配信先: 総務省記者クラブ、テレコム記者会、
文部科学記者会、科学記者会、
大阪科学・大学記者クラブ、
学研都市記者クラブ

プレスリリース
2026 年 2 月 26 日
SB Intuitions 株式会社
国立研究開発法人情報通信研究機構

SB Intuitions と情報通信研究機構、高性能 LLM の安全性技術に関する共同研究を開始

～高度な安全性と優れた日本語性能の両立により、信頼できる AI の社会実装を目指す～

SB Intuitions 株式会社（本社：東京都港区、代表取締役社長 兼 CEO：井尻 善久、以下「SB Intuitions」）と、国立研究開発法人情報通信研究機構（本部：東京都小金井市、理事長：徳田 英幸、以下「ＮＩＣＴ（エヌアイシーティー）」）は、高性能 LLM（大規模言語モデル）の安全性技術に関する共同研究（以下「本共同研究」）を 2026 年 2 月 18 日に開始しました。

本共同研究では、NICT が長年にわたり蓄積してきた高品質な言語資源ならびに不適切出力の検出・抑制といった LLM の安全性確保のための技術と、SB Intuitions が国産 LLM 「Sarashina（さらしな）」の開発等で培ってきた高度なモデル構築のノウハウを組み合わせることで、より高い安全性を備えた高性能な LLM の開発を目指します。なお、開発にあたっては、ソフトバンク株式会社の AI 計算基盤を最大限に活用します。

【背景】

生成 AI の急速な普及に伴い、その利便性が注目される一方で、AI が事実と異なる情報を生成する「ハルシネーション（幻覚）」や、不適切な表現の出力、著作権侵害のリスクなどに対する安全性の確保が喫緊の課題となっています。特に国内の企業活動や公的機関での活用においては、優れた日本語能力に加え、日本の法制度や倫理観に基づいた「信頼できる AI」への期待が高まっています。

こうした背景を受け、国内屈指の言語資源を有する NICT と、大規模な計算基盤を活用し、高度な LLM 開発を推進する SB Intuitions が連携し、データの構築からモデルの学習・評価、さらには安全性確保までを推進する共同研究を実施することになりました。

【研究概要】

・LLM の安全性向上に向けたアライメントや評価指標の開発

LLM が人の価値観や倫理観に沿って適切に動作をするように調整する技術（アライメント）や、その安全性を測るための評価指標などの開発に取り組みます。

・不適切な表現を検知するガードレール技術の開発

LLM への入力文および LLM が生成する出力文に含まれる不適切な表現を検出して、フィルタリングやブロックなどの制御を行うガードレール技術の研究に取り組みます。

＜各者の役割＞

SB Intuitions 株式会社

・高性能 LLM のベースモデル開発および学習の実施
・アライメントやガードレール技術など、安全性確保技術の研究開発

NICT

・長年にわたり蓄積してきた言語資源の提供と、LLM の安全性向上に資するデータの強化
・不適切出力の検知や抑制など、LLM の安全性確保・評価のための基盤技術の開発

＜各者概要＞

SB Intuitions 株式会社

ソフトバンクの子会社として、日本語に強い大規模言語モデル「Sarashina」シリーズを核に生成 AI の研究と周辺サービスの開発に力を入れ、国産の大規模言語モデルで日本語性能 No.1 を目指しています。国内最大規模の計算基盤と国内データセンターでの厳格なデータ管理を強みに、高い日本語性能を誇る 700 億パラメーターのモデル「Sarashina2-70B」や視覚言語モデル「Sarashina 2-Vision」等を公開し、産学連携によるオープンイノベーションで生成 AI の社会実装を加速しています。

https://www.sbintuitions.co.jp/

国立研究開発法人情報通信研究機構

国立研究開発法人情報通信研究機構（NICT）は、情報通信分野を専門とする我が国唯一の公的研究機関です。情報通信技術の研究開発を基礎から応用まで統合的な視点で推進し、同時に、大学、産業界、自治体、国内外の研究機関などと連携して、研究開発成果を広く社会に還元し、イノベーションを創出することを目指しています。

特に AI 分野に関しては、けいはんな地区に所在するユニバーサルコミュニケーション研究所において長年にわたり取り組んできており、多言語音声翻訳システム VoiceTra、大規模 Web 情報分析システム WISDOM X、対災害 SNS 情報分析システム DISAANA、D-SUMM、防災チャットボット SOCDA、高齢者介護支援用マルチモーダル音声対話システム MICSUS 等を様々な企業、組織と協力しながら、開発、社会実装してまいりました。近年では、過去 20 年近くにわたって収集してきた大量の日本語 Web ページを用いて日本語特化型の独自 LLM をフルスクラッチで開発してきました。加えて、多様な AI システムを組み合わせ可能なプラットフォームを構築し、様々な LLM に対して、自動合成した評価用プロンプトで LLM の出力を自動評価してリスクや弱点を特定、さらには追加学習データを自動合成して強化も図る能動的評価基盤の研究開発を、国内 AI 開発企業等と連携して進めています。

https://www.nict.go.jp/

＜ 本件に関する問合せ先 ＞

SB Intuitions 株式会社
広報部
E-mail：sbint-pr@sbintuitions.co.jp

国立研究開発法人情報通信研究機構
広報部 報道室
E-mail: publicity@nict.go.jp