News Release

RICOH
imagine. change.

2024 年 8 月 21 日

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

～高速処理と省コストを同時実現し、環境負荷低減にも貢献～

株式会社リコー（社長執行役員：大山　晃）は、お客様の業務効率化や課題解決での活用を目的に、企業ごとのカスタマイズを容易に行える 700 億パラメータの大規模言語モデル[*1]（LLM）を開発[*2]しました。製造業で特に重視される日本語・英語・中国語に対応したほか、お客様のニーズに合わせてオンプレミス・クラウドのどちらの環境でも導入可能です。入力された文章を単語などの細かい単位に分割するトークナイザーの独自改良により、高速処理と省コストを実現し、環境負荷低減にも貢献します。ベンチマークツールを用いた検証[*3]の結果、優れた性能を確認しました（2024 年 8 月 9 日時点）。2024 年秋から、まずは日本国内のお客様より提供を開始し、今後海外のお客様への提供も目指します。

【プライベート LLM としてのユースケース例】

- 社内でも厳しいアクセス制御が求められる機密情報を取り扱う業務
  - -金融業：　融資審査業務等
  - -自治体：　行政サービス等
  - -流通・小売業：　顧客情報分析やマーケティング等
  - -教育・医療：　長時間労働が課題となっている教員や医師の文章作成等の周辺業務等
- 日本語・英語・中国語で日々更新される社内文書のデータを利活用する業務
  - -製造業：　RAG を活用した社内情報の検索や要約等

【リコーが開発した 700 億パラメータ LLM の特徴】

① 高い日本語性能を持ち、英語・中国語にも対応可能

リコーの LLM は、AI が自然言語の学習に利用するコーパスの選定や、誤記や重複の修正などのデータクレンジング、学習するデータの順序や割合を最適化するカリキュラム学習などリコー独自の方法で学習されています。これにより日本語による安定した回答を実現しました。また、AWS（Amazon Web Services）と共同で開発した学習スクリプトに基づいて訓練されており、日本語、英語、中国語の多様な表現を学習済みです。

さらに、独自開発を含む約 1 万6千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することにより、広範なタスクに適応する能力を獲得しました。これによりお客様のご要望に合わせてプライベート LLM を構築する際の追加学習で生じる破滅的忘却による性能低下を抑制し、高品質なプライベート LLM を開発することができます。

② トークナイザーの独自改良により、日本語の処理効率が同ベースモデルと比較して 43%向上

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?

リコーは、テキストをトークン[*4]に分割しLLMが理解できる形に変換するトークナイザーを独自に改良することで処理効率を向上させました。これにより、リソース削減、レスポンス時間の短縮、省コストを実現しました。LLM は処理に多くの電力が消費され環境負荷が大きいという社会課題に直面するなか、本技術は省エネルギー・環境負荷低減にもつながります。

③ **セキュリティを確保したオンプレミス環境で、学習～推論までご提供可能**

通常、700 億パラメータの LLM の運用や学習には、複数のサーバをネットワークで繋ぐ大規模なクラスタシステムが必要となります。リコーの LLM は独自の語彙置換技術やその他の最新技術を活用することによりモデルサイズを保ったまま学習が可能です。セキュリティ面でデータを自社内で保有したいお客様向けには、お客様先のクローズドな環境下での機密情報含めた追加学習が可能です。

④ **従来手法の開発と比較し、およそ 50%のコスト低減および最大 25%の電力消費量の削減を実現**

「AWS LLM 開発支援プログラム」と「AWS 生成 AI イノベーションセンター（AWS Generative AI Innovation Center）」によるサポート提供のもと、AWS Trainium アクセラレーターを搭載した Amazon Elastic Compute Cloud Trn1 インスタンスを利用することで、効率的な開発を実現しました。お客様向けカスタムLLMを開発する際にも、より安価・短納期でのご提供が可能です。また、学習に際して Trn1 インスタンスを活用することで、同等のアクセラレーテッドコンピューティング EC2 インスタンスよりもエネルギー効率を最大 25% 改善しました。

【評価結果（ELYZA-tasks-100）】

複雑な指示・タスクを含む代表的な日本語のベンチマーク「ELYZA-tasks-100」において、リコーの LLM は平均で 4 を超える高いスコアを示しました。また、比較した他の LLM はタスクによって英語で回答するケースが見られましたが、リコーの LLM は全てのタスクに対して日本語で回答して高い安定性を示しました。さらに、回答速度の面でも他の LLM を大きく上回り、トークナイザーの改良の効果を確認しました。

企業/組織	モデル	ELYZA-tasks-100		
		スコア	英語で回答されたタスクの割合 [%]	総回答時間 [秒] (リコーモデルを基準とした割合)
OpenAI	gpt-4-0613	4.45	-	-
rinna	llama-3-youko-70b-instruct	4.11	3	1,070 (1.5)
Tokyotech	Llama-3-Swallow-70B-Instruct-v0.1	3.88	7	819 (1.1)
Meta	Meta-Llama-3-70B-Instruct	3.63	>70	1,242 (1.7)
Ricoh	Llama-3-Ricoh-70B-Instruct	4.02	0	738 (1.0)

ベンチマークツール（ELYZA-tasks-100）における他モデルとの比較結果[*3]（リコーは最下段）

【リコーの LLM 開発の背景】

労働人口減少や高齢化を背景に、AI を活用した生産性向上や付加価値の高い働き方が企業成長の課題となっており、その課題解決の手段として、多くの企業が AI の業務活用に注目しています。しかし、

AI を実際の業務に適用するためには、企業固有の用語や言い回しなどを含む大量のテキストデータを LLM に学習させ、その企業独自の AI モデル（プライベート LLM）を作成する必要があります。

リコーは国内でもトップクラスの LLM の開発・学習技術をベースに、企業向けプライベート LLM の提供や、社内文書の活用を後押しする RAG の導入支援等、様々な AI ソリューションの提案が可能です。

製造業や金融業などセキュリティ要件の高い業界では、オンプレサーバに導入可能な省リソースでありながら、700 億パラメータの規模で、対応言語は日英中を選択可能なプライベート LLM に対する強い要望がありました。こうした要望を踏まえ、リコーが開発した高性能な LLM に企業独自の情報や知識を取り入れることで、お客様ごとの業種・業務に合わせた高精度な AI モデル（プライベート LLM）を、低コスト・短期間で容易に構築することが可能です。

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。

＊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 このたびリコーが開発した LLM は、米 Meta Platforms 社が提供する「Meta-Llama-3-70B」の日本語性能を向上させた「Llama-3-Swallow-70B」をベースモデルに採用し、日本語と英語、中国語のオープンコーパス（自然言語の文章や使い方を大規模に収集し、一般公開されているデータセット）を追加学習させて開発したものです。

＊3 2024 年 8 月 9 日時点の評価結果。「Meta-Llama-3-70B」と公開されているその日本語継続事前学習モデルを比較対象として選定。参考として「GPT-4」の「スコア」も併記。「スコア」の算出に際して、生成文の評価には「GPT-4」（gpt-4-0613）を使用し、英語での回答による減点は行っていない。「総回答時間」は NVIDIA DGX H100 において GPU を 2 枚用いて計測。「英語で回答されたタスクの割合」は 100 タスクのうち英語で回答されたものの割合。但し、「Meta-Llama-3-70B-Instruct」は回答の大部分が英語または英語交じりの日本語だったため概算値。

＊4 LLM はテキストデータを「トークン」と呼ばれる単位で処理します。トークンとは、単語、文字セット、または単語と句読点の組み合わせです。

■関連ニュース

インストラクションチューニング済みの 130 億パラメータの日本語 LLM を開発
https://jp.ricoh.com/release/2024/0603_1

日本語精度が高い 130 億パラメータの大規模言語モデル（LLM）を開発
https://jp.ricoh.com/release/2024/0131_1

| 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（2024年3月期グループ連結売上高2兆3,489億円）。

“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/