

オープンかつ日本語に強いGPT-3級大規模言語モデルの構築・事業成果概要

実施者	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）
事業概要	1.35兆トークンのコーパスを用いて1750億パラメータ規模の大規模言語モデルを事前学習し、高い日本語性能を持つモデルを構築する。コーパスやモデルはオープンな形で公開する。

事前学習用コーパス構築

日本語、英語、プログラムコード、中国語、韓国語を対象に1兆3500億トークンのコーパスを開発

事前学習用コーパスに対するトークナイザなどのツール開発

単一のトークナイザモデルに上記各言語の語彙を順次追加。約10万トークンの語彙をもち、従来よりもトークン化効率が向上

基盤モデル構築

1720億パラメータのモデルを開発し、事業期間内に約4000億トークンの事前学習を完了(当初予定していた1兆3500億トークンの事前学習は他の計算資源上で継続し、終了後のモデルを公開予定)

社会実装イメージ

オープンなデータやモデルの公開を通して、日本の研究コミュニティ知識基盤のかさ上げ、生成AI開発力強化、および革新的なイノベーションの創出に貢献

1720億パラメータ・モデルの事前学習データを公開

<https://gitlab.llm-jp.nii.ac.jp/datasets/llm-jp-corpus-v3>

1720億パラメータ・モデル(7000億トークン学習済み)を公開

<https://huggingface.co/llm-jp/llm-jp-3-172b-beta1>

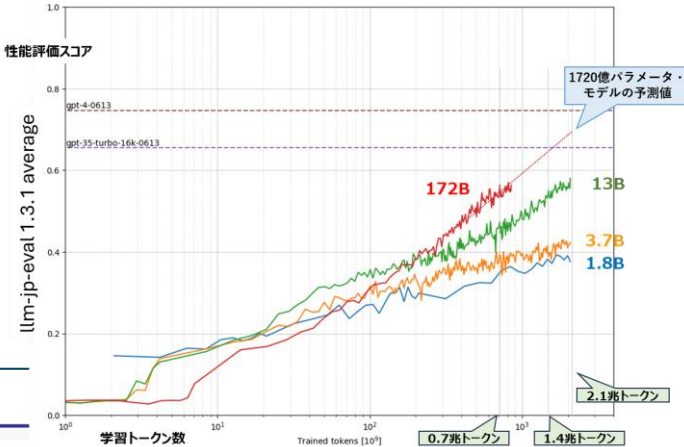
130億パラメータ・モデルなど、その他の規模のモデルも公開

<https://huggingface.co/llm-jp/llm-jp-3-13b>

事業成果

- 3兆6000億トークンのコーパスを収集・構築し、2兆1000億トークンの事前学習用コーパスを開発
- 1720億パラメータの大規模言語モデル開発。事業終了時点(8月15日)で2兆1000億トークンの事前学習の約1/6を終え、現在も別の計算資源を利用してモデル構築を継続中
- 事前学習時に大きな影響をもつ学習パラメータ(adam-eps)を特定

右図は現在事前学習中の1720億パラメータ・モデル(赤線)の性能評価結果を示したものの。事前学習終了時には130億パラメータ(緑線)やGPT-3.5 turboなどのモデルを超え、GPT-4に迫る性能を達成できると予想



ニュースリリース

2024/09/17

約1720億パラメータ（GPT-3級）の大規模言語モデルのフルスクラッチ学習を行い、プレビュー版「LLM-jp-3 172B beta1」を公開
～学習データを含めすべてオープンにしたモデルとしては世界最大～

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、東京都千代田区の大規模言語モデル研究開発センター（LLMC）は、主宰するLLM勉強会（LLM-jp）の成果として、これまでのデータ活用社会創成プラットフォームmdx⁽¹⁾での130億パラメータ・モデルの学習、国立研究開発法人産業技術総合研究所の第2回大規模言語モデル構築支援プログラムによるAI橋渡しクラウド（ABCI）での1750億パラメータ・モデルの学習トライアルの成果を踏まえ、パラメータ数(2)約1720億（GPT-3級）の大規模言語モデル（LLM）のフルスクラッチ学習を完了し、プレビュー版「LLM-jp-3 172B beta1」を公開する。

2024年9月17日に公開した1720億パラメータ・モデル(7000億トークン学習済み)のプレスリリース記事。上図の他の3つのモデルも公開済み

<https://www.nii.ac.jp/news/release/2024/0917.html>