研究開発構想（個別研究型）
人工知能（AI）が浸透するデータ駆動型の経済社会に必要なAIセキュリティ技術の確立

「大規模言語モデルのミスアライメントに対するレッドチーミング基盤」

研究開発実施報告書（年次）

令和6（2024）年度

研究代表者
佐久間 淳
東京科学大学 情報理工学院・教授

１．当該年度における研究開発の実施概要

（１）研究開発概要

本研究開発は、大規模生成モデルの生成コンテンツのミスアライメント（人間の期待や倫理観から外れた挙動）を検出し、これを軽減・抑制するためのセキュリティ技術基盤の構築を目的としています。大規模生成モデルの利用においては、その生成コンテンツに有害情報・偽情報・差別的内容が含まれていたり、その生成コンテンツによって機密漏えい 、プライバシー侵害、著作権侵害などが発生したりする可能性があり、このようなミスアライメントへの対策が不可欠です。一般の開発者が外部から入手した大規模生成モデルをそのまま利用したり、これを手元のデータで改変して利用したりする状況では、このようなミスアライメントのリスクを独力で評価することは簡単ではありません。本研究開発では、生成モデルのこのようなミスアライメントに関するリスクの評価を支援するための技術基盤を提供し、ミスアライメントの抑制につなげます。

（２）実施内容と成果の概要（研究開発開始から当該年度末まで）

令和 6(2024)年度

- 研究計算基盤: NVIDIA 社製 B200 GPU 8 枚を搭載した DGX サーバ 1 台の調達を開始しました。
- 評価データ基盤: 「偽情報」についてのミスアライメントのデータ構築のため、アノテーション企業と協力しながらアノテーションガイドラインの作成とトライアルを実施しました。
- 敵対的評価データ基盤: 主要国際会議(AI/ML 3 会議、NLP 2 会議、セキュリティ 4 会議)について、LLM の攻撃に関連する文献を網羅するデータベースを整備し、敵対的評価データ基盤の実装対象となる攻撃アルゴリズムのスクリーニングプロセスを設計しました。
- レッドチーミングフレームワーク:システム設計について議論を進め、入札のための仕様を検討しました。

２．主たる研究分担者一覧

原 聡（電気通信大学　 大学院情報理工学研究科 教授）