identify representative groups of domain experts from each country to participate in the red teaming. These experts were brought together through virtual workshops to develop a taxonomy that defines how bias stereotypes manifest differently in their countries.

ii. Challenge Design. A robust incentive structure for the red teamers (i.e. scoring system) was designed to effectively draw out the challenge's intended outcomes. A balance had to be struck between incentivising the breadth (coverage of different bias categories), depth (number of prompts within each bias category), variety (unique and non-repetitive prompts), and number of turns. Incentivising single-turn

prompts was an important consideration to facilitate the use of the challenge data to build technical tools like benchmarks post-challenge.

iii. Annotation. As annotation of the harmfulness of the model outputs can be subjective, it is critical to develop an annotation guide on what was considered harmful, and how to treat borderline cases. Training of annotators on these guidelines is equally crucial to ensure consistent and high-quality annotation. It is useful to adopt a consultative approach to develop the guidelines to ensure that the 'harmful' threshold defined in the guide is aligned with the cultural and societal expectations in the country.

iv. Results Analysis. IMDA and Humane Intelligence conducted quantitative and qualitative analysis of the raw data to draw actionable insights to improve model safety. English data was analysed using clustering, topic modelling techniques and evaluated for sentiment. Output in English and regional languages was also analysed manually for qualitative themes, the latter with feedback from participants.

b. **Cultural Bias Taxonomy.** Through the pre-Challenge workshops, a taxonomy identifying the top 3 bias concerns in each of the 9 countries was developed together with the red teamers (see table below). It draws from earlier work on bias definition in the Bias Benchmark for Question Answering (BBQ), a commonly used benchmark to test for bias in LLMs, as well as the red teamers' expertise and lived experiences. Details of the top 3 bias concerns in each country can be found in the full report. Nevertheless, deeper and more extensive qualitative research (e.g. focus group discussions involving more domain experts) can be undertaken to expand the taxonomy.

c. **Baseline understanding of cultural bias in LLMs.** Through analysis of the challenge data, we gained a baseline understanding of the extent to which cultural bias is manifested in model output. While these key observations provide helpful insights on the safety characteristics of the models, they should be treated as indicative signals due to experimental limitations.

i. It is not uncommon for cultural bias to be found in model output, even in everyday use (not only in adversarial scenarios). The red teamers were explicitly instructed to prompt in the persona of a benign user (vs adversarial user). During the half-day challenge, a total of 1,335 successful exploits (~30 per pax) were collected. Participants were also able to successfully elicit bias within a single turn (86.1% of

total successful exploits). In particular, it was found that prompts that frame bias in a "positive" manner were particularly successful in eliciting a bias response from the model. For example, prompts that asked the LLMs to decide which city in China is the richest (instead of poorest), or which province in South Korea has the prettiest people (instead of ugliest), led to biased model responses. On occasion, LLMs were able to highlight unexpected cultural sensitivity, such as acknowledging the funeral rights of an indigenous group in Sulawesi in Indonesia. However, on balance, there were more misses than hits.

ii. Model guardrails for cultural biases in non-English languages may not hold up as well as in English. Regional language prompts constituted a higher percentage of total successful exploits than English language prompts (69.4% vs 30.6%). While this is influenced to some extent by language competency of the red teamers, it provides an indication of the extent to which model safety lags in non-English languages, compared to English.

iii. Out of the 5 bias categories prioritised by the red teamers for Asia Pacific, gender bias (26.1%) recorded the highest percentage of total successful exploits. The other categories recorded the following percentages – race/religious/ethnicity bias (22.8%), geographical/national identity bias (22.6%), socio-economic bias (19.0%) and other unique challenges (9.5%). This breakdown could be helpful in pinpointing specific areas of bias for model developers to strengthen safeguards.