NEWS RELEASE

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

LLM勉強会
LLM-jp

October 20, 2023

Development of the Large Language Model "LLM-jp-13B" with 13 Billion Parameters

~The NII-hosted LLM Study Group (LLM-jp) releases initial results to contribute to academic and industrial research and development ~

The Research Organization of Information and Systems, National Institute of Informatics (NII, Director-General: Sadao Kurohashi, located in Chiyoda-ku, Tokyo) has been hosting the LLM Study Group (LLM-jp), consisting of over 500 participants including researchers in natural language processing and computer systems from universities and corporations, since May of this year. In July, utilizing the platform for building data-empowered society mdx[*1] as a computational resource, we commenced the development of a large language model (LLM) with 13 billion parameters[*2]. We are pleased to announce that the pre-training and tuning phase of the LLM is now complete and the model is released.

Although this model represents an early stage in LLM research and development and its performance metrics are comparable to previously released domestic models, all of its components, including its corpora, are made openly available to benefit future academic and industrial R&D efforts.

From now on, in collaboration with the National Institute of Advanced Industrial Science and Technology (AIST) and Tokyo Institute of Technology (Tokyo Tech), we aim to advance our research and development towards constructing a world-class performance LLM. As a first step, we have begun the construction of a more sophisticated LLM with 175 billion parameters (equivalent to GPT-3) using AIST's computational resource, the AI Bridging Cloud Infrastructure (ABCI)[*4].

In order to utilize LLMs in society, it is essential to ensure their transparency and reliability. As the models become more advanced, the consideration of safety becomes increasingly crucial. We will advance research using this model and future models to enhance the transparency, reliability, and safety of LLMs, contributing to the promotion of LLM research and development.

National Institute of Informatics

Research Organization of Information and Systems,
Inter-University Research Institute
National Institute of Informatics
Public Relations Team, Planning Division, General
Affairs Department
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430,

Web: https://www.nii.ac.jp
Twitter: @jouhouken
facebook: https://www.facebook.com/jouhouken

(*1) mdx (a platform for building 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.
(*2) 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.
(*3) Corpora: A database that consists of large amounts of structured natural language texts.
(*4) AI Bridging Cloud Infrastructure (ABCI): The largest computational resource currently available in Japan for AI, provided by the National Institute of Advanced Industrial Science and Technology (AIST).

1． Overview of LLM Study Group (LLM-jp)

（1） NII hosts the LLM-jp, where over 500 participants, including researchers in natural language processing and computer systems from universities and corporations, participate. LLM-jp utilizes hybrid meetings, online conferences, and Slack, among others, to share information on LLM research and development and collaboratively work on building LLMs. Specifically, activities are conducted for the following purposes:

- Construction of an open LLM that is proficient in Japanese and the promotion of related research and development
- Regular information exchange on model building insights and recent developments in research for researchers interested in natural language processing and related fields
- Promotion of cross-organizational collaborations among researchers, predicated on the sharing of data and computational resources
- Public release of outcomes, including models, tools, and technical materials

（2） For LLM construction, LLM-jp has established working groups such as "Corpora Construction WG," "Model Construction WG," and "Tuning & Evaluation WG." Each group, led respectively by Professor Daisuke Kawahara of Waseda University, Professor Jun Suzuki of Tohoku University, and Professor Yusuke Miyao of the University of Tokyo, is actively engaged in research and development activities. Additionally, the initiative is propelled by the contributions of many, including Professor Kenjiro Taura, the director of the University of Tokyo Information Technology Center, and Associate Professor Yohei Kuga (for the utilization of computational resource mdx), and Professor Rio Yokota of Tokyo Tech (parallel computation methods), among others.

(3) For more details, please refer to the homepage https://llm-jp.nii.ac.jp/.

2. Overview of the Newly Built LLM "LLM-jp-13B"

(1) Computational Resources Used:

- mdx (a platform for building data-empowered society): Utilized 12 nodes (A100 96 pieces).
- Funded by: NII, RIKEN Center for Advanced Intelligence Project (AIP), and Joint Usage/Research Center for Interdisciplinary Large-scale Information Infrastructures (JHPCN).
- Model Construction: Microsoft's DeepSpeed technology was employed.
- Monitoring and Logging: Weights & Biases were used for real-time monitoring and logging during model construction.

(2) Corpora Used for Model Training:

- A specialized tokenizer and web corpora filtering tool were developed to construct the corpora.
- Training Data Volume: Approximately 300 billion tokens (145 billion tokens in Japanese including Japanese mC4 and Japanese Wikipedia; 145 billion tokens in English including English Pile and English Wikipedia; and about 10 billion tokens of program code).

(3) Number of Model Parameters:

- 13 billion (13B)

(4) Tuning and Evaluation:

- Tuning: Conducted tuning experiments using 12 types of Japanese instruction data and English instruction data translated into Japanese. The evaluation results were analyzed, and the tuned models are now released.
- Evaluation: Developed 9 types of evaluation data using existing Japanese language resources and built a cross-sectional evaluation framework (approximately 50 types of evaluation data are expected to be constructed, including currently developed data).

（5） URL for the Released Models, Tools, and Corpora:
https://llm-jp.nii.ac.jp/release

Note: The model released this time is still in the initial stages of research and development. It has not been tuned to produce outputs aligned with human intentions from a safety perspective. Therefore, it is not intended to be provided as is for practical service.

3． Future Plans

- We are currently developing a search tool for the model training corpora, which is crucial for analyzing the behavior of the model, and plan to release it shortly.
- In cooperation with AIST and Tokyo Tech, we aim to advance our research and development towards constructing an LLM with world-class performance. As a first step, we were selected for AIST's 2nd ABCI Large-scale Language Model Building Support Program in September 2023, and building upon the insights gained from constructing the 13 billion parameter LLM, we embarked on the development of a 175 billion parameter LLM in October—a model not yet built or released by any other organization in Japan.
- In order to utilize LLMs in society, it is essential to ensure their transparency and reliability. As the models become more advanced, the consideration of safety becomes increasingly crucial. We will advance research using this model and future models to enhance the transparency, reliability, and safety of LLMs, contributing to the promotion of LLM research and development.

[Media inquiries should be directed to.]
Research Organization of Information and Systems,
Inter-University Research Institute
National Institute of Informatics
Public Relations Team, Planning Division,
General Affairs Department
TEL: 03-4212-2164 E-mail: media@nii.ac.jp