News Release

RICOH
imagine. change.

2025 年 12 月 8 日

リコー、「Gemma 3 27B」ベースにオンプレミス導入に最適な日本語 LLM を開発
～エフサステクノロジーズの「Private AI Platform on PRIMERGY」に搭載して提供開始～

株式会社リコー（社長執行役員：大山　晃）は、自社で開発・提供する日本語大規模言語モデル[*1]（以下、LLM）シリーズの次世代モデルとして、Google が提供するオープンモデル「Gemma 3 27B[*2]」をベースに、オンプレミス環境への導入に最適な高性能 LLM を開発しました。

本 LLM は、リコー独自のモデルマージ[*3]技術を活用し、ベースモデルから大幅な性能向上を実現しています。具体的には、独自開発を含む約 1 万 5 千件のインストラクションチューニングデータで追加学習した Instruct モデルから抽出した Chat Vector[*4]など複数の Chat Vector を開発し、「Gemma 3 27B」に対して独自技術でマージしています。

同規模パラメータ数の LLM とのベンチマーク評価の結果、米 OpenAI のオープンウェイトモデル「gpt-oss-20b[*5]」をはじめとする最先端の高性能モデルと同等の性能を確認しました。さらに、本モデルは、ユーザー体験を重視した非推論モデル[*6]ならではの高い初期応答性[*7]を実現しながら、高い執筆能力も兼ね備えており、ビジネス用途での活用に適しています。

また、モデルサイズは 270 億パラメータとコンパクトでありながら高性能を実現しており、PC サーバ[*8]等で構築でき、低コストでのプライベート LLM 導入を可能にします。LLM は高い電力消費による環境負荷が課題となっていますが、コンパクトで高性能な本 LLM は省エネルギー・環境負荷低減にも寄与します。

本 LLM は、お客様のご要望に応じて個別提供が可能です。さらに、2025 年 12 月下旬からは、エフサステクノロジーズ株式会社が提供するオンプレミス環境向けの対話型生成 AI 基盤「Private AI Platform on PRIMERGY（Very Small モデル）」に、本 LLM の量子化モデルと生成 AI 開発プラットフォーム「Dify（ディフィ）」をプリインストールし、LLM 動作環境を構築したうえで、リコージャパン株式会社から提供します。本 LLM と Dify を活用することで、お客様は自社の業種・業務に合わせた生成 AI アプリケーションなどをノーコードで作成できます。さらに、リコージャパンが提供する「Dify 支援サービス」による伴走支援も可能なため、社内に AI の専門人材がいない場合でも安心して生成 AI の業務活用を開始できます。

今後は、推論性能[*9]や業種特化モデルの開発を進めるとともに、リコーが強みとするマルチモーダル性能と合わせて、リコーの LLM ラインアップをさらに強化してまいります。

リコーは、お客様に寄り添い、業種業務に合わせて利用できる AI サービスの提供により、お客様が取り組むオフィス／現場のデジタルトランスフォーメーション（DX）を支援してまいります。

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

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?

【評価結果】

複雑な指示やタスクを含む代表的な日本語ベンチマーク「ELYZA-tasks-100」、日本語のマルチターンの対話能力を評価する「Japanese MT-Bench」により、性能を評価しました。その結果、リコーが開発したLLMは、日本語ベンチマークにおいて米OpenAIが開発したオープンウェイトモデル「gpt-oss-20b」をはじめとする最先端の高性能なモデルと同等レベルの高いスコアを示しました。

企業/組織	モデル名	推論モデル/非推論モデル	Japanese MT-Bench	Elyza-tasks-100	平均スコア
Google	gemma-3-27b-it	非推論	8.90	8.63	8.76
Alibaba Cloud	Qwen3-32B (/no_think)	非推論	8.92	8.95	8.93
Alibaba Cloud	Qwen3-32B (/think)	推論	9.26	8.98	9.12
Open AI	gpt-oss-20b	推論	9.48	8.92	9.20
Ricoh	gemma-3-Ricoh-27b-20251030	非推論	9.26	9.03	9.15
Ricoh	gemma-3-Ricoh-27b-20251030-gptq	非推論	9.01	9.05	9.03

ベンチマークツールにおける他モデルとの比較結果
（今回開発したモデルが下から2段目、その量子化モデルが最下段）

各ベンチマーク・データセットの概要は次の通りです。

- **Japanese MT-Bench**：マルチターン対話設定のデータセット。タスクはコーディング、抽出、人文、数学、推論、ロールプレイ、STEM、ライティングから成る。スコアの範囲は1（最低）から10（最高）。
- **Elyza-tasks-100**：複雑な指示・タスクを含むデータセット。要約の修正、意図の汲み取り、複雑な計算、対話生成など広範なタスクから成る。スコアの範囲は1（最低）から5（最高）。ここではJapanese MT-Benchとの平均スコアを算出するため、スコアを2倍にして比較。

エフサステクノロジーズ株式会社 代表取締役社長 CEO 保田 益男様のコメント

株式会社リコーの開発した高性能な LLM と小規模から大規模まで幅広いラインナップを持つ対話型生成 AI 基盤「Private AI Platform on PRIMERGY」を組み合わせたオンプレミス AI ソリューションをリコーグループの販売網を通じて、より多くのお客様にご提供できることを大変嬉しく思います。
私たちは開発から製造・品質保証・サポートまでを一貫して担う体制を活かしたハードウェアを通して、株式会社リコーと連携し、お客様のデジタルトランスフォーメーションを支援するとともに夢のある未来の共創に取り組んでまいります。

株式会社リコー リコーデジタルサービス BU AI サービス事業本部 本部長 梅津 良昭からのコメント

この度、優れた基本性能を持つ Google の先進的な基盤モデル Gemma 3 27B をもとに、オンプレミス導入に最適な日本語 LLM を開発しました。エフサステクノロジーズ様による迅速な製品化により、Private AI Platform on PRIMERGY へのこの日本語 LLM の搭載が実現しました。3 社の技術と強みが結集した本製品を、リコージャパンの提供力で多くのお客様にお届けし、伴走支援することで課題解決に貢献できることを確信しております。

*1 Large Language Model（大規模言語モデル）：人間が話したり書いたりする言葉（自然言語）に存在する曖昧性やゆらぎを、文章の中で離れた単語間の関係までを把握し「文脈」を考慮した処理を可能にしているのが特徴。「自然文の質問への回答」や「文書の要約」といった処理を人間並みの精度で実行でき、学習も容易にできる技術。

*2 https://ai.google.dev/gemma/docs/core?hl=ja

*3 モデルマージ：複数の学習済みの LLM モデルを組み合わせて、より性能の高いモデルを作る新たな方法のこと。GPU のような大規模な計算リソースが不要で、より手軽にモデル開発ができるとして、近年注目されています。

*4 Chat Vector：指示追従能力を持つモデルからベースモデルのウェイトを差し引き、指示追従能力のみを抽出したベクトル。

*5 https://openai.com/ja-JP/index/introducing-gpt-oss/

*6 非推論モデル：学習済み知識から直接回答を生成する思考プロセスを持つモデル。推論のステップを省略するため、明確な指示を与えれば、迅速に回答生成が可能。

*7 初期応答性 Time to First Token: TTFT：ユーザーが AI にプロンプト（質問や指示）を入力してから、モデルが最初の出力テキスト（トークン）を生成し始めるまでにかかる時間を測定する応答速度の指標。ユーザー体験（UX）に直接影響する指標。

*8 PC サーバ：一般的なパソコン製品と共通の技術や仕様、部品などを用いて設計、製造されたサーバコンピュータ。サーバに比べて、一般的には安価に導入が可能。

*9 推論性能：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。

■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、推論性能強化により GPT-5 と同等の高性能な日本語大規模言語モデルを開発
https://jp.ricoh.com/release/2025/1010_1

リコー、金融業務特化型 LLM（大規模言語モデル）を開発、10 月末から個別提供開始
https://jp.ricoh.com/release/2025/1002_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、OpenAI の「gpt-oss-120B」をオンプレミス環境でいち早く検証完了、顧客への個別提供を開始
https://jp.ricoh.com/release/2025/0808_2

リコー、モデルマージによって GPT-4o 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/2025/0403_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

■関連情報：

エフサステクノロジー株式会社が提供するオンプレミス環境向けの対話型生成AIソリューション「Private AI Platform on PRIMERGY」
https://www.fsastech.com/ja-jp/products/primergy/solution/private-ai-platform/

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/