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

December 24, 2024

A Fully Open Large Language Model with Approximately 172 Billion Parameters (GPT-3 Level): "llm-jp-3-172b-instruct3" Now Publicly Available

-- Achieving Performance Beyond GPT-3.5 --

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 developed a large language model (LLM) with approximately 172 billion parameters (*1) (Comparable in scale to the number of parameters in GPT-3) trained from scratch using 2.1 trillion tokens of training data, and released the model to the public under the name "llm-jp-3-172b-instruct3." This model is the largest in the world among those that make not only model parameters but also training data publicly available. The model has surpassed GPT-3.5 in performance on benchmarks such as "llm-jp-eval," which measures various language understanding capabilities in Japanese, and "llm-leaderboard" developed as part of the GENIAC project (*2), a program by the Ministry of Economy, Trade, and Industry (METI) and the New Energy and Industrial Technology Development Organization (NEDO) to support the development of generative AI.

This model builds upon the results of training a 13-billion-parameter LLM on the mdx(*3) Platform for Building Data-Empowered Society, and a trial for developing 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. LLMC plans to utilize "llm-jp-3-172b-instruct3" to promote research and development to ensure the transparency and reliability of LLMs.

1. Overview of the Released LLM

(1) Computing Resources

- Approximately 0.4 trillion tokens of pre-training were processed using cloud resources provided by Google Cloud Japan under the support of the Ministry of Economy, Trade and Industry's GENIAC project.

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

- Subsequently, up to about 2.1 trillion tokens of pre-training and tuning were processed using cloud resources from SAKURA Internet, procured with a grant from the Ministry of Education, Culture, Sports, Science and Technology (MEXT).

(2) Model Training Corpus (*4)

- Pre-training was conducted using the following corpus (approximately 2.1 trillion tokens)
  - Japanese: Approximately 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 achieved a score of 0.613 in evaluations conducted with "llm-jp-eval v1.4.1," an evaluation framework developed by LLM-jp based on 26 existing Japanese datasets. This result exceeded GPT-3.5's performance score of 0.590 by 0.023 points.

- The model also achieved a score of 0.669 in evaluations on "llm-leaderboard (g-leaderboard branch)," an evaluation framework employed for performance evaluation in the GENIAC project. This result outperformed GPT-3.5's score of 0.653 by 0.016 points.

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

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

Notes :

- While this model has been fine-tuned for safety to the extent possible with current technology, it is technically challenging to guarantee safety. Depending on the input, the output may not always be appropriate.
- An evaluation of safety was conducted using 181 items from the safety dataset (AnswerCarefully v1). The results showed that 7 responses did not meet the safety criteria. This evaluation outcome is better than that of cutting-edge systems such as OpenAI's gpt-4-0613.

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.
- Additionally, apart from the final checkpoint (at 292,812 training steps), data from intermediate checkpoints at every 1,000 training steps leading up to the final checkpoint have also been preserved. We plan to make these checkpoints available in the future.

(Reference 1) Overview of LLM-jp

1. LLM-jp, organized by NII, consists of over 1,900 participants (as of December 24, 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 KAWAHARA Daisuke of Waseda University, Professor SUZUKI Jun of Tohoku University, Professor MIYAO Yusuke of the University of Tokyo, Project Professor SEKINE Satoshi of NII, Professor OKAZAKI Naoaki of Institute of Science Tokyo, and Professor OGATA Tetsuya of Waseda University, is engaged in research and development activities.
   Additional contributions come from many individuals, including: Professor TAURA Kenjiro and Associate Professor KUGA Yohei of the University of Tokyo (on computational infrastructures), and Professor YOKOTA Rio of Institute of Science Tokyo (on parallel computing methods).
3. For more details, visit the official website: https://llm-jp.nii.ac.jp/en/

(Reference2)

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 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 lead to higher performance.

(*2) GENIAC (Generative AI Accelerator Challenge): A program jointly conducted by the Ministry of Economy, Trade, and Industry (METI) and NEDO to strengthen domestic capabilities in developing generative AI. The initiative primarily focuses on providing computational resources for the development of foundation models, which are core technologies of generative AI, as well as supporting pilot studies for utilizing data and AI.

(*3) mdx (a platform for building a data-empowered society): A high-performance virtual environment focused on data utilization, jointly operated by a consortium of nine universities and two 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.

(*4) Corpus: A database that stores large amounts of natural language texts in a structural manner.