2024年10月29日
株式会社オルツ

ニュースリリース

オルツの「LHTM-OPT2」、日本語RAG（検索拡張生成）で軽量型LLMとして世界最高の精度と推論速度を実現

～国内一の日本語推論能力を有する軽量型言語モデルで日本語AIの新たなユースケースを創出～

株式会社オルツ（本社：東京都港区、代表取締役：米倉 千貴、以下、オルツ）は、当社が開発する軽量大規模言語モデル「LHTM-OPT」シリーズの最新バージョン「LHTM-OPT2（ラートム・オプト2）」をリリースいたしました。「LHTM-OPT2」は、RAG（検索拡張生成）の精度を最適化する軽量型LLMであり、日本語RAG精度において、軽量型LLMで世界最高精度※1 を達成したことをお知らせいたします。

※ご参考：LHTM-OPTについて：https://alt.ai/news/news-2300/
※ご参考：LHTM-OPT、AWS Marketplace上に日本語LLMとして世界初公開：https://alt.ai/news/news-2553/

「LHTM-OPT」は、小規模GPUマシンで実用的な、パラメータ数が最適化された新たな軽量型大規模言語モデルです。この度、「LHTM-OPT」シリーズの最新バージョン「LHTM-OPT2」の日本語RAG精度にあたり、オルツが独自に開発したWikipediaデータからのRAG質問・回答のデータセット（Wiki RAGデータセット）と、東京大学入学試験の国語科目データセットを用いて評価を行いました。

Wiki RAGデータセットを作成する手法は、日本語Wikipediaから特定の段落を抽出し、その段落に基づく質問を生成し、[段落、 質問、 正解] の３つ組を作成することです。このデータは、専門家の手によって再度確認、修正を行うことで、高品質のRAGベンチマークになります。
また、東京大学入学試験の国語科目データセット評価では、東京大学入学試験※2 における国語大学科目問題の前提テキスト（段落）とその設問をRAGの入力とし、LLMがその段落と設問から生成した回答を専門家が評価しました。
評価結果では、Wikipedia RAGデータセットでは、「LHTM-OPT2」が、GPT-4oと同等レベルの精度（LHTM-OPT2：91.0%、GPT-4o：90.8%）を達成しました。また、東大入試国語科目におけるRAGに関する質問では、「LHTM-OPT2」が、GPT-4oの94%の精度を達成しました。

さらに、RAG評価においては、国内の全ての軽量型LLM（パラメータ数が10B以下のLLM）を上回る高い精度を達成し、「JGLUE（Japanese General Language Understanding Evaluation）」ベンチマークや「Japanese MT-Bench（MTベンチ）※3」でも、軽量型LLMとしての最高スコアを記録しました。
推論速度に関しては、SambaNova社の協力を得て、日本語推論において平均速度500TPS（トークン／秒）、最大速度796TPSを確認しました。この速度は、日本語LLM推論速度の最高記録※4 です。

※1 世界最高精度・最高スコア：
「弊社が独自に開発した日本語Wikipediaデータによる RAGデータセット」というLLM・RAGベンチマークによる評価で、軽量型LLMとして、パラメータ数が10B以下のモデルの中での比較として国内トップスコアを達成。
（2024年10月15日時点。自社調べ）

※2 これまでの東京大学入学試験問題及び解答等：https://www.u-tokyo.ac.jp/ja/admissions/undergraduate/e01_04.html

※3 「Japanese MT-Bench」はStability AI社が提供しているベンチマークテストです。2024年10月15日に性能評価した結果、軽量型LLMとして最高点の評価を得ました。ベンチマークテストとは、定められた基準を元にその性能を測定する方法で、「Japanese MT-Bench」はGPT-4を評価者としたものです。

※4 最高記録：
ArtificialAnalysis.ai によると、既存のLLMでは、Cerebrasが最速の2148 TPS で、SambaNovaが2番の速度（462 TPS)。ただし、日本語専用のLLM超高速推論は、オルツとSambaNova社が初めて実現しました。
（2024年10月15日時点。自社調べ）
https://artificialanalysis.ai/#providers

オルツは、今後も「LHTM-OPT」シリーズの開発と提供を通じて、より高精度で効率的な言語モデルの開発を進め、世界水準の技術で "アジアにおけるOpenAI" の地位確立を目指します。そして、顧客へ最高品質のソリューションを提供することで、日本企業の労働生産性向上に資する取組みを推進してまいります。

▶LHTM-2／LHTM-OPT／GPT など大規模言語処理ソリューションに関するお問い合わせ先
https://alt.ai/aiprojects/gpt/

■株式会社オルツについて
2014年11月に設立された当社は、「P.A.I.」（パーソナル人工知能）、AIクローンをつくり出すことによって「人の非生産的労働からの解放を目指す」企業です。また、AIの対話エンジンの開発から生まれた音声認識テクノロジーを活用したCommunication Intelligence「AI GIJIROKU」を筆頭として、そのほか、PoC（Proof of Concept、概念実証）により様々なビジネス上の課題を切り口としてソリューション展開するプロダクト（「altBRAIN」、「AIコールセンター」、「CLONEdev」など）を開発・提供しています。
https://alt.ai/

商号　　：株式会社オルツ
代表者　：代表取締役社長　米倉 千貴
事業内容：デジタルクローン、P.A.I.の開発を最終目的とした要素技術の研究開発とそれらを応用した
　　製品群（Communication Intelligence「AI GIJIROKU」等）の展開、AIソリューションの提供
創業　　：2014年11月
所在地　：東京都港区六本木七丁目15番７号

＜報道関係者からのお問い合わせ先＞

株式会社オルツ　広報 西澤
e-mail：press@alt.ai

<アライアンスに関するお問い合わせ先>

株式会社オルツでは、IT・金融・建設・物流・メディア・製造・小売・サービス業など、
ジャンルを問わずAIソリューションの提供および支援を行っております。
お気軽にお問い合わせください。

株式会社オルツ 事業本部 小村
e-mail：gptsolutions@alt.ai