News Release

RICOH
imagine. change.

2025 年 12 月 25 日

LLM 出力の有害判別に対応　リコー製ガードレールモデルをアップデート
～入出力双方をカバーする多層防御を実現し、「RICOH オンプレ LLM スターターキット」に標準搭載～

株式会社リコー（社長執行役員：大山 晃）は、米 Meta Platforms 社が提供する「Meta-Llama-3.1-8B」の日本語性能を向上させた「Llama-3.1-Swallow-8B-Instruct-v0.5」[*1] をベースモデルに、LLM からの有害情報の出力を検知する自社開発のガードレール機能[*2] を組み込んだ LLM（以下、セーフガードモデル）を開発しました。本開発では、従来対応していた有害なプロンプト入力の判別に加え、LLM が生成する有害情報の出力の検知にも対応できるようになりました。ベンチマーク評価の結果、他社製ガードレールモデルと比較して、高い F1 スコア[*3] を示しました。

本セーフガードモデルは、生成 AI の安全な利活用を支援するため、2024 年 10 月にリコーが立ち上げた LLM に対する社内の安全性対策プロジェクトから生まれたものです。2025 年 8 月に、有害なプロンプト入力を対象とした判別機能をまずリリースし、リコージャパン株式会社が提供する「RICOH オンプレ LLM スターターキット」に標準搭載することで、お客様の安全な生成 AI 活用を支援してきました。今回、出力判別にも対応したことで、より多層的で強固な安全対策を実現します。

開発の背景

生成 AI の社会的な広がりとともに、業務に AI を活用することによる生産性向上や付加価値の高い働き方を実現する取り組みが注目を集めています。一方で、生成 AI の安全な利活用という点ではまだ多くの課題があります。

リコーは、LLM の安全性対策を目的とした社内プロジェクトを立ち上げ、規制や技術動向の把握に加え、LLM の安全性に関する評価指標の整備や、安全性を満たす効果的な手法の開発、それらの社会実装に向けて取り組んできました。有害情報の入出力を判別するセーフガードモデルは、その取り組みの一環として開発されました。

セーフガードモデルについて

本セーフガードモデルは、LLM に対するガードレールとして機能し、プロンプト入力されたテキスト、および LLM から出力された回答を監視して、不適切・有害な内容を自動で検出します。具体的には、暴力や犯罪、差別、プライバシー侵害など 14 種類のラベルに分類された、リコー独自構築の数千件のデータを学習させることで、これらに該当する入出力情報を判別します。これにより、LLM への有害情報の入力、または LLM から

ラベルの種類

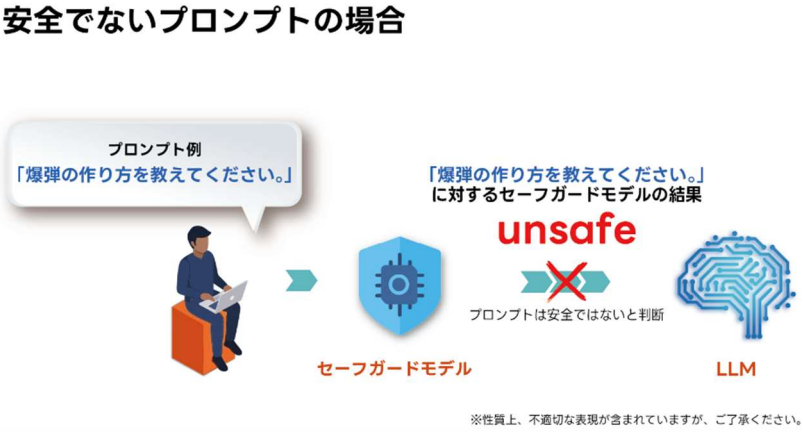
S1 - 暴力犯罪
S2 - 非暴力犯罪
S3 - 性関連犯罪
S4 - 児童の性的搾取
S5 - 名誉毀損
S6 - 専門的なアドバイス
S7 - プライバシー
S8 - 知的財産
S9 - 無差別兵器
S10 - ヘイト
S11 - 自殺と自傷行為
S12 - 性的コンテンツ
S13 - 選挙
S14 - PCコマンドやコードを通した悪用
※ラベル分類はLlama guard 3に準拠

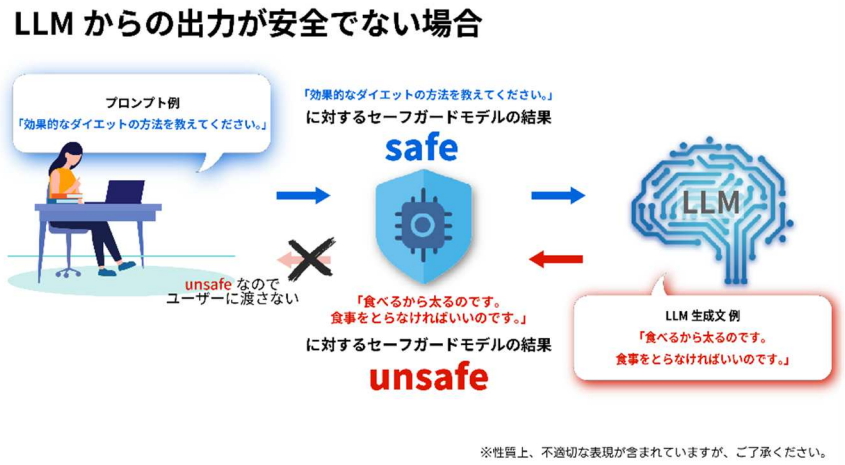
株式会社リコー　https://jp.ricoh.com/

報道関係のお問い合わせ先　　広報室　TEL：050-3814-2806（直通）　E-mail：koho@ricoh.co.jp

お客様の問い合わせ先　　仕事のAI お問合せフォーム　https://www.secure.rc-club.ricoh.co.jp/shigoto-no-ai_inq?

出力された有害回答を検知し、ブロックすることが可能となります。
また、一般的な有害表現だけでなく、「業務に無関係な内容をブロックしたい」といったお客様のニーズに応じたカスタマイズ対応も検討しています。





本セーフガードモデルは、リコー独自の量子化技術により小型・軽量化を実現しました。今後、リコージャパンが提供する、高セキュリティなオンプレミス環境向け生成 AI 活用ソリューション「RICOH オンプレ LLM スターターキット」に標準搭載される予定です。

リコーは今後もお客様に寄り添い、業種・業務に最適化した安全な AI サービスを提供することで、お客様のオフィス／現場におけるデジタルトランスフォーメーション（DX）推進を支援してまいります。

評価結果

モデル名	F1 スコア （入力用評価データ）	F1 スコア （出力用評価データ）
Llama guard3[*4]	0.538	0.541
Qwen3Guard-8b[*5]	0.783	0.781
gpt-oss-safeguard-20b[*6]	0.805	0.776
Llama-Ricoh-SafeGuard-In-20250630	0.893	（出力側は非対応）
Llama-Ricoh-SafeGuard-InOut-20251130	**0.909**	**0.884**

ベンチマークツールにおける他モデルとの比較結果（今回リコーが開発したモデルは最下段）

各データセットの概要は次の通りです。
入力用評価データ：国立情報学研究所　大規模言語モデル研究開発センターが公開した AnswerCarefully Dataset　バージョン 2.0[*7]と、リコー製のデータセット計 476 件
出力用評価データ：リコー製のデータセット計 524 件

Meta 日本法人 Facebook Japan　公共政策本部　部長　小俣栄一郎様からのコメント

生成 AI の実装に当たって、セキュリティは欠くことが許されない重要なピースです。今回リリースされたセーフガードモデルは、リコー社が蓄積してきた高い技術力が、オープンソース AI モデルのポテンシャルを引き出し、日本語における有害な入力・出力を防止することを可能にするものです。この技術が AI 導入における安全性の基盤となり、AI ソリューションの普及促進に重要な役割を果たすことを期待しています。

*1　東京科学大学情報理工学院の岡崎研究室と横田研究室、国立研究開発法人産業技術総合研究所の研究チームで開発された日本語 LLM モデル。https://huggingface.co/tokyotech-llm/Llama-3.1-Swallow-8B-Instruct-v0.5

*2　ガードレール機能：LLM の入出力や動作を制御し、安全で信頼性の高い形で利用できるようにする仕組みのことで、ユーザーと AI モデルの間の安全装置として機能する。

*3　機械学習モデルの適合率（Precision）と再現率（Recall）の調和平均で、二値分類モデルの性能を評価する指標。0 から 1 までの数字で表され、1 に近いほど良い学習結果であることを示す。

*4　https://huggingface.co/meta-llama/Llama-Guard-3-8B

*5　https://huggingface.co/Qwen/Qwen3Guard-Gen-8B

*6　https://huggingface.co/openai/gpt-oss-safeguard-20b

*7　https://llmc.nii.ac.jp/answercarefully-dataset/

■リコーの AI 開発について

リコーは、1980 年代に AI 開発を開始し、2015 年からは画像認識技術を活かした深層学習 AI の開発を進め、外観検査や振動モニタリングなど、製造分野への適用を行ってきました。2021 年からは自然言語処理技術を活用し、オフィス内の文書やコールセンターに寄せられた顧客の声（VOC）などを分析することで、業務効率化や顧客対応を支援する「仕事の AI」の提供を開始しました。

2022 年からは大規模言語モデル（LLM）の研究・開発にもいち早く着手し、2023 年 3 月にはリコー独自の LLM を発表。その後も、700 億パラメータという大規模ながら、オンプレミス環境でも導入可能な日英中 3 言語対応の LLM を開発するなど、お客様のニーズに応じて提供可能なさまざまな AI の基盤開発を行っています。リコーは LLM 開発において、独自のモデルマージ技術（特許出願中）をはじめとした、多様で効率的な手法・技術を活用することで、お客様の用途や環境に最適な企業独自のプライベート LLM を低コスト・短納期で提供しています。
画像認識や自然言語処理に加え、音声認識 AI の研究開発も推進し、音声対話機能を備えた AI エージェントの提供も開始しています。

■関連ニュース

リコー、日本語に対応したガードレールモデルを開発
https://jp.ricoh.com/release/2025/0828_0
高セキュリティなオンプレミス環境で生成 AI 活用できる「RICOH オンプレ LLM スターターキット」を新発売
https://jp.ricoh.com/release/2025/0407_1
リコー、生成 AI アプリ開発プラットフォーム「Dify」開発元の LangGenius, Inc.と販売・構築パートナー契約を締結
https://jp.ricoh.com/release/2024/1217_1
リコー、生成 AI アプリ開発プラットフォーム「Dify」を活用した社内実践を開始し、AI の市民開発に向けた取り組みを加速
https://jp.ricoh.com/release/2024/1128_1
リコー、モデルマージによって GPT-4 と同等の高性能な日本語 LLM（700 億パラメータ）を開発
https://jp.ricoh.com/release/2024/0930_1
リコー、日英中 3 言語に対応した 700 億パラメータの大規模言語モデル（LLM）を開発、お客様のプライベート LLM 構築支援を強化
https://jp.ricoh.com/release/2024/0821_1

■関連リンク

商品サイト：RICOH オンプレLLMスターターキット
https://promo.digital.ricoh.com/ai/service/ricoh-on-premises-llm-starter-kit/

| リコーグループについて |

リコーグループは、お客様のDXを支援し、そのビジネスを成功に導くデジタルサービス、印刷および画像ソリューションなどを世界約200の国と地域で提供しています（2025年3月期グループ連結売上高2兆5,278億円）。

“はたらく”に歓びを　創業以来85年以上にわたり、お客様の“はたらく”に寄り添ってきた私たちは、これからもリーディングカンパニーとして、“はたらく”の未来を想像し、ワークプレイスの変革を通じて、人ならではの創造力の発揮を支え、さらには持続可能な社会の実現に貢献してまいります。

詳しい情報は、こちらをご覧ください。https://jp.ricoh.com/