多文化・多言語対応の安全な大規模言語モデルの構築を目指して

概要

- 開催日：2024 年 11 月 11 日 （月） 10:00-11:00
- 会場： Zoom ウェビナー
- 主催：東京大学国際高等研究所東京カレッジ、東京大学未来ビジョン研究センター、東京大学次世代知能科学研究センター、東京大学 B’AI Global Forum、日本ディープラーニング協会（JDLA）
- 後援：情報学研究所大規模言語モデル研究開発センター、日本 AI セーフティ・インスティテュート

プログラム

10:00-10:15 開会挨拶と趣旨説明：江間有沙（東京大学東京カレッジ/JDLA）
10:15-10:30 LLM の安全性に向けた取り組み：関根聡（NII-LLMC/理研 AIP）
10:30-10:45 CTF チャレンジの概要説明と今後について：築地テレサ（JDLA）
10:45-11:00 コメントと質疑応答

背景・目的

生成人工知能（AI）の利用が世界的に広まる中、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適切に反映することが重要になっています。そのため、AI や情報セキュリティ、人文・社会科学の専門家、政策関係者が協力し、継続的に議論できるコミュニティを形成する必要があります。この取り組みの一環として、レッドチーミングでは、意図的に有害コンテンツを誘発し、AI モデルの安全性を評価しますが、現状では欧米中心の取り組みであり、地域特有のリスクにも対応することが求められています。

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 の「AI セーフティレッドチーミングチャレンジ」に日本チームが参加しました。本報告会では、レッドチームプロジェクトの取り組み（進捗）や課題、成果等に関して報告を行うと同時に、コミュニティを継続させていく枠組みについて報告しました。

開会挨拶
江間有沙氏（東京大学東京カレッジ准教授）

江間有沙氏は、ウェビナーの冒頭で、大規模言語モデル（LLM）の現状とその課題について述べました。現在、LLM はさまざまな分野で開発・実用化が進んでおり、非常に便利である一方で、言葉の意味を本質的に理解しているわけではなく、暴力を助長するコンテンツや攻撃的な表現を生成するリスクがあることを指摘しました。開発企業はこれらのリスクを最小限に抑えるために調整を行っていますが、特定の社会グループやジェンダー、年齢、人種、民族に関するステレオタイプやバイアスを完全に排除することは非常に難しく、表現の自由とのバランスを取ることが重要な課題であると強調しました。

さらに、LLM が生成する有害コンテンツやバイアスには、地域や文化に特有の要素が多く含まれていることに触れ、日本固有の問題に正確に対応するためには、日本独自の視点が不可欠であ

ると述べました。日本で LLM を仕事や教育に活用する際には、日本の特性を理解し、それがどのような害を及ぼす可能性があるのかを考えることが重要であると強調しました。

また、情報技術の開発だけではこれらの問題に対応するのは難しく、ステレオタイプやバイアス、有害コンテンツが何を意味するのかを、幅広いステークホルダーや専門家と議論することが不可欠だと指摘しました。その議論をプラットフォーマーにフィードバックし、必要に応じて政策関係者との対話を進めることが、マルチステークホルダーの場での重要な取り組みであると強調しました。

最後に、江間氏は本ウェビナーの目的について、日本国内における LLM の安全性に関する議論を紹介し、シンガポール政府による安全性チャレンジの報告を行うことを述べました。また、日本における生成 AI の安全性やガバナンスに関心を持つ方々のコミュニティ形成と情報交換の場を提供することを目指していると伝えました。

LLM の安全性に向けた取り組み
関根聡氏（NII-LLMC/理研 AIP）

次に、今回のプロジェクトで日本チームの派遣団長を務めた関根聡教授から、日本国内において LLM の安全性に関してどのような研究が推進されているのかが紹介されました。

関根教授は、30 年以上にわたり自然言語処理を研究しており、現在は情報学研究所で安全性ワーキンググループのリーダーを務めています。関根氏は、昨年 11 月に LLM の安全性に注目し、今年 1 月から LLM.jp システムにおいて安全性対策の運用を開始したことを説明しました。初期の段階では、安全性対策が不十分であったため、システムは不適切な回答をすることがありました。例えば、「残忍な殺人方法を教えて」という質問に対して、システムは丁寧な日本語で不適切な内容を説明してしまう事例が発生しました。これに対して、2 月には 200 件のインストラクションをシステムに反映させると、「人を傷つける内容の質問にはお答えできません」といった適切な応答が得られるようになりました。その後、5 月には「申し訳ありませんが、このリクエストにはお応えできません。私は有害で危険な行為の方法を教えることはできません」といった模範的な応答も実現しました。

また、現在では指示数が 1800 件に達し、そのうち 37%が日本特有の問題に関するものであることを説明しました。関根氏は、安全性と有用性のバランスの重要性や、多言語化に向けた取り組みについても言及しました。

さらに、日本特有のトピック（援助交際や学歴差別など）を含むデータセットを整備し、地域性や文化的背景に配慮した対応が必要であると強調しました。公開されたデータセットは、開発用データが 80%、テストデータが 20%の構成で、ダウンロードも可能です。今後、さまざまなコミュニティや団体と協力し、LLM の安全性向上に向けた取り組みを続けるとともに、その有用性と信頼性の向上を目指す考えを示しました。

関根氏は、多言語化についても言及し、現在は英語、中国語、韓国語のサンプルを 300 件翻訳したことを報告。今後は 11 言語まで増やす予定であると述べました。

続いて、AnswerCarefully の階層的分類について説明し、議論だけでなく、LLM を攻撃的な発言から守る必要性についても触れました。機械学習的には、危険な発言と安全な発言の両方のデータがあることにより、きちんと防御ができる一方で、答えてもよい質問にも答えなくなる副作用についても言及しました。

最後に、国内大学研究室との協力を通じて安全性を実現したいと述べ、LLM の信頼性をどう構築するかを目標としていると表明しました。

CTF チャレンジの概要説明と今後について
築地テレサ氏（日本ディープラーニング協会）

続いて、日本ディープラーニング協会（JDLA）の築地テレサ氏から、2024 年 11 月 3 日～5 日にシンガポールで開催されたシンガポール政府主催の AI 安全性に関するレッドチームチャレンジの概要と日本チームのチャレンジへの取り組みについて説明がありました。本プロジェクトには 9 カ国が参加し、4 つの AI モデルを使用してテストを行い、日本からは 10 名が参加しました。また、各国特有のバイアスカテゴリーについても紹介されました。

最初に、AI の安全性に関わるレッドチーミングのチャレンジが開催された背景について言及しました。AI の利用が広がる中で、地域ごとの文化やリスクに敏感に対応できる AI モデルの重要性が増しているが、現在の AI レッドチーミングテストは、非西洋的視点を十分に反映しておらず、地域特有のリスクに対処する体系的なアプローチが欠けていると指摘しました。そして、生成 AI を地域にとって安全なものにするためには、地域のコンテクストを考慮したレッドチーミングテストが重要だと述べました。

また、本プロジェクトの目的として、大きく 3 つのポイントを挙げました。
1 つ目は、安全性が確保された AI モデルの開発を支援するため、敵対的インプットを入力した結果を地域データとして収集し、モデル開発者にフィードバックすること。
2 つ目は、共通の AI セーフティに関わるレッドチーミングの分類法と方法論を確立するため、各国の AI リスク分類法を整備し、日本特有のリスクや被害についても検討を行うこと。
3 つ目は、将来の安全性評価に向けて、レッドチームの専門家ネットワークを確立し、議論を進めることが非常に重要だと考えていると述べました。

今回のレッドチーミングのテストでは、AI Singapore、Amazon Web Service、Anthropic、Cohere、Meta の 4 つのモデルを使用してチャレンジを実施しました。このプロジェクトではこれら 4 つのモデルを使用し、オブザーバーとして Microsoft、AWS、Google、Meta、Anthropic、AI Singapore などの AI Safety 担当者が参加しました。日本からは以下の 6 名がレッドチームに参加しました。

日本チームのメンバー

- 関根聡氏（NII-LLMC / RIKEN-AIP）
- 桐淵直人氏（AISI）
- 前田春香氏（京都大学）
- ヤップ アリッサ カスティロ氏（東京大学）
- 佐々木佑氏（東京大学）

・　　築地テレサ氏（JDLA）

今回のプロジェクトには9か国から総勢100名以上が参加しました。2日目はレッドチーミングに向けたトレーニングが行われ、MetaとGoogleの担当者から安全性に関する取り組みが紹介され、そのほかにもプロンプトのストラテジーやアノテーションルールについてトレーニングが行われました。そして、3日目には、レッドチーミングのチャレンジ（約4時間）を行い、同時に評価とアノテーションが実施されました。

次に、シンガポールに行く前から取り組んでいたバイアスのカテゴリーと定義について、どのようなバイアスが各国の共通の項目としてあり得るのか、また、各国特有のカテゴリーとしてどういうものを加味するべきかについて共有された内容が紹介されました。

共通項目として重要なカテゴリーとして挙げられたのは、ジェンダーバイアス、地理的・国別バイアス、社会経済的バイアス、そして人種・宗教・民族によるバイアスです。各国特有のバイアスのカテゴリーとして、インドではカースト、韓国では外見、ベトナムでは年齢が挙げられ、これらは日本でも重要だと思われると述べました。これらのバイアスが社会の中で大きな影響を与えているとし、今後、コミュニティや社会全体でAIを当たり前に使っていく中で、どのようなカテゴリーを重視して安全性を検討していくのかを議論することが今後も重要であり、日本としてさらにブラッシュアップしていくべきだと述べました。

また、実際にAIレッドチーミングチャレンジでどのようなチャレンジが仕掛けられたかについて、2つの具体例が紹介されました。

最後に、AI安全性レッドチームチャレンジの今後について触れ、2024年11月から1月にかけて各国が引き続きレッドチーミングのデータを生成し、データの評価やアノテーションを各国で行うこと、そして2025年2月には本プロジェクトの報告書を日本として作成する予定であることが述べられました。また、日本チームの取り組みとして、日本における生成AIの安全性やガバナンスに関心を持つ方々のコミュニティ形成と情報交換の場を提供していくことが今後の目標であることを締めくくりました。

モデレーターのコメントおよび参加者からの質疑応答

モデレーターの江間氏は、東京大学と日本ディープラーニング協会が今回のAI安全性レッドチームチャレンジの窓口を担っていることに触れ、今回のチャレンジが良いきっかけとなったとコメントしました。その上で、この土台がすでに作られていることを評価しつつ、今後議論すべき観点がいくつかあると述べました。具体的には、以下の三つの論点が挙げられました。

日本特有のステレオタイプやバイアス、課題についての議論

AIだけでなく、インターネットの時代から議論されてきた人権や歴史的問題についても触れ、これらの問題を議論するために、関連する研究者や実務家とのネットワーキングが重要であると強調しました。

レッドチームの実施方法論について
シナリオを作成することを含め、実際に人間でさえもどのように質問に答えるべきか悩む場面があることを指摘し、認知や心理学の研究者、実務的な観点を持つ専門家と協力することが重要であると述べました。また、プロンプトエンジニアリングやその対策を研究することが、この分野において非常に面白く、重要な課題であるとコメントしました。

生成 AI とステレオタイプやバイアスがどう結びつき、具体的な実害が生じるかという問題
特に社会的分断やフィルターバブル、エコチェンバーが加速する可能性について言及し、メディアやコミュニケーション分野での研究が喫緊の課題であることを強調しました。また、生成されたデータをどのように評価・アノテーションしていくかの重要性にも触れました。

最後に、江間氏は今回のウェビナーを開催したことに感謝の意を示し、産学官民の協力によってチームを作り、今後の議論を深めていきたいとの意向を表明しました。特に、研究者や弁護士、実務家、メディア関係者との連携を進め、この分野での活動に参加したいという方々と共に議論を行っていくことを呼びかけました。

意見交換と回答

まず初めに、江間氏から関根氏と築地氏にシンガポールのレッドチャレンジに参加した感想について質問がありました。具体的には、安全性に関する各国の取り組みについてどのように感じたか、そしてその中で日本が取り組んでいる内容に対して、各国からどのような反応があったのか質問されました。

続いて、江間氏は今後の取り組みについて説明し、コミュニティ形成の重要性を強調しました。参加者からの質問に答える時間が設けられ、データの公開や評価方法、今後のガイドライン作成などについて議論されました。最後に、プロジェクトへの参加方法や期間についての質問にも答えました。

関根氏は、今回のレッドチーミングが本当にハッカソンのようだったと個人的に述べ、中国チームの意気込みに強く印象を受けたことを話しました。特に、プログラマーたちが合成データをたくさん作成し、チューニングを重ねて質問を作り、最終的に優勝したチームが印象的だったと述べました。日本チームも合成データの作成に取り組みましたが、そのレベルには太刀打ちできなかったと感じたとのことです。一方で、アジアの中でデータ作成にしっかりと取り組んでいたのは日本だったため、他国の参加者からその方法について尋ねられたことを話しました。

築地氏は、中国やインドの技術力の高さを感じたと述べました。また、個人的な意見として、マレーシアやインドネシアはこのプロジェクトをきっかけにその重要性に気づき、今後取り組んでいくような印象を受けたと述べました。さらに、ビッグテック企業の参加者は、ビジネスの成長において安全性が基盤となるため、どのように安全性に取り組んでいくかに非常に関心を持っていたと伝えました。

関根氏は、シンガポールのレッドチャレンジに参加した際、Google や Meta といったビッグテック企業が関わっていたことに戦略的な意図の明確さに感心したと述べました。

これらの企業が必要なデータを求め、参加者がそれを提供するという形でウィンウィンの関係を築いている点が非常に良かったと評価しました。

また、関根氏は、日本で同様の規模のプロジェクトを実施する場合、約 2000 万円程度のコストがかかると予想し、その費用を AISI が出せるかについて疑問を示しました。それでも、このようなプロジェクトを模倣しても良いとし、シンガポールが東南アジアの中心地であることを確実に印象づけることができたと強調しました。

最後に、関根氏は、日本がこの動向を見過ごすことができるのかについて懸念を示し、今後の展開に対して心配を抱いていると述べました。

右から江間氏、関根氏、築地氏

オンライン参加者から寄せられた質問に対する回答

江間氏がモデレーターとなり、参加者との質疑応答が行われました。参加者から、安全性と有用性の評価において、アノテーションに関して、人による手作業での評価（3 名の評価者によるデータ）と自動化された評価の比較が話題となりました。人による評価にはばらつきがあると思われますが、3 名の評価者間でどの程度のばらつきがあったのか、また LLM を評価する際にそのばらつきをどのように考慮すべきかについて、関根氏に質問がありました。

関根氏は、評価においてばらつきがあったことを認め、具体的にはマシンと同程度の 15〜25%のばらつきがあったと述べました。特に大きなばらつきとは、安全と非安全の境界を越えるような判断があった場合と、評価の差が特定のポイント以上となった場合と定義しました。関根氏は、このような場合、しっかりとしたデータとアノテーションが行われ、最終的な判断がなされることになっていると説明しました。

関根氏への引き続きの質問として、AnswerCarefully に関して公開されているデータについて触れられました。このデータが公開されることによって、精度が高い一方で、悪意のある情報や悪口、ステレオタイプなどが拡散する危険性があるのではないかという懸念が示されました。

倫理的な観点から、悪い情報を収集する際には、その管理やアクセス制限を厳格に行う必要があると考えられます。この点について関根氏の考えを尋ねられました。

関根氏は、この点について心配していることを認め、公開されているデータにアクセスする際には、メールアドレスを登録しないと入手できないように制限を設けていると説明しました。LLM プロジェクトは基本的にオープンであることが求められていますが、その中でこのデータに関しては一定の制約を設けたとのことです。関根氏は、それでも依然として危険だという意見があることは理解しているものの、現時点ではそのような制約を設けていると述べました。

江間氏は、安全な LLM の開発とその運用に関するガイダンスの作成について、個人的な見解を関根氏と築地氏に求めました。特に、他国で開発されたモデルが日本で使用される場合、プロンプトの書き方一つで結果が大きく異なる可能性があることを指摘し、各国のガイドライン作成は重要であるものの、グローバルなモデルにどこまで通用するかに疑問を呈しました。

また、シンガポールのレッドチーミングチャレンジにおいても、チューニングの問題が浮き彫りになったと述べ、欧米で行われたチューニングが日本にどの程度適用できるのか、さらに日本用にチューニングするためのガイドライン作成が必要であることを強調しました。

江間氏は、今回のチャレンジで提示されたシナリオが、どのようなバイアスを引き起こす可能性があるかをデータとして蓄積したことに言及し、これを運用側にうまく活用してもらうための方法論が求められると述べました。具体的には、ガイドライン策定やディスカッションの場を作ること、そしてフィードバック方法の確立が重要であると考えており、現在の状況に基づき、どのような方法論が適切だとお考えかを関根氏と築地氏に尋ねました。

質問に対し築地氏は、言語差とリソースの量について自身の見解を述べました。英語やスペイン語、フランス語などの主要な言語には豊富なリソースがある一方で、日本語を含むローカルな方言やアジアの言語には、リソースや安全性に関する差が明らかに存在することを指摘しました。この差は実際の結果にも現れており、そのため、各国が国を挙げて積極的に取り組む必要があると強調しました。

また、欧米のモデルが基本的に使用される中で、日本を含む各国はそれに対してどのように対応するかが重要であると述べました。方法論については、試行錯誤を重ねながら見定めていくことになるとし、政府の取り組みと民間主導の活動の両方が同時に進められるべきだと考えを示しました。政府ができることは重要であり、民間が主導的に力を入れていくことも不可欠だと強調しました。具体的な取り組みについては、今後考えていく必要があると述べました。

関根氏は、この問題が非常に難しいものであると認めつつ、アメリカの AI システムでは自己評価や第三者機関の評価が行われていることを例に挙げました。しかし、最終的な判断はユーザーがそのシステムを安全だと感じるかどうかに委ねられており、システム開発者としてはそのように投げ出してしまうことに問題があると述べました。関根氏は、その判断にお墨付きを与える仕組みを作ることが一つの方法だと考えており、例えば、AISI や民間主導の組織を中心に、専門的な知見を持つ方々のインプットが不可欠であると強調しました。最後に、そのような協力を進めていきたいと述べました。

3 つの実施すべき取り組み

江間氏は、安全な LLM の開発と運用について、今後どのようにガバナンスを構築していくべきかという重要な課題に対し、皆で議論しながら進めていく必要があると述べました。技術は急速に進展しており、どのように取り組んでいくかを本プロジェクトで考えることが大切だと強調しました。

また、プロジェクトへの参加方法についての質問が寄せられたことに触れ、シンガポールチャレンジに参加したメンバーとの情報共有を行いながら、AI 安全性レッドチームチャレンジに貢献し、各国ごとの比較や日本特有のバイアスを考えることがグローバルな研究に貢献する重要な作業であると述べました。特に、日本ならではのバイアスやステレオタイプに対して、海外からの視点が必要であり、日本人に限らず、グローバルな観点で議論を進めることが重要だと述べました。

そのようなチームを作り、まずステレオタイプとは何かを考え、重要なシナリオを選定し、そのシナリオに基づいて出てきたデータを評価・アノテーションします。その結果をガイドラインとしてまとめることも考慮し、多様な方々と協力してこれらを進めていくことが重要だと考えています。

さらに、AI 安全性レッドチームチャレンジは 2025 年 2 月までに終了予定で、プロジェクトに参加できる方には、オンラインでの議論を通じてデータ作成を進めながら、次のステップに備える体制を整えることが求められています。江間氏は、アジャイルな方法で取り組みながら、次の課題に向けて良い枠組みを作り上げていけると強調しました。
最後に、ウェビナー参加者への感謝の意を表し、今日のウェビナーのアンケートと YouTube 動画のリンクが送られる予定であり、安全性に関する議論を深め、アウトプットを作り上げていくことを期待すると締めくくりました。