NEWS RELEASE

大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
National Institute of Informatics

2024 年（令和 6 年）4 月 30 日

LLM勉強会
LLM-jp

大規模言語モデル「LLM-jp-13B v2.0」を構築
～NII 主宰 LLM 勉強会（LLM-jp）が「LLM-jp-13B」の後続モデルとその構築に使用した全リソースを公開～

大学共同利用機関法人 情報・システム研究機構 国立情報学研究所（ＮＩＩ、所長：黒橋 禎夫、東京都千代田区）は、昨年 5 月から、自然言語処理及び計算機システムの研究者を中心として、大学・企業等から 1,200 名以上（2024 年 4 月末現在）が参加する LLM 勉強会（LLM-jp）を主宰しています。同年 10 月には、初期の開発成果として、パラメータ数(*1)130 億の大規模言語モデル（LLM）「LLM-jp-13B v1.0」を公開しました。

この開発経験を踏まえ、今年 1 月から、計算資源としてデータ活用社会創成プラットフォーム mdx(*2)を活用し、コーパス(*3)やモデル構造の改善、安全性を考慮したチューニングの導入などを行い、後続モデルとして「LLM-jp-13B v2.0」を構築しました。本日、同 LLM を公開しましたのでお知らせします。また、今後のアカデミアや産業界の研究開発に資するため、コーパスを含む、同 LLM の構築に使用したすべてのリソースもあわせて公開しています。

LLM の性能が向上し、社会での利活用が本格化するにしたがって、LLM の透明性・信頼性の確保、安全性の配慮がより一層重要となります。今回のモデルや今後構築するモデルを活用してそれらの研究を進め、LLM 研究開発の促進に貢献します。

1． 今回構築した LLM「LLM-jp-13B v2.0」の概要

（1） 「LLM-jp-13B v1.0」からの主な変更点

- 日本語ウェブコーパスの改善：大規模ウェブアーカイブ Common Crawl の全量から「Uzushio」を用いて日本語テキストを抽出、フィルタリングしたコーパス「日本語 Common Crawl」を新たに構築して使用。「LLM-jp-13B v1.0」の学習に使用した日本語ウェブコーパス「日本語 mC4」と比較して品質が大幅に改善。
- モデルアーキテクチャの改善：種々の改善が加わった現代的なモデルアーキテクチャを採用。最大トークン長を 2,048 から 4,096 に拡張し、より長い文脈を処理できるように改善。
- 安全性に配慮したチューニング：安全性の観点に基づくデータセットを新たに構築し、モデルのチューニングに使用。

（2） 利用計算資源等

- データ活用社会創成プラットフォーム mdx（16 ノード、NVIDIA A100 GPU 128 枚）を利用

- mdx の利用料金は NII、理化学研究所革新知能統合研究センター（AIP）、学際大規模情報基盤共同利用・共同研究拠点（JHPCN）の 3 組織が負担
- モデル構築には NVIDIA が開発している LLM の学習フレームワーク「Megatron-LM」を利用
- モデル構築時の評価指標の監視やログの保存には実験管理プラットフォーム「Weights & Biases」を利用

（3） モデル学習用コーパス

- 学習データ量：約 2,600 億トークン
  - 日本語：約 1,300 億トークン（日本語 Common Crawl、日本語 Wikipedia）
  - 英語：約 1,200 億トークン（英語 Pile、英語 Wikipedia）
  - プログラムコード：約 100 億トークン

（4） モデル

- パラメータ数：130 億個（13B）
- モデルアーキテクチャ：LLaMA ベース
- 最大トークン長：4,096

（5） チューニング

- 日本語インストラクションデータおよび英語インストラクションデータの和訳データ 8 種類を用いてチューニング実験を実施

（6） 評価

- 既存の日本語言語資源を利用した 22 種類の評価データを整備し、横断的に評価を行うフレームワーク「llm-jp-eval v1.3.0」を構築・使用
- GPT-4 による生成テキストの自動評価フレームワーク「日本語 Vicuna QA」および「日本語 MT Bench」を使用
- 人手による生成テキストの安全性に関する評価を実施
- いずれの評価においても「LLM-jp-13B v1.0」と比較して大幅な性能向上を確認

（7） 開発モデル・ツール・コーパスの公開 URL

https://llm-jp.nii.ac.jp/release

注：今回公開するモデルは、安全性の観点に基づくチューニングを行ったものではありますが、まだ研究開発の初期段階のものであり、そのまま実用的なサービスに供することを想定しているものではありません。

２．　LLM 勉強会（LLM-jp）の概要

（1）　NII が主宰する LLM-jp では、自然言語処理及び計算機システムの研究者を中心として、大学・企業等から 1,200 名以上（2024 年 4 月末現在）が集まり、ハイブリッド会議、オンライン会議、Slack 等を活用して LLM の研究開発について情報共有を行うとともに、共同で LLM 構築等の研究開発を行っています。具体的には、以下の目的で活動しています。

- オープンかつ日本語に強い LLM の構築とそれに関連する研究開発の推進
- 上記に関心のある自然言語処理および関連分野の研究者によるモデル構築の知見や最近の研究の発展についての定期的な情報交換
- データ・計算資源等の共有を前提とした組織横断的な研究者間の連携の促進
- モデル・ツール・技術資料等の成果物の公開

（2）　LLM 構築にあたっては、「コーパス構築 WG」「モデル構築 WG」「チューニング・評価 WG」「安全性 WG」「マルチモーダル WG」等を設置し、それぞれ、早稲田大学　河原大輔教授、東北大学　鈴木潤教授、東京大学　宮尾祐介教授、国立情報学研究所　関根聡特任教授、東京工業大学　岡崎直観教授を中心に研究開発活動に取り組んでいます。このほか、東京大学　田浦健次朗教授、空閑洋平准教授（計算資源 mdx の利用）、東京工業大学　横田理央教授（並列計算手法等）等、多数の方々の貢献により、活動を進めています。

（3）　詳細については、ホームページ https://llm-jp.nii.ac.jp/ をご参照ください。

３．　今後の予定

LLM を社会で利活用していく上では LLM の透明性・信頼性の確保が必要であり、モデルの高度化に伴い、安全性の配慮もより重要となります。そのため、NII は、文部科学省の補助事業「生成 AI モデルの透明性・信頼性の確保に向けた研究開発拠点形成」（https://www.mext.go.jp/content/20240118-ope_dev03-000033586-11.pdf の p.7）の支援を受け 2024 年 4 月に大規模言語モデル研究開発センターを設置しました。

今回公開したモデルや、現在経済産業省の「GENIAC」の支援のもと構築している 1750 億パラメータ規模の LLM などの今後構築するモデルを活用してそれらの研究を進め、LLM 研究開発の促進に貢献します。

〈メディアの皆様からのお問い合わせ先〉
大学共同利用機関法人 情報・システム研究機構 国立情報学研究所
総務部企画課 広報チーム
TEL:03-4212-2164　E-mail：media@nii.ac.jp

(*1) **パラメータ数**：大規模言語モデルは言語を学習した大規模なニューラルネットワークで、パラメータはニューラルネットワークの規模を示す指標のひとつ。パラメータ数が多いほど高い性能であるといわれている。

(*2) **データ活用社会創成プラットフォーム mdx**：9 大学 2 研究所が連合して共同運営する、データ活用にフォーカスした高性能仮想化環境。研究環境を用途に合わせてオンデマンドで短時間に構築・拡張・融合できる、データ収集・集積・解析のためのプラットフォーム。今回のモデル構築に当たっては、NII のほか、理化学研究所革新知能統合研究センター（AIP）と学際大規模情報基盤共同利用・共同研究拠点（JHPCN）の資金拠出により利用。

(*3) **コーパス**：自然言語の文章を構造化し大規模に集積したデータベース。