NEC's LLM with Superior Japanese Language Proficiency

OYAMADA Masafumi, AKIMOTO Kosuke, DONG Yuyang, YANO Taro, TAKEOKA Kunihiro, MAKIO Junta

Abstract

NEC have developed our own LLM (Large Language Model) with superior Japanese language proficiency and accelerating its use for internal operations and business applications. Despite its compact design capable of operating on a single GPU, this model boasts world-class Japanese language proficiency, achieved through long-time training with large amounts of high-quality data, a robust architecture, and meticulous tuning of instructions. Furthermore, with the motto, "Usable in business," we identified the elements necessary for LLMs in practical applications, such as high-speed inference and processing long texts of more than 200,000 characters. This paper provides an overview of the design philosophy, development process, and performance that we focused on strengthening.

Keywords

large language model, LLM, generative AI, general purpose

1. Introduction

In early 2023, we developed a proprietary LLM with a high performance and superior Japanese language proficiency and announced its development in July of that year. Since then, we have been accelerating its use for internal operations and expanded business applications while enhancing its performance and functionality. The LLM is designed to provide highly accurate responses to user instructions. **Fig. 1** provides an overview of the LLM's development process.

This paper first explains the design of the model and then follows with descriptions of how to prepare data for pre-training, perform chat-tuning with instruction data, and human feedback (alignment), and an evaluation of the model obtained by this process.

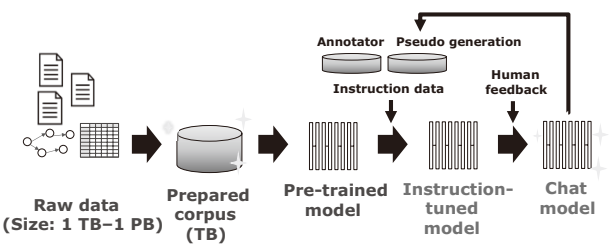

Fig. 1 Overview of the development process of the LLM.

2. Model Design and Pre-training

Most of the current LLMs are based on an architecture called the Transformer[1]. We have also based the design of our LLM's architecture on the Transformer.

(1) Model architecture

Considering the balance between performance and speed, the size of the LLM was set to have approximately 13 billion (13B) parameters. The number of layers, the number of hidden dimensions, and the number of heads were set to 40, 5120, and 40, respectively. They are the same as those of LLaMA 13B[2].

(2) Tokenizer

The tokenizer was trained using the BPE (byte pair encoding) algorithm[3] on a corpus totaling 10 GB of Japanese and English text, followed by post-processing to exclude inappropriate tokens.

(3) Pre-training

For pre-training, we used Megatron-DeepSpeed[4] and 64 servers, each equipped with eight A100 80 GB GPUs for computation. The batch size was set to 4 million tokens, and the learning rate was varied from 10^{-4} to 10^{-5} with a cosine scheduler. The sequence length during pre-training was initially set to 2048. Extensions for longer sequence lengths are discussed next.

(4) Long text support

To enhance the LLM's ability to infer information from long texts, we conducted additional pre-training using a long-text corpus after the aforementioned pre-training. Initially, when attempting to improve performance on long texts using standard additional pre-training methods, we found that the performance on short texts decreased as a result. To address this issue, we adopted a method called Positional Interpolation[5] for NEC's LLM. This method successfully improved performance on both short and long texts, as shown in **Fig. 2**. In the figure, the positions of tokens (x-axis) represent different parts of the text, with larger positions indicating longer texts and smaller positions indicating shorter ones.

3. Preparation of Data for Pre-Training

To achieve high performance with a parameter size of 13 billion, it is necessary to prepare an extremely large-scale and high-quality corpus for training data. To create our own corpus we collected and then processed a large amount of data primarily from the web, ensuring that knowledge from various fields was evenly included. In addition to Japanese corpus, we included English and source code corpus in a certain ratio to improve overall performance on those languages. Furthermore, to ensure that the LLM can generate coherent and natural texts, it is necessary not only to adjust the proportion of the data but also to clean the data for improved quality. Therefore, we conducted multi-stage cleaning for Japanese data using both rule-based and learning-based filtering. For learning-based filtering, we used a machine learning model that was trained to distinguish between clean and dirty data.

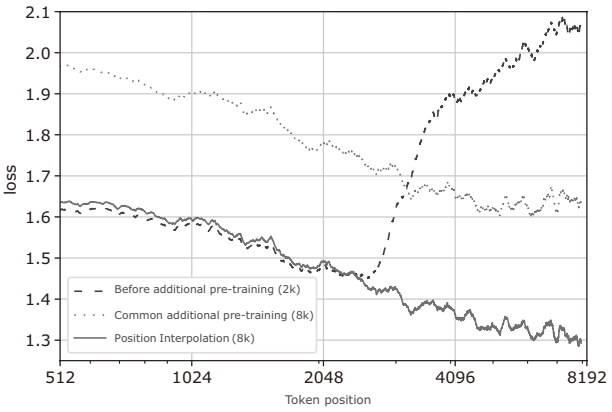

Fig. 2 Differences in text prediction error (loss) of the language model among various learning methods.

4. Chat-Tuning with Instruction Data

Through pre-training, the LLM can generate continuations of given text, but it lacks the ability to engage in meaningful conversations. To enable it to interact in a conversational manner, similar to ChatGPT, post-training called chat-tuning is required. In chat-tuning, we fine-tune the LLM using a dataset comprising multiple rounds of conversations between a user and assistant. This allows the LLM to function as an assistant and provide appropriate responses to user queries, taking into account the conversational context and generating responses based on the chat history between the user and the assistant.

As a refinement during chat-tuning, we applied a markup language called Chat Markup Language (ChatML) that was developed by OpenAI to structure the conversation data.

ChatML enables the semi-structured representation of system inputs, user inputs, and assistant outputs as follows.

<|im_start|>system
You are an AI assistant helpful to humans.
<|im_end|>
<|im_start|>user
Tell me who is the president of NEC.
<|im_end|>
<|im_start|>assistant
The president of NEC Corporation (NEC) is Takayuki Morita. Mr. Morita assumed office on April 1, 2021.
<|im_end|>

In addition to the large-scale dataset created by NEC employees and other people, we developed and utilize a mechanism to semi-automatically augment conversation data using the LLM we developed to enhance the ability to generate responses to more complex instructions and diverse questions.

5. Human Feedback (Alignment)

Through chat-tuning, the LLM learns to follow the user's instructions. However, chat-tuned LLMs may

occasionally generate harmful outputs or unhelpful responses. Reinforcement learning from human feedback (RLHF)[6] is a training method that suppresses undesirable responses from the LLM and encourages desirable ones. RLHF is typically a two-stage process that involves training of reward model and proximal policy optimization (PPO)[7] after chat-tuning. However, there can be instability issues during training depending on how the PPO hyperparameters are selected. Recently, Direct Preference Optimization (DPO)[8] has gained attention as a way to address this issue. DPO offers stable training in a single stage and achieves good performance. Therefore, we used DPO for alignment.

DPO is an algorithm that directly optimizes model parameters by using data that consists of a set of triplets in the format <prompt, positive example response, negative example response> (preference data) for a given prompt, where positive examples represent desirable responses and negative examples represent undesirable responses, to increase the likelihood of positive examples and decrease the likelihood of negative examples. By training on many such triplets, the LLM becomes more likely to output patterns of desirable responses than patterns of undesirable responses.

The preference data was created by generating answer candidates using chat-tuned LLMs in response to prompts collected from our employees and external companies.

Implementing such alignment processes has been confirmed to result in the improved quality of the human evaluation of LLM responses and the suppression of harmful output.

6. Benchmarks

The evaluation of the LLM's performance is multifaceted, but here we evaluate it from two of the most common perspectives: the inference and information processing ability as a pre-trained LLM, which is referred to as Japanese language proficiency and includes common sense reasoning and document comprehension; and the information processing ability as a chat-tuned LLM, which requires complex capabilities.

6.1 Evaluation as a pre-trained LLM

To evaluate the pre-trained LLM, we used JSQuAD and JCommonsenseQA (JCQA) from JGLUE[9], which is a commonly used Japanese language benchmark. JSQuAD is a task that involves extracting an answer string for the provided question from a given document. It uses two-shot in-context learning, and the evaluation score is based on whether the predicted answer exactly matches the correct answer string. JCQA is a multiple-choice question answering dataset for the task of asking questions about common-sense knowledge with answers selected from five options. During inference, it uses three-shot in-context learning, and the evaluations are based on accuracy. The baseline LLMs include globally top-tier LLMs from other countries and high-performance Japanese LLMs available to us as of the end of October 2023, denoted as (A, B, C, and so forth).

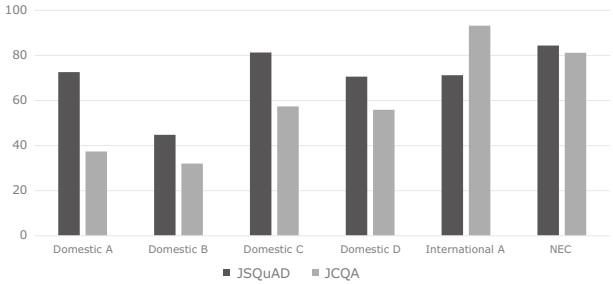

Fig. 3 JSQuAD and JCQA experiment results: NEC's LLM vs other LLMs.

Fig. 3 shows the evaluation results. Among the LLMs compared, our LLM achieved the highest performance in JSQuAD, and in JCQA, it achieved the top performance among domestic LLMs and the second-highest performance overall, including LLMs from other countries.

6.2 Evaluation as a chat-tuned LLM

Next, we adopted RAKUDA[10] as a Japanese language benchmark for evaluating the performance as a chat-tuned LLM. The task of RAKUDA not only requires knowledge and comprehension as mentioned earlier but also seeks valid responses in a conversational context. RAKUDA involves a task responding to 40 Japanese free-form queries, and the quality of responses is evaluated by GPT-4. It employs an evaluation method that compares the quality of responses generated by two different LLMs, in order to determine which one is better. Here, we evaluate the LLM developed by NEC in comparison with the other Japanese-specialized LLMs that were accessible to NEC as of the end of October 2023 (referred to as X, Y, and Z).

Fig. 4 shows the evaluation results for the Rakuda benchmark. The LLM we developed achieves high performance even when compared to globally top-tier LLMs, demonstrating its capability of dialogue. In particular, it significantly outperforms Japanese-specialized LLMs in terms of win rates. However, it is important to note that

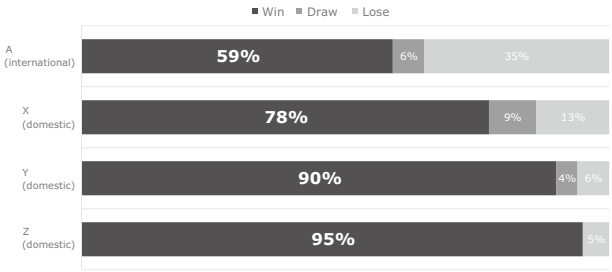

Fig. 4 Win rates of NEC's LLM against other LLMs in the Rakuda benchmark.

the Rakuda benchmark evaluates single-turn conversations with a small amount of data so it measures only one aspect of dialogue performance.

7. Conclusion

In this paper, we introduced the 13B LLM developed by NEC. In the future, we will start with this LLM and develop a variety of generative AI technologies as a contribution to society.

* ChatGPT is a trademark of OpenAI Inc. in the United States.
* All other names of companies names and products that appear in this paper are trademarks or registered trademarks of their respective companies.

References

1) Ashish Vaswani, et al.: Attention Is All You Need, 2017 https://arxiv.org/abs/1706.03762
2) Hugo Touvron et al.: LLaMA: Open and Efficient Foundation Language Models, 2023 https://arxiv.org/abs/2302.13971
3) Rico Sennrich, Barry Haddow, and Alexandra Birch: Neural Machine Translation of Rare Words with Subword Units, 2015 https://arxiv.org/abs/1508.07909
4) GitHub: Megatron-DeepSpeed https://github.com/bigscience-workshop/Megatron-DeepSpeed
5) Shouyuan Chen et al.: Extending Context Window of Large Language Models via Positional interpolation, 2023 https://arxiv.org/abs/2306.15595
6) Long Ouyang et al.: Training language models to follow instructions with human feedback, 36th Conference on Neural Information Processing Systems, 2022 https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf
7) John Schulman et al.: Proximal Policy Optimization Algorithms, 2017 https://arxiv.org/abs/1707.06347
8) Rafael Rafailov et al.: Direct Preference Optimization: Your Language Model is Secretly a Reward Model, 2023 https://arxiv.org/abs/2305.18290
9) Kentaro Kurihara et al.: JGLUE: Japanese General Language Understanding Evaluation, LREC 2022, pp. 2957-2966, 2022 http://www.lrec-conf.org/proceedings/lrec2022/pdf/2022.lrec-1.317.pdf
10) GitHub: Japanese-llm-ranking https://github.com/yuzu-ai/japanese-llm-ranking

Authors' Profiles

OYAMADA Masafumi
Research Fellow and Group head
Data Science Laboratories

AKIMOTO Kosuke
Manager
Data Science Laboratories

DONG Yuyang
Special Researcher (Assistant Manager)
Data Science Laboratories

YANO Taro
Assistant Manager
Data Science Laboratories

TAKEOKA Kunihiro
Special Researcher (Assistant Manager)
Data Science Laboratories

MAKIO Junta
Data Science Laboratories

Information about the NEC Technical Journal

Thank you for reading the paper.
If you are interested in the NEC Technical Journal, you can also read other papers on our website.

Link to NEC Technical Journal website

Japanese

English

Vol.17 No.2 Special Issue on Revolutionizing Business Practices with Generative AI
– Advancing the Societal Adoption of AI with the Support of Generative AI Technologies

Remarks for Special Issue on Revolutionizing Business Practices with Generative AI
Approaches to Generative AI Technology: From Foundational Technologies to Application Development and Guideline Creation

Vol.17 No.2
June 2024

Special Issue TOP

Papers for Special Issue

Market Application of Rapidly Spreading Generative AI
NEC Innovation Day 2023: NEC's Generative AI Initiatives
Streamlining Doctors' Work by Assisting with Medical Recording and Documentation
Using Video Recognition AI x LLM to Automate the Creation of Reports
Understanding of Behaviors in Real World through Video Analysis and Generative AI
Automated Generation of Cyber Threat Intelligence
NEC Generative AI Service (NGS) Promoting Internal Use of Generative AI
Utilization of Generative AI for Software and System Development
LLMs and MI Bring Innovation to Material Development Platforms
Disaster Damage Assessment Using LLMs and Image Analysis

Fundamental Technologies that Enhance the Potential of Generative AI
NEC's LLM with Superior Japanese Language Proficiency
NEC's AI Supercomputer: One of the Largest in Japan to Support Generative AI
Towards Safer Large Language Models (LLMs)
Federated Learning Technology that Enables Collaboration While Keeping Data Confidential and its Applicability to LLMs
Large Language Models (LLMs) Enable Few-Shot Clustering
Knowledge-enhanced Prompt Learning for Open-domain Commonsense Reasoning
Foundational Vision-LLM for AI Linkage and Orchestration
Optimizing LLM API usage costs with novel query-aware reduction of relevant enterprise data

For AI Technology to Penetrate Society
Movements in AI Standardization and Rule Making and NEC Initiatives
NEC's Initiatives on AI Governance toward Respecting Human Rights
Case Study of Human Resources Development for AI Risk Management Using RCModel

NEC Information

2023 C&C Prize Ceremony