

2024年6月3日

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 ～理化学研究所との共同研究のデータや、リコー独自開発のデータを追加学習させ、指示追従性能や要約性能が向上～

株式会社リコー（社長執行役員：大山 晃）は、国立研究開発法人理化学研究所 革新知能統合研究センター 言語情報アクセス技術チーム（以下、理研 AIP）が主催する日本語インストラクションデータ作成プロジェクトに参画しています。リコーは、同チームとの共同開発で得られたインストラクションデータをリコー製130億パラメータの日本語LLM^{*1}に追加学習させ、LLMの指示追従性能が向上するという結果を得ました。また、リコー独自開発のインストラクションデータ^{*2}を追加学習させた結果においても、指示追従性能^{*3}の向上を確認し、要約タスクでの優位性を確認しました。

今回、インストラクションデータ「ichikara-instruction」（10,329件）を用いてリコー製LLMにインストラクションチューニング^{*4}を行いました。結果、複雑な指示・タスクを含む代表的なベンチマーク「ELYZA-tasks-100」において、チューニング前と比較し、指示追従性能の大幅なスコア向上が確認できました。また、リコーが独自開発した3,556件のインストラクションデータを用いたチューニング結果でも、同ベンチマークにおいて、同様にスコアが大きく向上しました。

これらの結果から、「ichikara-instruction」はインストラクションデータとして高品質なデータセットであること、また、リコー製インストラクションデータにおいても高スコアが得られたことから、LLMの性能向上にはデータ量だけでなく、データの品質が重要だということが示唆されました。（表1）

モデル	スコア
リコー製130億パラメータLLM（インストラクションチューニングなし）	1.19
リコー製130億パラメータLLM + 理研 AIPデータセット（10,329件）	3.02
リコー製130億パラメータLLM + リコー製データセット（3,556件）	2.87

表1. 評価結果（ELYZA-tasks-100）

また、要約タスクを独自評価^{*5}したところ、特に長文要約においては、リコー製データセットの優位性を確認できました。（表2）

AIによる要約生成はお客様のニーズが高く、リコーが強化していく領域です。リコーは継続的にデータ開発を進めており、2024年5月末時点では、5,000件超のインストラクションデータの開発を完了しています。今後、これらをリコーが提供するさまざまなAIソリューションに活用することで、より高品質なサービスの提供を目指します。

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?

モデル	ニュース要約	論文要約
リコー製130億パラメータLLM + 理研 AIPデータセット (10,329件)	24.82	25.01
リコー製130億パラメータLLM + リコー製データセット (3,556件)	24.63	30.42

表2. 評価結果 (Rouge-Lでの自動評価) *5

(背景)

労働人口減少や高齢化を背景に、生産性向上や付加価値の高い働き方の実現に向けて、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、その業種・業務の情報や、企業固有の用語や言い回しなどを含む大量のデータをLLMに追加学習させ、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。

(リコー製 LLM の特徴)

リコーでは、お客様の想定用途に合わせてさまざまなデータ(企業独自の情報や知識を含む)を使ってドメイン適用された高精度なAIモデル(プライベートLLM)の個別開発を行っています。リコー製LLMは、日本企業の業務での活用を目的に開発され、企業ごとのカスタマイズを容易に行うことができます。独自の学習上の工夫が組み込まれており、日本語としての文法や回答が正確で日本語精度が高く、日本企業が持つ情報資産の活用に適したモデルになっています。特にNLI(自然言語推論能力)において高性能という評価結果が出ています。2024年4月から、プライベートLLMをクラウド環境で提供開始しています。

今後、リコーはインストラクションデータの品質をさらに向上させ、インストラクションチューニング済みの高精度なプライベートLLMをご提供していくことで、お客様のAI活用を支援します。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。

*1 LLM: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 リコー独自開発のインストラクションデータ:2024年3月末時点で3,556件の開発を完了し、評価実施。その後、開発を継続し、24年5月末には5,091件のデータを開発。

*3 指示追従性能: ユーザの指示や質問に対して自然な回答ができる能力。ユーザの意図を適切に理解し、それに応じた対応をできるかどうかを測る指標。

*4 インストラクションチューニング:タスクに対して、指示(プロンプト)と正しい回答(インストラクションデータ)をセットで与えて受け答えを学習させる手法。

*5 独自評価:評価データセットはリコーで独自に準備。ニュース要約:200件(要約対象の文字数:200~3000文字程度)。論文要約:100件(要約対象の文字数:2500~5500文字程度)。

■ 関連情報

理研 AIP :ichikara-instruction: LLM のための日本語インストラクションデータ

<https://liat->

aip.sakura.ne.jp/wp/llm%E3%81%AE%E3%81%9F%E3%82%81%E3%81%AE%E6%97%A5%E6%9C%AC%E8%AA%9E%E3%82%A4%E3%83%B3%E3%82%B9%E3%83%88%E3%83%A9%E3%82%AF%E3%82%B7%E3%83%A7%E3%83%B3%E3%83%87%E3%83%BC%E3%82%BF%E4%BD%9C%E6%88%90/

■ 関連ニュース

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

※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、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/>