



Press Release

2024 年 5 月 10 日

東京工業大学

東北大学

富士通株式会社

理化学研究所

名古屋大学

株式会社サイバーエージェント

Kotoba Technologies Inc.

スーパーコンピュータ「富岳」で学習した 大規模言語モデル「Fugaku-LLM」を公開 ー日本語能力に長け、研究・ビジネス利用にも期待ー

【要点】

- 日本の計算機技術を用いて開発した日本語能力に優れた大規模言語モデルを公開
- スーパーコンピュータ「富岳」の性能を最大限に活用した分散並列学習を実現
- AI 基盤モデルを科学研究に活用する「AI for Science」など革新的な研究やビジネスにつながる

【概要】

東京工業大学 学術国際情報センターの横田理央教授の研究チームと東北大学 大学院情報科学研究科の坂口慶祐准教授、富士通株式会社 人工知能研究所の白幡晃一シニアプロジェクトディレクター、理化学研究所 の Mohamed Wahib（モハメド・ワヒブ）チームリーダー、名古屋大学 大学院工学研究科の西口浩司准教授、株式会社サイバーエージェント AI 事業本部 AI Lab の佐々木翔大リサーチサイエンティスト、Kotoba Technologies Inc. の小島熙之 CEO は、理化学研究所のスーパーコンピュータ「富岳」を用いて学習した日本語能力に優れた**大規模言語モデル**（用語 1）「Fugaku-LLM」を 2024 年 5 月 10 日に公開した。

今回、深層学習フレームワークを「富岳」に移植して **Transformer**（用語 2）の性能を「富岳」上で最適化するとともに、並列分散学習手法を開発し適用することで、「富岳」を用いて大規模言語モデルの学習を行う際の**演算速度を 6 倍**（参考文献 1 と比較）に高速化した。さらに、「富岳」向けに **Tofu インターコネクト D**（用語 3）上での集団通信の最適化を行うことにより、**通信速度を 3 倍**（参考文献 2 と比較）高速化することに成功した。これにより「富岳」の CPU を用いて、現実的な時間内で大規模言語モデルを学習することが可能になった。

Fugaku-LLM は国内で多く開発されている 70 億**パラメータ**（用語 4）より一般に高性能、かつ現在の計算機環境下で扱いやすい 130 億パラメータのモデルとなっている。今回の学習では、サイバーエージェントが収集した独自の日本語学習データと英語のデー

タなどを用いているため、透明性と安全性を確保しつつ日本語性能にも優れている。国産かつ独自のデータで学習を行っているオープンなモデルの中では日本語ベンチマーク **Japanese MT-Bench**（用語 5）で最高性能を達成し、特に人文社会系のタスクでは高いベンチマーク性能を発揮する。

Fugaku-LLM は、**GitHub**（用語 6）や **Hugging Face**（用語 7）を通じ公開しており、ライセンスに従う限りにおいては、研究および商業目的での利用が可能である。

今後、多くの研究者や技術者が基盤モデルの改善や新たな応用研究に参画することで、さらに効率的な学習方法が創出され、科学シミュレーションと生成 AI の連携や、数千の AI によるバーチャルコミュニティの社会シミュレーションなど、次世代の革新的な研究やビジネスでの応用につながることを期待される。

（公開リンク）

モデル：<https://huggingface.co/Fugaku-LLM/Fugaku-LLM-13B>

ソースコード：<https://github.com/Fugaku-LLM/DeepSpeedFugaku>

●背景

近年、米国を中心に大規模言語モデル(LLM)の開発が活発に行われ、研究開発、経済社会、安全保障などあらゆる場面において大きな変革が起きている。米国以外の国々においても、自国で LLM を開発するために莫大な人的資源や計算資源を投入している。日本においても、AI 研究のための計算資源として、日本のスーパーコンピュータのフラッグシップシステムである「富岳」に寄せられる期待は大きく、「富岳」における大規模な分散並列計算を実施するための環境整備が求められていた。

そこで、東京工業大学、東北大学、富士通、理化学研究所は、2023 年 5 月より大規模言語モデルの共同研究開発を開始し、2023 年 8 月からは、名古屋大学、サイバーエージェント、Kotoba Technologies が参加した。

●各機関・企業の役割

東京工業大学：全体総括、大規模言語モデルの並列化および通信の高速化（3 種類の並列化を組み合わせた通信性能の最適化、Tofu インターコネクト D 上での集団通信の高速化）

東北大学：学習用データの収集、学習モデルの選択

富士通：演算高速化と通信高速化（Tofu インターコネクト D 上での集団通信の高速化、パイプライン並列の性能最適化）、事前学習と学習後のファインチューニング

理化学研究所：大規模言語モデルの分散並列化・通信の高速化（Tofu インターコネクト D 上での集団通信の高速化）

名古屋大学：3D 形状生成 AI への Fugaku-LLM の応用方法の検討

サイバーエージェント：学習用データの提供

Kotoba Technologies：深層学習フレームワークの「富岳」への移植



図1 理化学研究所 スーパーコンピュータ「富岳」

●研究成果

1. 「富岳」における大規模言語モデル学習の計算性能を大幅に向上

本研究では、「富岳」を用いることで、大規模言語モデルを学習する際の演算速度を既存技術の 6 倍、通信速度を 3 倍に高速化することに成功した。演算高速化については Transformer の性能を「富岳」上で最適化するため、深層学習フレームワーク Megatron-DeepSpeed を「富岳」へ移植、密行列積ライブラリの Transformer 向け高速化を行った。通信高速化については 3 種類の並列化を組み合わせた「富岳」向け通信性能最適化と Tofu インターコネクト D 上での集団通信の高速化を行った。

通常、大規模言語モデルの学習には **GPU**（用語 8）が用いられるが、大規模言語モデルの学習のために世界中で GPU 不足が起きており、最新の GPU を大量に入手することは困難となっている。そのような中、GPU ではなく富士通製の国産 CPU を中央演算処理装置とした「富岳」を用いて大規模言語モデルを学習できたことは、日本の半導体技術の活用や、経済安全保障の観点からも重要な成果である。

また、今回の取り組みで得られた知見は「富岳」の後の次世代計算基盤の設計に活かされうるものであり、AI 分野における日本の優位性確立に寄与していく。

2. 透明性と安全性を担保し、使いやすく日本語性能に優れた 130 億パラメータの大規模言語モデル

2023 年には国内の企業で多くの大規模言語モデルが開発されたが、その多くは 70 億パラメータのモデルとなっている。大規模言語モデルは一般にパラメータ数が増えるに

従ってその性能も向上するため、今回開発した 130 億パラメータの Fugaku-LLM は高性能なモデルと言える。国外ではさらに大規模なモデルも開発されているが、大規模な言語モデルでは、使用する際にも大規模な計算資源が必要になるため、あまりにもパラメータ数が大きいものは使用しにくい。2024 年現在の計算機環境と照らし合わせ、Fugaku-LLM では高性能かつバランスの取れた 130 億パラメータとした。

また、日本語が扱える多くのモデルは、国外で開発されたオープンなモデルに対してさらに日本語データを学習させる**継続学習**（用語 9）の方法を採っている。それに対し、今回開発した Fugaku-LLM は一から独自のデータを用い学習を行っているため、学習の全学習工程を把握でき、透明性と安全性の観点から優れている。

「富岳」の 13,824 台の計算ノードを用いて、約 4,000 億トークンを学習したが、学習データの約 60%が日本語コンテンツであり、その他英語、数学、コードと組み合わせて学習した。他言語での学習をベースに日本語を継続学習するのではなく、元から多くの情報を日本語で学習した Fugaku-LLM は、Japanese MT-Bench で平均スコア 5.5 と、国産で独自のデータで学習を行っているオープンなモデルの中では最高性能を達成した。特に人文社会系のタスクでは 9.18 と高いベンチマーク性能を発揮しており、敬語など日本語の特徴を踏まえた自然な対話を行えることなどが期待される。

Fugaku-LLM

チャット履歴をクリア

🗨️ なにかお困りですか？

🔴 日本語の敬語の使い分け方と、ビジネスシーンでの敬語使用の重要性について説明してください。

📄 日本語の敬語は、話し手と聞き手の社会的地位や関係によって、さまざまな種類があります。敬語は、尊敬や丁寧な気持ちを表現するために使われます。

ビジネスシーンでは、上司、同僚、顧客など、すべての人に対して敬語を使うのが一般的です。敬語には、丁寧語、謙譲語、尊敬語の3種類があります。

丁寧語は、話し手が聞き手に対して丁寧な言葉や態度をとることで、相手を敬う表現です。丁寧語は、語尾に「です」「ます」をつけて丁寧な印象を与えるのが特徴です。

謙譲語は、話し手が自分の立場を低くすることで、聞き手に対して敬意を表す表現です。謙譲語では、話し手は自分のことをへりくだって表現し、相手を立てます。

尊敬語は、話し手が聞き手のことを高めることで、尊敬の気持ちを表す表現です。尊敬語では、話し手は相手のことを敬い、丁寧な言葉や丁寧な言い回しを使います。

日本のビジネス社会では、敬語を正しく使うことが重要です。敬語を適切に使い分けすることで、相手に敬意を表し、良好な人間関係を築くことができます。また、敬語を使うことで、日本のビジネス文化における礼儀や作法を守ることができます。

図 2 Fugaku-LLM デモの様子

●今後の展開

研究者やエンジニアが大規模言語モデルの開発に活用できるよう、今回の取り組みで得られた研究成果を GitHub や Hugging Face を通じ公開している。ライセンスで定めた条件下で、誰もが研究および商業目的での利用が可能である。さらに、富士通は Fugaku-LLM を、富士通の先端技術が無償で試せる「Fujitsu Research Portal」を通じて 2024 年 5 月 10 日より提供開始する。

公開されたモデルを用い、多くの研究者や技術者が基盤モデルの改善や新たな応用研究に参画することで、効率的な学習方法や言語モデルが創出され、科学シミュレーションと生成 AI の連携による科学研究サイクルの自動化のように AI 基盤モデルを科学研究に活用する「AI for Science」、数千を超す AI によるバーチャルコミュニティの社会シミュレーションなど、次世代の革新的な研究やビジネスの成果につながることを期待される。

●付記

本成果は、「富岳」政策対応課題「『富岳』を活用した大規模言語モデル分散並列学習手法の開発」（課題番号：hp230254）によるものである。

【用語説明】

- (1) **大規模言語モデル**：テキストの現れやすさをモデル化したもので、与えられた文脈（問いかけ）に対して続くテキスト（応答）を予測できる。
- (2) **Transformer**：単語の系列などを変換するためのニューラルネットワークで、現在の大規模言語モデルでは最も多く用いられている。2017 年 6 月に Google が発表した「Attention Is All You Need」という論文で登場した深層学習モデルであり、主に自然言語処理の分野で使用される。
- (3) **Tofu インターコネクト D**：「富岳」のノード間接続に用いられる 6 次元トーラスのトポロジーを持つ高速なネットワーク。Tofu とは、Torus fusion を短縮した名称であり、記号「D」は、高密度(High-density)を意味する。
- (4) **パラメータ**：大規模言語モデルなどのニューラルネットワークの規模を表す指標の一つ。パラメータが多いほど高性能なモデルになるが、学習にはその分多くのデータが必要になる。
- (5) **Japanese MT-Bench**：Stability AI 社が提供している日本語のベンチマークテスト
- (6) **GitHub**：世界中でオープンソースソフトウェアの公開に使われているプラットフォーム。<https://github.com/>
- (7) **Hugging Face**：世界中でオープンソースソフトウェアの公開に使われているプラットフォーム。<https://huggingface.co/>
- (8) **GPU**：元々は描画のための演算加速器として生産されていたが、近年では深層学習の高速化に用いられている。
- (9) **継続学習**：すでに学習されている大規模言語モデルに対し、追加で学習を行う手法。異なる言語やドメインで言語モデルを活用するときに用いられる。

【参考文献】

- (1) COOL Chips 27(2024 年 4 月 17 日～19 日)発表「Implementation of Batch Matrix Multiplication for Large Language Model Training on A64FX CPUs」
- (2) 第 193 回 HPC 研究発表会(2024 年 3 月 18 日～19 日)発表「富岳上の大規模機械学習における All-reduce 通信の高速化」