

2026 年 10 月 15 日

**リコー、MoE 型マルチモーダル大規模言語モデル
「リコーLMM_35B-A3B」を開発
～高い日本語リーズニング性能と高速・省リソースな運用を両立、
オンプレミス向け LLM ラインアップを拡充～**

株式会社リコー（社長執行役員：大山 晃）は、日本語の文書や図表を高精度に理解し、高速・省リソースで動作するマルチモーダル大規模言語モデル（LMM）「リコーLMM_35B-A3B」を開発しました。本モデルは、アリババクラウドが開発・提供する MoE（Mixture of Experts）^{*1} 構造を採用した「Qwen3.6-35B-A3B」をベースに、リコー独自の強化学習^{*2} およびカリキュラム学習^{*3} 技術を活用して、日本語でのリーズニング性能^{*4}を高めたものです。

本モデルは、総パラメータ数約 350 億のうち、推論時に約 30 億パラメータのみを選択的に稼働させる MoE 構造を採用しています。これにより、同規模の Dense 型^{*5} モデルと比べて推論時の計算量を抑え、高速・省リソースな処理を可能にしています。

図表を含む日本語ドキュメントのリーズニング性能を評価するリコー独自のベンチマークツール「JDocQA Reasoning Benchmark」^{*6} による評価の結果、ベースモデルを上回り、商用クラウド AI モデル「GPT-5.6 Sol」と同等水準の性能を確認しました。

あわせて、NVIDIA の Blackwell 世代以降の GPU で効率的に動作するよう量子化^{*7} したリコー LMM_35B-A3B モデルの NVFP4 版を開発し、超小型デスクサイド AI 用サーバー「NVIDIA DGX Spark」1 台で実用レベルの生成 AI 処理を実現しました。

本モデルは 11 月から順次「RICOH オンプレ LLM スターターキット」に搭載し、リコージャパン株式会社から提供予定です。2026 年 6 月に発表した高精度志向の Dense 型モデル「リコーLMM_27B」、より軽量の「リコーLMM_9B」に、速度と省リソース性に優れた本モデルを加えることで、求められる精度や処理速度、利用可能な計算資源など、お客様の業務要件や導入環境に応じて最適なモデルを選択できるラインアップを拡充します。

1. 背景

生成 AI の企業活用が本格化するなか、製造業、金融・保険、公共分野などでは、請求書、経営資料、設計図書、報告書、約款、申請書類など、図表を含む多様な社内ドキュメントを生成 AI で活用したいニーズが高まっています。一方で、セキュリティやガバナンスの観点から、クラウド AI の利用が難しく、オンプレミス環境で安全に利用できる高性能な LMM への期待が高まっています。

また、生成 AI の高性能化に伴い、推論に必要な GPU リソースや電力、応答時間が導入・運用上の課題となっています。特に、大量文書を一括で処理する業務や、多数の同時アクセスが発生する環境では、精度に加えて、限られた計算資源で高速に処理できることが重要になります。

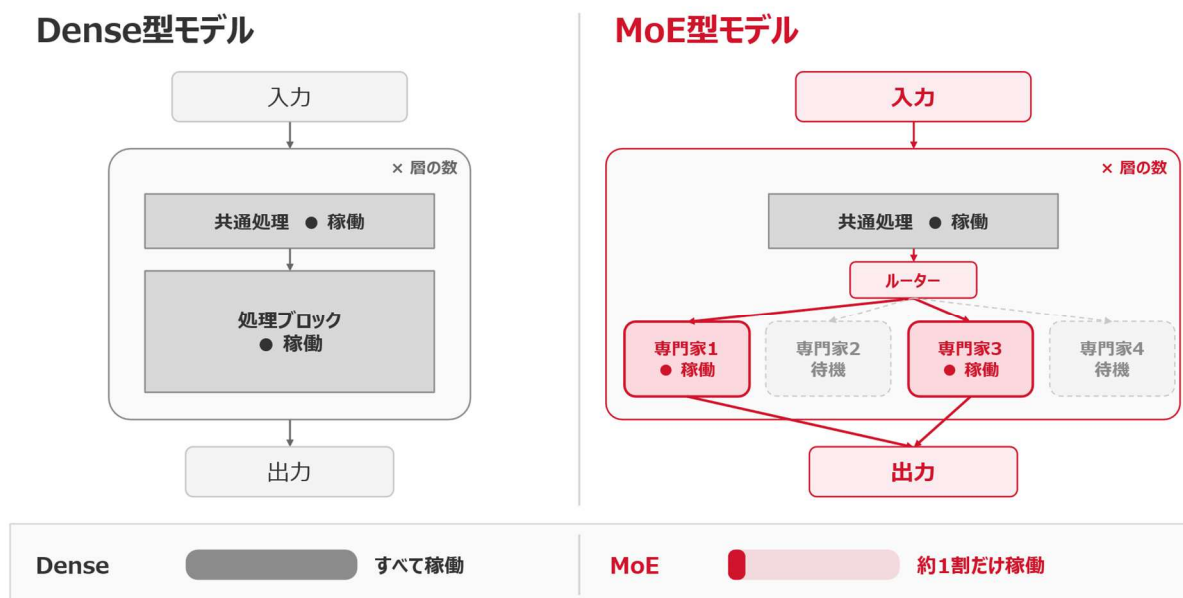
こうした課題に対し、必要な部分のみを選択的に稼働させることで計算量を抑える MoE 構造は、高い

性能と処理効率を両立する手法として注目されています。リコーは今回、初めて MoE 構造を採用したモデルをベースとし、独自の追加学習を通じて、日本企業の業務で実用的に活用できるモデルを開発しました。

2. 本モデル「リコーLMM-35B-A3B」の特徴

① MoE 構造による高速・省リソースな推論

本モデルは、総パラメータ数約 350 億のうち、推論時に約 30 億パラメータのみを稼働させる MoE 構造を採用しています。MoE 構造では、モデル内部の処理を複数の「専門家」に分担させ、入力に応じて必要な専門家だけを選んで動かします(模式図)。これにより、同規模の Dense 型モデルと比べて 1 トークンあたりの計算量を抑え、高速な処理を実現します。最大 256K トークンの長文コンテキストに対応しており、長大なドキュメントを一度に読み込み、推論することが可能です。画像に加えて動画入力にも対応し、外部ツールを呼び出すツール利用 (Tool use) 機能も備えています。



Dense 型モデルと MoE 型モデルの推論時に稼働するパラメータの違い (模式図)

② 日本語リーズニング性能の向上

リコー独自の強化学習およびカリキュラム学習技術を MoE 向けに発展させ、図表を含む企業ドキュメントの読解性能をベースモデル「Qwen3.6-35B-A3B」から向上させました。

リコー独自ベンチマーク「JDocQA-Reasoning Benchmark v2.0」において、本モデルはベースモデルの 0.816 から 0.841 に向上しました。これは、推論時に稼働するパラメータが約 9 分の 1 でありながら、同じ Qwen3.6 系列の Dense 型モデル「Qwen3.6-27B」(0.841)と同等の水準です。また、商用クラウド AI モデル「GPT-5.6 Sol」(0.821)と同等以上の性能を確認し、同規模の MoE 型オープンモデル「gemma-4-26B-A4B-it」(0.784)を上回りました。

加えて、ELYZA-tasks-100 や Japanese MT-Bench による日本語 LLM 性能評価では、ベースモデ

ルと同水準を維持しており、図表読解能力の向上と汎用的な日本語性能の両立を実現しています。

③ 日本語 AI エージェント^{*8} 性能の強化

本モデルは、文書の読解や回答生成に加え、業務システムやツールを活用しながらタスクを自律的に実行する AI エージェント用途にも対応しています。日本語 AI エージェント評価ベンチマーク「j-tau-bench」(SB Intuitions 公開)の通信事業者カスタマーサービス(telecom)シナリオにおいて、業務ポリシーに従ってツールを利用しながらユーザーと対話し、依頼されたタスクを正しく遂行できた割合(正答率)が、ベースモデルの 62.4%から 77.7%へと 15.3 ポイント向上しました。これにより、社内文書の読解・活用だけでなく、問い合わせ対応や各種業務手続きの自動化を支援します。

④ ループ回答^{*9}の抑制による応答時間の安定化

リーズニング型モデルでは、同じ内容を繰り返して出力し続ける「ループ回答」が実運用上の課題となることがあります。リコーは、過度に長い出力を抑制する報酬設計を強化学習に取り入れることで、回答の品質を維持しながら不要な繰り返しを抑制しました。これにより、応答時間の安定性を高めています。

⑤ 小型 AI サーバーで動作する量子化版も開発

NVIDIA Blackwell 世代以降の GPU 向けに量子化した「リコーLMM_35B-A3B」の NVFP4 量子化版を開発しました。小型 AI サーバー「NVIDIA DGX Spark」1台で動作し、量子化後もベースモデルの量子化版を上回るリーズニング性能を確認するとともに、BF16 版と比べて 1 問あたりの応答時間を約 3 分の 1(MTP 併用時は約 4 分の 1)に短縮しています。

評価結果含めて、詳細は技術ページをご覧ください。

<https://jp.ricoh.com/technology/ai/LLM>

3. 想定される活用シーン

- ・製造業: 設計図と要求仕様の適合確認、トラブル対応時の社内ドキュメント参照
- ・金融・保険: 約款や報告書の高精度な読解と要点抽出
- ・公共・自治体: 申請書類や図表を含む行政文書の処理支援
- ・業務文書理解: 会議資料や提案資料からの情報抽出・分析
- ・大量文書の一括処理: 過去の帳票・報告書のデータ化やナレッジ化など、高い生成速度が求められる業務

4. 今後の展開

リコーは、2026 年 6 月に発表した高精度志向の Dense 型モデル「リコーLMM_27B」、軽量モデル「リコーLMM_9B」に、今回開発した速度・省リソース性に優れた MoE 型モデルを加えることで、お客様の業務要件や導入環境に応じて最適なモデルを選択できるラインアップを拡充します。

本モデルは 11 月から順次「RICOH オンプレ LLM スターターキット」に搭載し、リコージャパン株式会社から提供予定です。リコーは、オンプレミス環境で安全に活用できる LMM の拡充を通じて、日本企業のドキュメント利活用と業務革新を支援してまいります。

- *1 MoE (Mixture of Experts): モデル内部に複数の「専門家 (エキスパート)」と呼ばれるサブネットワークを持ち、入力ごとに一部の専門家のみを選んで計算するアーキテクチャ。総パラメータ数に対して実際に稼働するパラメータ数 (アクティブパラメータ数) が少ないため、高速・省リソースな推論が可能。本モデルの「35B-A3B」は総パラメータ数約 350 億、アクティブパラメータ数約 30 億を示す。
- *2 強化学習: AI による機械学習の一分野で、AI が環境と相互作用しながら最適な行動を学習する方法。AI が与えられた目標に対して、試行錯誤して得る報酬を最大化するため、最適な行動戦略を見つけ出すことが特徴。
- *3 カリキュラム学習: AI による機械学習の一分野で、学習の順序を工夫した学習方法。
- *4 リーズニング性能: LLM が単に情報を検索したりテキストを生成したりするだけでなく、複数のステップからなる論理的な思考プロセスを経て結論を導き出す性能。
- *5 Dense 型: モデル全体を用いて計算を行う従来型のアーキテクチャ。入力ごとに一部の専門家 (エキスパート) のみを利用する MoE モデルと異なり、すべてのパラメータが常に動作するため、構造がシンプルで学習が安定しやすい一方、パラメータ数に比例して推論時の計算量が大きくなる。
- *6 JDocQA Reasoning Benchmark: https://jp.ricoh.com/release/2026/0529_1
- *7 量子化 / NVFP4: モデルの重みなどをより少ないビット数で表現し、必要なメモリ量を削減して推論を高速化する技術。NVFP4 は NVIDIA Blackwell 世代 GPU で導入された 4 ビット浮動小数点形式。
- *8 エージェント (AI エージェント): LLM が与えられた目標に対し、外部のツールやシステムを呼び出し、利用者に対話しながら複数のステップからなるタスクを自律的に遂行する仕組み。
- *9 ループ回答: LLM が同じ語句や文を繰り返し生成し続け、回答が収束しなくなる現象。出力長の上限に達するまで生成が続くため、応答時間の長期化や計算リソースの浪費につながる。

■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、マルチモーダル大規模言語モデル「Qwen3.6-Ricoh-27B-20260522」および「Qwen3.5-Ricoh-

9B-20260522」を開発

https://jp.ricoh.com/release/2026/0605_1

リコー、生成 AI の推論性能を測る独自ベンチマークを無償公開

https://jp.ricoh.com/release/2026/0529_1

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

リコーとライズ・コンサルティング・グループ、AX 実現を支援する合弁会社設立に向け基本合意

https://jp.ricoh.com/release/2026/0303_1

■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>

リコーの AI 開発の最前線: 企業の AI 活用を広げる「マルチモーダル LLM」

<https://jp.ricoh.com/news/stories/articles/multimodal-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/>