

2026 年 10 月 20 日

リコー、生成 AI の安全な業務活用を支援する 日本語向けセーフガードモデルを刷新

～Gemma 4 をベースに小型化、安全性判定基準も見直し安全性と実用性を両立～

株式会社リコー（社長執行役員：大山 晃）は、生成 AI への入力や、生成 AI が出力する内容が安全かどうかを判定する、ガードレール機能^{*1}を組み込んだ日本語向け LLM（以下、セーフガードモデル）の改良版を開発しました。セーフガードモデルは、大規模言語モデル（LLM）の入出力に含まれる有害情報を検知し、生成 AI の安全な利用を支援する、安全性に特化した LLM です。

今回の新モデルは、「gemma-4-E4B-it」^{*2}をベースにリコーが追加開発したものです。従来モデル（「Llama-Ricoh-SafeGuard-20260520」）と同等水準の安全性判定性能を維持しながら、モデルの小型化により、より少ない計算資源での利用を可能にしています。あわせて、近年の LLM の性能向上や実際の利用シーンの広がりを踏まえ、安全性判定基準を見直しました。これにより、企業や組織における生成 AI 活用において、必要以上の利用制限を抑えながら、安全性と実用性を両立した運用を支援します。

新モデルは、11 月から順次、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、お客様へ提供します。

1. セーフガードモデル開発の背景

生成 AI の業務活用が広がる一方、不適切な情報の生成や機密情報の漏えいなど、安全性への対応が重要になっています。企業や組織が生成 AI を業務で安心して活用するためには、LLM の性能向上だけでなく、入力内容や生成結果を適切に監視・制御する仕組みが求められています。

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、LLM の安全性に関する評価指標の整備、安全性を高めるための技術開発と社会実装に取り組んできました。2025 年 8 月には、有害なプロンプト入力を判別する機能をリリースし、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。2026 年 5 月からは、セーフガードモデルを Hugging Face 上で無償公開しています。今回、これまでの開発で培った知見をもとに、より実用性を高めた改良版モデルを開発しました。

2. 本モデルの特徴

①モデルサイズを大幅に小型化、少ない計算資源で利用可能

今回の新モデルは、ベースモデルを従来の「Llama-3.1-Swallow-8B-Instruct-v0.5」から「gemma-4-E4B-it」に刷新しました。安全性判定性能は従来モデルと同等水準を維持しながら、モデルの小型化により、より少ない計算資源で利用できます。これにより、オンプレミス環境など、利用できる計算資源に制約のある環境にも導入しやすくなります。

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

評価結果

モデル名	F1 スコア ^{*3} (入力用評価データ)	F1 スコア ^{*3} (出力用評価データ)
Llama guard3	0.556	0.618
Qwen3Guard-8b	0.834	0.851
Llama-Ricoh-SafeGuard-20260520	0.891	0.952
本セーフガードモデル(ベース: Gemma 4)	0.900	0.942

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^{*4}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 508 件

②LLM の進化と実利用を踏まえた安全性判定基準の見直し

近年、LLM の性能向上により、従来は安全上の懸念から一律に制限していた内容についても、利用目的や文脈に応じて有用かつ適切な回答が可能なケースが増えています。そこで今回の新モデルでは、実際の利用シーンを踏まえて安全性判定基準を見直しました。

この見直しにより、必要以上に利用を制限するのではなく、安全性を確保しながら LLM 本来の有用性を活かす運用が可能になります。これにより、企業や組織における生成 AI の業務活用範囲の拡大を支援します。

ラベルの種類

S1: 暴力犯罪
S2: 非暴力犯罪
S3: 性犯罪
S4: 児童搾取
S5: 誹謗中傷
S6: 危険な専門的アドバイス
S7: プライバシー侵害
S8: 知的財産権侵害
S9: 無差別兵器
S10: ヘイト
S11: 自傷行為
S12: 性的表現
S13: コードインテグリティの乱用

※分類はLlama guard 3をベースにリコーで設計

③顧客ごとのポリシーに応じた簡易的なカスタマイズにも対応

本セーフガードモデルは、あらかじめ学習された 13 種類の一般的な有害カテゴリに加え、ユーザーが用途や運用ポリシーに応じて、任意の判定ラベルを追加できます。モデルの追加学習を行わずに、業界固有のルールや企業ごとのガイドラインなどのラベルを追加することで、簡易的なカスタマイズが可能です。

[>詳細は技術ページをご覧ください。](#)

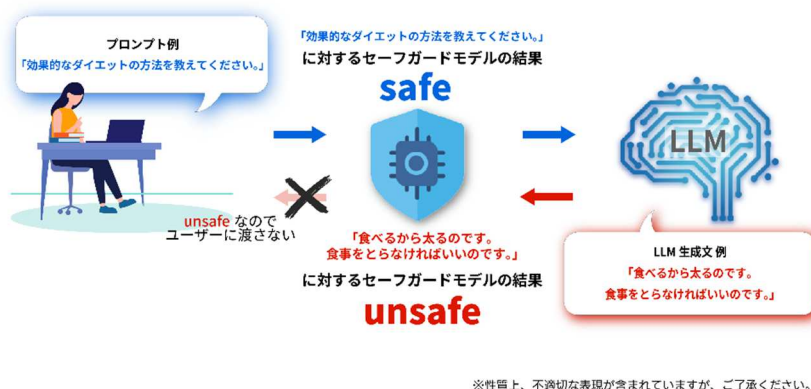
3. 今後の展望

新モデルは、11 月から順次、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、お客様へ提供します。今後もリコーは、安全性と実用性を両立する AI 技術の開発と実装を進め、お客様が生成 AI を安心して業務に活用できる環境づくりを支援していきます。

安全でないプロンプトの場合



LLM からの出力が安全でない場合



*1 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。

*2 <https://huggingface.co/google/gemma-4-E4B-it>

*3 F1 スコア: 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。

*4 <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。企業内に蓄積された文書・画像・業務データに内在する暗黙知を AI によって可視化し、業務成果へとつなげる AI 開発を推進しています。独自のドキュメント理解技術と LLM/LMM を組み合わせることで、テキストや図表を高精度に解析し、実務で活用可能な知識として展開します。さらに、オンプレミス対応やセキュリティを重視した提供により、安心して利用できる環境を提供しています。

また、リコーは AI を自社業務において積極的に活用し、その知見をサービスに反映しています。社員自らが AI エージェントを構築・活用し、業務の高度化を実践していることが強みです。こうした実践知と、顧客の現場に入り込み業務を深く理解してきた実装力を掛け合わせることで、AI を構想にとどめることなく、継続的に価値を創出・提供していきます。

■関連ニュース

リコー、自社開発のセーフガードモデルを無償公開

https://jp.ricoh.com/release/2026/0520_1

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

■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>

※社名、製品名は、各社の商標または登録商標です。

｜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/>