イベントレポート：シンガポール AI セーフティレッドチーミングチャレンジ

日時	2024 年 11 月 3 日 - 5 日
会場	マリーナ・ベイ・サンズ、シンガポール パークロイヤル アット ビーチロード、シンガポール
主催	シンガポール政府 Infocomm Media Development Authority 米 Humane Intelligence

“シンガポール AI セーフティレッドチーミングチャレンジ "について

生成人工知能（AI）の利用が世界的に広まるにつれ、AI モデルが地域ごとの文化や言語におけるリスクや懸念を敏感に反映できることがますます重要になっています。そのためには、何がリスクや有害なコンテンツなのかを地域・文化ごとに特定する作業を更新し続けていくことが必要となります。これに対する 1 つのアプローチとして展開されている「レッドチーミング」は、生成 AI を評価するため、安全性に違反するような暴力を煽る有害コンテンツ、攻撃的な言葉、汚い言葉などをあえて誘発し、モデルを「壊そう」とするものです。しかし、現在の AI の安全性に関するレッドチーミングは、欧米中心であり、地域的な被害（モデルが特定の民族を差別するリスクなど）に対処する方法を考える必要があります。

このような問題意識のもと 2024 年 11 月、シンガポール政府 Infocomm Media Development Authority（以下 IMDA）が「AI セーフティレッドチームチャレンジプロジェクト」を開始しました。文化や言語の専門家のコミュニティを集め、AI モデルをレッドチーム化することを目的として組成しており、生成 AI を地域にとってより安全なものにするための重要な第一歩と捉え、同プロジェクトには日本も協力しています。11 月 3 日 - 5 日に開催されたシンガポールでの「AI レッドチーミング・チャレンジ」に、日本から東京大学、京都大学、国立情報学研究所大規模言語モデル研究開発センター（NII LLMC）、日本 AI セーフティ研究所（AISI）、日本ディープラーニング協会（JDLA）から 6 名が参加しました。

このイベントはシンガポール政府 IMDA が主催し、アメリカニューヨークを拠点とする Humane Intelligence の協力のもと、中国、韓国、ベトナム、インドネシア、タイ、マレーシア、シンガポール、インド、日本の 9 カ国から参加した人たちが英語と選択した母国語で大規模言語モデル（LLM）の安全性のテスト「AI レッドチーミング」に挑戦しました。このチャレンジでは、LLM の規制と改善の必要性が高まっていることが強調され、各チームは 4 つのモデル（Meta の Llama 3.2 1B、Cohere の Aya Expanse-8B、Anthropic の Claude 3.4、Sealion-9B）のうち 3 つのモデルをテストしました。

イベント	期間
プレ・コンペティション	1 日目（11 月 4 日）
トレーニング 1（レッドチーミング入門）	1.5 時間

Meta と Google によるゲスト・プレゼンテーション	0.5 時間
トレーニング 2（プロンプト・ストラテジー）	1 時間
トレーニング 3（アノテーション）	1 時間
コンペティション	2 日目（11 月 5 日）
課題 1：偏見の文化的表出（英語のみ）	2 時間
課題 2：多言語テスト（英語と母国語）	2 時間
コンペティション終了後	
シェアリング・セッション 1：レッドチーミングの評価と観察	1 時間
シェアリング・セッション 2：地域安全優先事項と次のステップ	1 時間
アノテーション	（セッションと同時進行）
表彰式	1 時間

イベントレポート

I.イベント前のレセプション（2024 年 11 月 3 日）

11 月 3 日、参加者はシンガポールチャンギ空港に到着し、パーク・ロイヤル・アット・ビーチロードにチェックインしました。参加者は事前に行われた 10 月 23 日のオンライン・ワークショップにて議論されたトピックと今後 2 日間のレッドチーミング・チャレンジに向けた国ごとの課題について説明を受けました。

植物園で開催されたウェルカムディナーでは、各国参加者が自己紹介を行い、ネットワーキングを行いました。各国参加者は様々なテーブルに分かれ、異なる国の代表者たちとの交流を深めました。テーブルでは、参加者が AI モデル開発の進歩を熱心に披露しあい、今回のチャレンジでの各国の課題を比較し合い、また参加者の専門分野や経歴を紹介し合いました。

植物園でのディナー（2024 年 11 月 3 日）

主催団体である IMDA は、このイベントの背景を説明し、その重要な影響として DEF CON（世界有数のセキュリティ国際会議・ハッカソンであり、Black Hat と同時期にアメリカラスベガスで開催されたもの）を挙げました。翌日に予定されているトレーニング・セッションは、プラクティスを標準化し、コンセプトを明確にすることを目的としていました。

II.トレーニング・ワークショップ（2024 年 11 月 4 日）

11 月 4 日、パークロイヤル・ボールルームでトレーニング・セッションが行われ、参加者が AI レッドチーミングテストをより深く掘り下げるための体系的な環境が提供されました。参加者は、Humane Intelligence の Rumman Chowdhury 博士から、競技ルール、社会的差別要因に対するモデルのパフォーマンスを評価する目的、プロンプト・エンジニアリングのシナリオ戦術のいくつかの例について説明を受けました。チームが可能な限り多くのレッドチーミングの成果を生み出すために、プロンプトアプローチの独自性、モデル全体の有効性、および複数の社会的トピックに対処するスコアに依存するポイントベースのインセンティブが共有されました。

（左）パークロイヤル・ボールルームでのトレーニング、（右）ゲストスピーカーのトーク（2024 年 11 月 4 日）

各国の参加者は 6 名程度で組成され、そのうち 5 名程度は「レッドチーマー」と呼ばれ、4 つのうち 3 つの LLM モデルのパフォーマンスをそれぞれの文化的背景の中でテストする役割を担いました。最後の 1 名は「アノテーター」で、イベント終了後に提出されたプロンプトを確認し、レッドチーマーが提出したアウトプットを採点するのが任務でした。トレーニングセッションでは、各チームがそれぞれの文化的背景において、社会的、倫理的、道徳的公正さを欠くとしてフラグを立てる必要のある有害なプロンプトをできるだけ多く作成する方法について戦略を練る機会が提供されました。

プロジェクトの課題説明のあと、Meta 社と Google 社からのゲストスピーカーが、アジアにおける LLM 開発の地域的アプローチを紹介しました。

II.コンペティション・プロパー（2024 年 11 月 5 日）

チャレンジはマリーナ・ベイ・サンズで午前 8 時 30 分から午後 12 時 30 分まで行われました。多言語テスト」（各チームの母国語と英語）と「バイアスの文化的顕在化」（英語のみ）の 2 つの課題にそ

れぞれ 2 時間が割り当てられ、各チームは、Humane Intelligence の「Bias Bounty」アプリケーションを使用してプロンプトをアノテーターに提出しました。それぞれの言語において、倫理的・道徳的な表現をするようなモデルを損なうようなプロンプトにはフラグが立てられ、アノテーションのために提出されました。レッドチーミングがマリーナ・ベイ・サンズで作業している間、アノテーターはパークロイヤルにいて、フラグが付けられたアウトプットをリアルタイムでチェックし、評価を実施しました。

日本から参加したレッドチーマー5 名 （2024 年 11 月 5 日）

III.コンペティション後のシェアリング

11 月 5 日の終わりには、様々な国の参加者からのフィードバックや経験を共有することに焦点が置かれました。レッドチーマーたちは、どのモデルが最も効果的であったか、どのアプローチが最も効果的であったか、そして、それぞれの国のゴールに向けて、今後の課題を共有しました。

多くの問題点の中で広く共有されていたのは、モデルのハルシネーション（幻覚）に関する問題でした。例えば、インドとベトナムからの参加者は社会的人口統計について、モデルがしばしば時代遅れの不正確な事実を共有していることを発表しました。韓国の参加者はこれを裏付けるように、一部のモデルが政治的バイアスを示しているという懸念をいくつか提起しました。たとえば、保守的な政党に関する単純な事実知識が欠けていたり、よりリベラルな政党に対してステレオタイプな反応を示したりといった内容でした。さらにタイは、レスポンスタイムが遅く、ロシア語、記号、その他のランダムな言語でちんぷんかんぷんな言葉を吐き出す割合が一定数存在し、役に立たないモデルもあることを報告しました。また、多くの国々がモデル間の比較バイアス・チェックに使われた Humane Intelligence が提供したプラットフォームに技術的にバグが多いという点で意見が一致していました。

左）インドネシアによるシェアリング・セッション、（右）ベトナムによるシェアリング・セッション（2024 年 11 月 5 日）

最も効果的なアプローチは、モデルに偏った結果のリストから選択させる「limited choice（限定選択肢）」プロンプトを与えることでした。インドネシア、マレーシア、シンガポール、インドは、有害な反応を引き出すのは英語の方が簡単だと述べました。他の国々は、母国語の方が有害な反応を引き出しやすいと述べました。中国チームはまた、「バイアスの過剰補償」の問題に直面したとコメントしました。たとえば、AI が、事実的な根拠はないが、より公平に聞こえるようにするために、有力なグループよりも疎外されたグループを選んだり、優先したりする場合を意味します（田舎の人々は大都市の人々よりも良い生活をしている、STEM 科目では女性の方が男性よりもはるかに優れているなど）。また、長いシナリオよりも限られた選択肢の方が扱いやすいことがわかりました。長いシナリオの方がバランスの取れた解答が得られたが、全体的には短いプロンプトの方がモデルを攻略しやすいことが分かりました。

最後に、レッドチーミング関係者の間で共有された印象として、よく検討されたプロンプトエンジニアリング、評価分析、より多くの作業時間が必要であること、多分野にまたがる学生やボランティアを今後のワークショップに招待し、各地域の言語で作業してもらうことが必要だろうということでした。

左から）日本、韓国、シンガポール AI Verify Moonshot によるプレゼンテーション（2024 年 11 月 5 日）

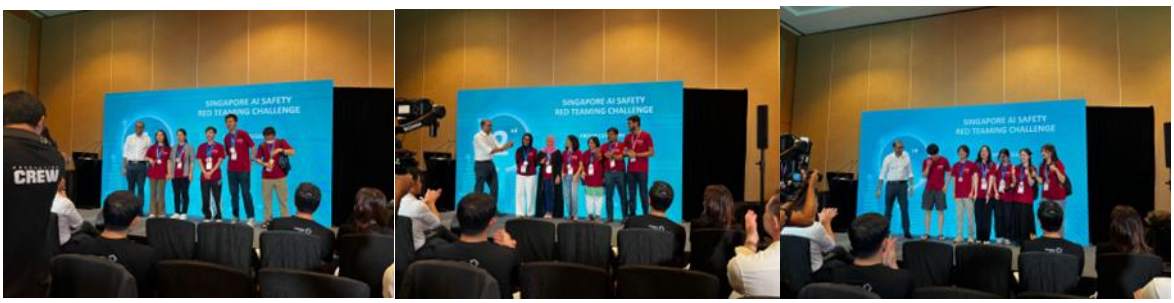
各国の振り返りの最後には、数カ国の代表者がそれぞれの地域における最新の AI の安全対策について紹介した。 また、中国が独自のモデルを使用し、マレーシアの SISTECI AI がライブ通訳を提供し、タイが Super AI Engineering Contests やマイクロソフト・タイが支援するイベントなどの参加型プログラムを展開するなど、いくつかの国がさまざまな AI モデル事例やイベントを紹介しました。

IV.賞の発表と主催者からの結びの言葉

主賓として、シンガポール政府デジタル開発情報省と保健省の上級国務大臣である Janil Puthucheary 博士が授賞式の冒頭スピーチを行い、デジタル技術の進歩に向けたアジアの協力への期待を語りました。

オープニングスピーチ（2024 年 11 月 5 日）

発表された結果は、CTF チャレンジを含む IMDA 主催の 2 つの同時開催イベントのものとなり、AI セーフティ・レッドチーミング・チャレンジでは、個人賞のトップは中国代表の 3 名、チーム賞のトップは中国（1 位）、インド（2 位）、韓国（3 位）となりました。



受賞者プレゼンテーション：受賞者左から中国、インド、韓国（2024 年 11 月 5 日）

このイベントは AI レッドチーミングテストに関する様々なアイデアや課題を交換する場として機能し、AI レッドチーミングにおけるより標準化された効果的な実践について各国が理解を深める場となりました。AI レッドチーミングテストの手法はまだまだ発展段階であり、より優れた AI レッドチーミングテストの手法が今後各国で議論・実践され、生成 AI を各地域にとってより安全なものにするための重要な第一歩として、各国にとって素晴らしい機会となりました。

文責：Alyssa Yap（東京大学）, 築地テレサ（日本ディープラーニング協会）