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

September 17, 2024

Full-Scratch Learning of a Large Language Model with Approximately 172 billion Parameters (GPT-3 Level) and Preview Release

-- The World's Largest Fully Open Model, Including Training Data --

The Research and Development Center for Large Language Models (LLMC) at the National Institute of Informatics (NII) (Director-General: KUROHASHI Sadao, Chiyoda-ku, Tokyo, Japan), part of the Research Organization of Information and Systems, has successfully conducted a full-scratch learning of a large language model (LLM) with approximately 172 billion parameters(*1) (GPT-3 level). This is an achievement of the NII-led LLM Research Group (LLM-jp), building upon previous achievements such as the training of a 13 billion parameter model on the mdx(*2) Platform for Building Data-Empowered Society, and the trial learning of a 175 billion parameter model using the AI Bridging Cloud Infrastructure (ABCI) with support from the 2nd Large-scale Language Model Building Support Program of the National Institute of Advanced Industrial Science and Technology. The preview version of the model has been released, and it is the world's largest fully open model, including its training data.

The preview version has completed learning on approximately one-third of the prepared training data, which consists of around 2.1 trillion tokens. The plan is to continue learning and release the fully trained model around December 2024.

LLMC aims to promote research and development to ensure the transparency and reliability of LLMs, leveraging this and previously released models.

1. Overview of the Released LLM

(1) Computing Resources

- Approximately 0.4 trillion tokens of pre-training were conducted using cloud resources provided by Google Cloud Japan under the support of the Ministry of Economy, Trade and Industry's GENIAC project.
- Subsequently, up to about 0.7 trillion tokens of pre-training and tuning were carried out using cloud resources from SAKURA Internet, procured with a grant from the Ministry of Education, Culture, Sports, Science and Technology (MEXT).

National Institute of Informatics

Research Organization of Information and Systems
National Institute of Informatics
Publicity Team
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo
101-8430 JAPAN
Direct: +81(0)3-4212-2164 FAX : +81(0)3-4212-2150
E-Mail: media@nii.ac.jp

Web: https://www.nii.ac.jp
X(旧 Twitter): @jouhouken
facebook: https://www.facebook.com/jouhouken

(2) Model Training Corpus[*3)]

- The corpus consists of approximately 2.1 trillion tokens, with pre-training completed for about one-third.
    - Japanese: About 592 billion tokens
        - Text extracted from the full Common Crawl (CC) archive in Japanese
        - Data crawled from websites based on URLs collected by the National Diet Library's Web Archiving Project (WARP), with URL lists provided by the Library
        - Japanese Wikipedia
        - The summary text of each research project in the KAKEN (Database of Grants-in-Aid for Scientific Research).
    - English: About 950 billion tokens (Dolma, etc.)
    - Other languages: About 1 billion tokens (Chinese and Korean)
    - Program code: About 114 billion tokens
    - These add up to around 1.7 trillion tokens, and approximately 0.4 trillion Japanese tokens are to be used twice, resulting in about 2.1 trillion tokens

(3) Model

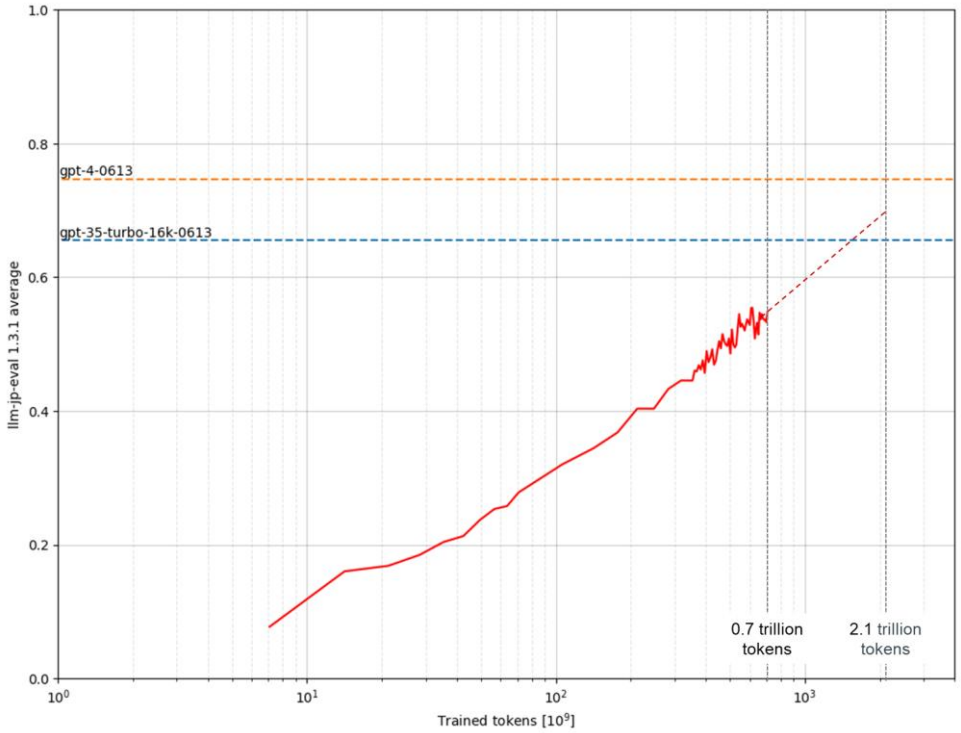
- Number of parameters: Approximately 172 billion (172B)
- Model architecture: LlaMA-2 based

(4) Tuning

- Tuning experiments were conducted using 13 types of Japanese instruction data and translations of English instruction data.

(5) Evaluation

- The model was evaluated using "llm-jp-eval v1.3.1," a cross-evaluation framework developed by LLM-jp, based on 22 types of existing Japanese language resources. The pre-training model, trained on 0.7 trillion tokens, achieved a score of 0.548.



- It was also evaluated using the "llm-leaderboard (g-leaderboard branch)," a framework used for performance assessment in the GENIAC project. The tuning model trained on 0.7 trillion tokens achieved a score of 0.529.

(6) URL for the Released Model, Tools, and Corpus

- https://llm-jp.nii.ac.jp/en/release/
- Note: Although the released model has undergone tuning for safety, it is still in the preview stage and not intended for direct use in practical services. The preview version will be provided under a limited license to approved applicants.

2. Future Plans

- To effectively utilize LLMs in society, ensuring transparency and reliability is essential. Additionally, as models become more advanced, safety considerations will become increasingly important. In response, NII established the Research and Development Center for Large Language Models in April 2024 with support from MEXT's project "R&D Hub Aimed at Ensuring Transparency and Reliability of Generative AI Models" (P7, https://www.mext.go.jp/content/20240118-ope_dev03-000033586-11.pdf, In Japanese). NII will continue to promote research and development using the released models and those yet to be built, contributing to the advancement of LLM research.
- With regard to the model released this time, all checkpoints up to the final checkpoint (100k steps), including every 1k step along the way, are saved and will be made available.

(Reference 1) Overview of LLM-jp

1. LLM-jp, organized by NII, consists of over 1,700 participants (as of September 17, 2024) from universities, companies, and research institutions, mainly focusing on researchers in natural language processing and computer systems. LLM-jp shares information on LLM research and development through hybrid meetings, online sessions, and Slack, while also conducting joint research on building LLMs. Specific activities include:
    - Promoting the development of open LLMs proficient in Japanese and related research.
    - Regular information exchange on model building expertise and the latest research developments.
    - Fostering collaboration across institutions by sharing data and computing resources.
    - Publishing outcomes such as models, tools, and technical documentation.
2. LLM-jp has established working groups such as "Corpus Construction WG," "Model Construction WG," "Fine-tuning & Evaluation WG," "Safety WG," "Multi-modal WG" and “Real Environment Interaction WG.” Each group, led respectively by Professor Daisuke Kawahara of Waseda University, Professor Jun Suzuki of Tohoku University, Professor Yusuke Miyao of the University of Tokyo, Project Professor Satoshi Sekine of NII, Professor Naoaki Okazaki of Tokyo Tech, and Professor Tetsuya Ogata of Waseda University, is engaged in research and development activities. Additionally, the initiative is propelled by the contributions of many others, including Professor Kenjiro Taura of the University of Tokyo, Associate Professor Yohei Kuga of the University of Tokyo (for utilization technologies of computational resource), and Professor Rio Yokota of Tokyo Tech (parallel computation methods).
3. For more details, visit the website: https://llm-jp.nii.ac.jp/en/

(Reference 2) Support for This Achievement

This achievement was made possible through a grant from the New Energy and Industrial Technology Development Organization (NEDO) and MEXT’s subsidy.

<Media Contact>
National Institute for Informatics Research Organization of Information and Systems
Publicity Team, Planning Division, General Affairs Department
TEL：+81-3-4212-2164　E-mail：media@nii.ac.jp

(*1) **Number of Parameters**：Large language models are massive neural networks trained on language data, and the number of parameters is one of the indicators of the network's size. It is generally believed that more parameters indicate higher performance.
(*2) **mdx (a platform for building a data-empowered society)**：A high-performance virtual environment focused on data utilization, jointly operated by a consortium of 9 universities and 2 research institutes. It is a platform for data collection, accumulation, and analysis that allows users to build, expand, and integrate research environments on-demand in a short amount of time, tailored to specific needs.
(*3) **Corpus**：A database that stores large amounts of natural language texts in a structural manner.