News Release

RICOH
imagine. change.

2026 年 5 月 20 日

リコー、自社開発のセーフガードモデルを無償公開

～LLM の入出力に含まれる有害情報を検出し、生成 AI の安全な利活用に貢献～

株式会社リコー（社長執行役員：大山　晃）は、大規模言語モデル（LLM）に対する有害情報の入出力を検知する自社開発のガードレール機能[*1] を組み込んだ LLM「Llama-Ricoh-SafeGuard-20260520」（以下、セーフガードモデル）を、本日から無償公開します。

本モデルは米 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」[*2] をベースに、リコーで追加開発を行ったものです。さらに、リコー独自の量子化技術により、小型・軽量化を実現しています。

これまで本モデルは、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」に標準搭載し、お客様へ提供してきました。今回、生成 AI の安全な利活用により一層貢献することを目的に無償で公開します。

【公開先】

https://huggingface.co/ricoh-ai/Llama-Ricoh-SafeGuard-20260520

1. セーフガードモデル開発の背景

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、2024 年 10 月に 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。

本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たものです。2025 年 8 月には、有害なプロンプト入力を対象とした判別機能をリリースし、同年 12 月には、LLM が生成する有害な出力情報の検知にも対応しました。

2. 無償公開の狙い

近年、LLM の活用が広がる一方で、日本においては LLM 分野におけるオープンモデルの選択肢が少ないという課題が指摘されています。

リコーはこれまで、経済産業省と国立研究開発法人新エネルギー・産業技術総合開発機構（NEDO）が推進する、国内における生成 AI の開発力強化を目的としたプロジェクト「GENIAC（Generative AI Accelerator Challenge）[*3]」の第 2 期、第 3 期に参画し、図表を含む多様なドキュメントを高精度に読み取るマルチモーダル大規模言語モデルを無償公開してきました。

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

また、ガードレール LLM においても重要性が高まる一方で、日本のビジネスの現場で実用的に利用できるモデルは少ない状況があります。リコーは、本セーフガードモデルをいち早く無償公開することで、その重要性を社会に提起するとともに、生成 AI の安全な利活用の推進に貢献していきます。

リコーは、企業の業務革新と付加価値の高い働き方を支援し、企業理念の使命と目指す姿として掲げる「“はたらく”に歓びを」の実現に向けて、引き続き取り組んでまいります。

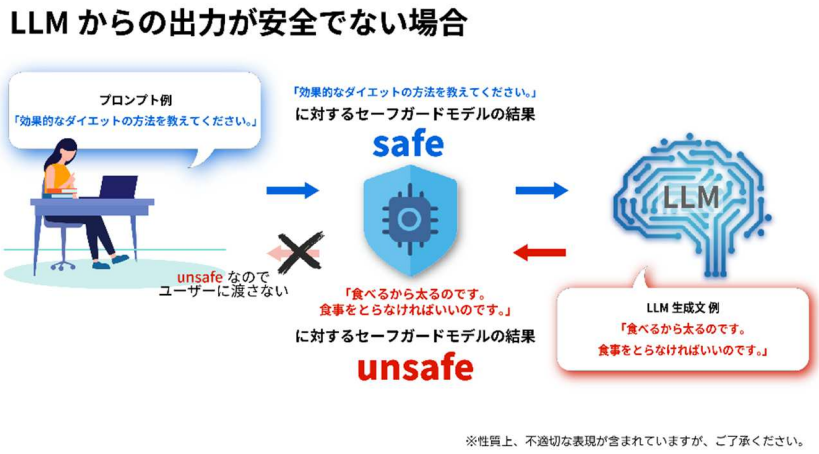
3. セーフガードモデルについて

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、入力されたプロンプト、および LLM が生成した回答を監視することで、不適切または有害な内容を自動的に検出します。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、リコー独自に構築した数千件規模のデータを学習させています。これにより、LLM への有害情報の入力や、LLM から出力される有害な回答を高精度に判別し、検知・ブロックすることが可能となります。

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

ベンチマーク結果など詳細はこちら
https://jp.ricoh.com/release/2025/1225_1



*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　GENIAC（ジーニアック/Generative AI Accelerator Challenge）：主に生成 AI のコア技術である基盤モデルの開発に対する計算資源の提供や、データや AI の利活用に向けた実証調査の支援等を実施するプロジェクト。

■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。

さらに、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。また、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

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
https://jp.ricoh.com/release/2025/1225_1

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

リコー、「GENIAC」第 3 期においてリーズニング性能を備えたマルチモーダル大規模言語モデルを開発
https://jp.ricoh.com/release/2026/0330_1

リコー、マルチモーダル LLM の基本モデルと評価環境を無償公開
https://prtimes.jp/main/html/rd/p/000000167.000043114.html

■関連リンク

技術ページ：“はたらく”を支えるリコーの大規模言語モデル（LLM）
https://jp.ricoh.com/technology/ai/LLM

Hugging Face：「Qwen3-VL-Ricoh-8B-20260227」を公開
https://huggingface.co/ricoh-ai/Qwen-3-VL-Ricoh-8B-20260227

リコーグループについて

リコーグループは、世界約200の国・地域で、AIをはじめとする先進テクノロジーと、長年培ってきたプリンティング領域の強みを基盤に、ワークプレイスにおける業務変革を支援するサービス・ソリューションを提供しています。また、商用・産業印刷事業や、インクジェット技術を応用した新たなソリューションの展開を通じて、お客様の価値創出を支えています（2026年3月期グループ連結売上高2兆6,083億円）。

“はたらく”に歓びを　創業以来90年にわたり、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/