Constructing Open-source Large Language Model Evaluator for Japanese

Yikun Sun, Sakiko Yahata
Fei Cheng, Yugo Murawaki, Chenhui Chu, Sadao Kurohashi
Kyoto University
{sun,yahata}@nlp.ist.i.kyoto-u.ac.jp
{feicheng, murawaki, chu, kuro}@i.kyoto-u.ac.jp

Abstract

Evaluating the performance of large language models (LLMs) remains a crucial research topic, and conducting a comprehensive and accurate evaluation of LLM performance efficiently is challenging. This challenge is particularly acute for non-English languages. GPT-4-based automated evaluation has proven effective, demonstrating high consistency with human preference. However, GPT-4-based evaluation still has several limitations, including its closed-source nature and general preference. This paper proposes an approach to construct an open-source Japanese LLM evaluator, which has demonstrated robust consistency on the Japanese Vicuna benchmark. We present a method for rapidly generating score rubrics that refer to specific instructions, enabling more diverse evaluation criteria when evaluating LLMs. Our Japanese LLM evaluator training data and models are available here.[1] [2]

1 Introduction

The quality assessment of text generated by large language models (LLMs) remains a significant challenge in the field of Natural Language Processing (NLP), as text quality directly reflects LLM performance [1, 2]. Currently, utilizing high-performance LLMs (such as ChatGPT) for automatic evaluation of LLM responses represents a viable approach and demonstrates comparable evaluation accuracy to human evaluation [3]. However, this approach faces several inherent limitations, including its closed-source nature and limitations on general preference [4].

PROMETHEUS [5] developed an open-source LLM evaluator for English LLM evaluation. This LLM can evaluate English LLM performance with diverse score rubrics. It demonstrates strong consistency with both human evaluations and GPT-4-based evaluations. However, PROMETHEUS exhibits limitations. It is restricted to evaluating English-language LLMs. Moreover, when it evaluates a specific instruction, it requires manually created score rubrics for that instruction. Both of these requirements impose constraints on evaluation feasibility.

To address these limitations, we propose an open-source Japanese LLM Evaluator that provides diverse rubrics of LLM performance evaluation. Additionally, to generate diverse score rubrics for different instructions, we propose to enable the LLM evaluator to automatically generate score rubrics for specific instruction evaluation.

To this end, we construct a dataset for training the Japanese LLM Evaluator. This dataset includes diverse score rubrics and corresponding instructions, along with responses ranging from quality scores 1 to 5 and responding feedback, designed to guide the LLM evaluator in conducting evaluations. Subsequently, we train an open-source Japanese LLM evaluator capable of efficiently assessing problems from multiple perspectives, meeting diverse evaluation requirements. Through experimentation, we demonstrated strong consistency with human evaluation on the Japanese Vicuna Benchmark [6].

Furthermore, we train an open-source Japanese rubrics generator on our dataset. It can automatically generate instructive and diverse score rubrics based on specific instructions. Instead of generating rubrics manually, these automatic generation rubrics can guide LLM evaluation across different evaluation dimensions.

[1] https://huggingface.co/ku-nlp/jp_llm_evaluator
[2] https://huggingface.co/datasets/ku-nlp/jp_llm_evaluator_training

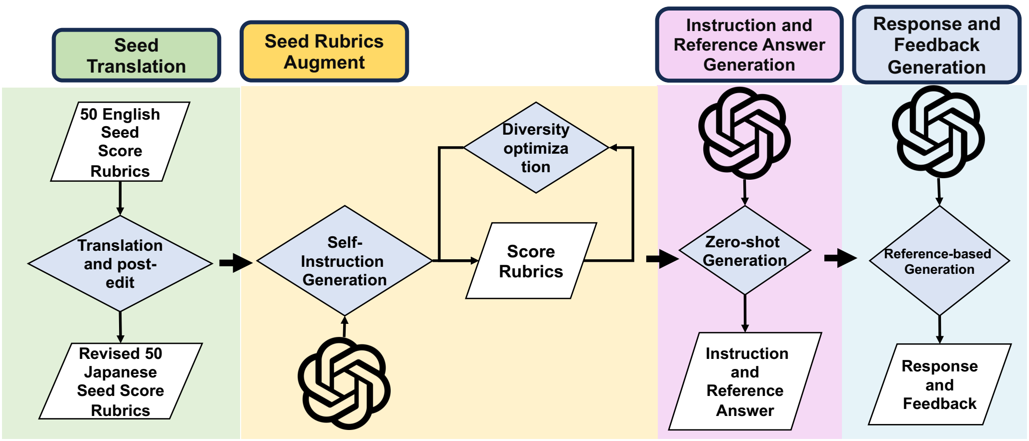
Seed Translation
Seed Rubrics Augment
Instruction and Reference Answer Generation
Response and Feedback Generation
50 English Seed Score Rubrics
Translation and post-edit
Revised 50 Japanese Seed Score Rubrics
Self-Instruction Generation
Diversity optimiza tion
Score Rubrics
Zero-shot Generation
Instruction and Reference Answer
Reference-based Generation
Response and Feedback

Figure 1 Japanese LLM evaluator training dataset construction

Our contributions are as follows:

- We generate Japanese LLM evaluator training data with diverse score rubrics and instruction using GPT-4o.
- We train an open-source Japanese LLM evaluator on our dataset.
- Experiment results show the consistency with Japanese LLM evaluator and GPT-based evaluation.

2 Methodology

Our work consists of three main parts. The first part utilizes self-instruct [7, 8, 9] to guide GPT-4o [10] to automatically generate training data for Japanese LLM evaluators. Through these high-quality training data, we can implement various functionalities of Japanese LLM evaluators by performing supervised fine-tuning (SFT) on different pre-trained LLMs. The second part is training open-source Japanese LLM evaluators using high-quality training data. The third part is to generate specific score rubrics for instructions. We will now elaborate on our work in detail.

2.1 Japanese LLM Evaluator Training Dataset Construction

In Figure 1, this pipeline demonstrates how we utilize a small number of seed tasks to guide GPT-4o in generating high-quality training data with instructional effectiveness. This pipeline consists of four main steps:

- **Seed Translation**
- **Seed Rubrics Augment**
- **Instruction and Reference Answer Generation**
- **Response and Feedback Generation**

You can find our detail process of Japanese LLM evaluator training dataset construction from Appendix A.

2.2 Japanese LLM Evaluator Training

In this section, we focus on the training process of the Japanese LLM evaluator. For model training, we employ low-rank adaptation (LoRA) [11] as our SFT training method. This approach ensures model performance while maintaining computational efficiency.

The training data for the Japanese LLM evaluators comprise the follow contents:

- **Instruction:** This is guidelines for directing model responses, covering various practically meaningful instructional questions.
- **Score Rubric:** This is evaluation criteria for specific instructions, considering different aspects of response quality evaluation. For example, for specific instruction, we may have multiple perspectives such as “cultural sensitivity,” “grammatical accuracy,” “humor sense” etc. Under different rubrics, the evaluation of responses will vary. Score rubrics set score from 1 to 5, enabling the LLM evaluator to understand and evaluate responses.
- **Response:** This is an answer obtained from different LLMs for specific instructions, serving as evaluation targets for the LLM evaluator.
- **Score:** This is the output component of the LLM evaluator, summarizing the evaluation of responses. The score is given directly with reference to score rubrics, ranging from 1 to 5 as integers.
- **Feedback:** This is the output component of the LLM evaluator, providing detailed explanations for evalua-

tion scores. It encompasses understanding and interpretation of both response and score rubrics, explaining why specific scores were assigned under the given rubrics.

2.3 Score Rubrics Generation during Evaluation

When evaluating responses, the LLM evaluator needs to choose diverse score rubrics for specific instructions. Traditionally, generating score rubrics has been done through handcraft. To more efficiently generate diverse score rubrics, we conducted SFT on pre-trained models to get an open-source rubrics generator. With our previously generated training data, we paired "instructions" and "score rubrics" as training samples, guiding the rubrics generator to automatically generate meaningful and diverse score rubrics for various instructions.

3 Experimental Settings

In this section, we present the experimental settings to evaluate our Japanese open-source LLM evaluator and score rubrics generation.

3.1 LLMs as Evaluator

In this experiment, we chose three pre-trained models for supervised fine-tuning. The target pre-trained models include:

- Llama-3.1-8B
- llm-jp-3-13b
- Llama-3.1-Swallow-8B-v0.2

LLama-3.1 model represents the current high-quality cross-lingual pre-trained model in terms of scale and quality. It shows excellent potential across various tasks.Llama-3.1-Swallow-8B-v0.2 and llm-jp-3-13b are base models that took continued pre-training using Japanese datasets, exhibiting awesome performance in Japanese while maintaining English language capabilities.

3.2 LLM Evaluator Evaluation

We referenced the Japanese Vicuna Benchmark as our evaluation benchmark. This benchmark contains 80 diverse questions designed to guide LLMs in generating meaningful responses.

Additionally, we referenced the score rubrics handcrafted generated by PROMETHEUS for the Vicuna bench [3], which were translated from English to Japanese through human translation as evaluation criteria.

Based on the Japanese Vicuna Benchmark, we generated 80 responses using the following LLMs as evaluation targets for the LLM evaluator:

- GPT-4o
- llm-jp-3-13b [12]
- Llama-3.1-Swallow-8B-v0.2 [13]
- Swallow-7b-hf [14, 15]

We evaluated the LLM evaluators' performance through consistency against the GPT–4-based evaluation. GPT-based evaluation has demonstrated exceptional evaluation performance and human consistency across many evaluation benchmarks. It also shows higher efficiency compared to manual evaluation. We referenced GPT-4o scoring of target responses to calculate the consistency between the LLM evaluator and GPT-4-based evaluation.

To calculate consistency, we chose Pearson Correlation to compute the consistency:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- x_i: The score of the i-th question assessed by LLM evaluator.
- y_i: The score of the i-th question assessed by GPT-4.
- $\bar{x}$: The mean score of all questions assessed by LLM evaluator (x).
- $\bar{y}$: The mean score of all questions assessed by GPT-4.
- r: The Pearson Correlation Coefficient

3.3 Score Rubrics Generation Evaluation

We referenced instructions and handcrafted score rubrics from the Japanese Vicuna Benchmark. We utilized the Japanese rubrics generator to generate score rubrics for specific instructions from this benchmark, a total of 80 different score rubrics. Subsequently, we calculated the average F-measure between these generated score rubrics and handcrafted score rubrics using ROUGE-L to demonstrate the diversity and feasibility of our Japanese LLM evaluator's generated score rubrics.

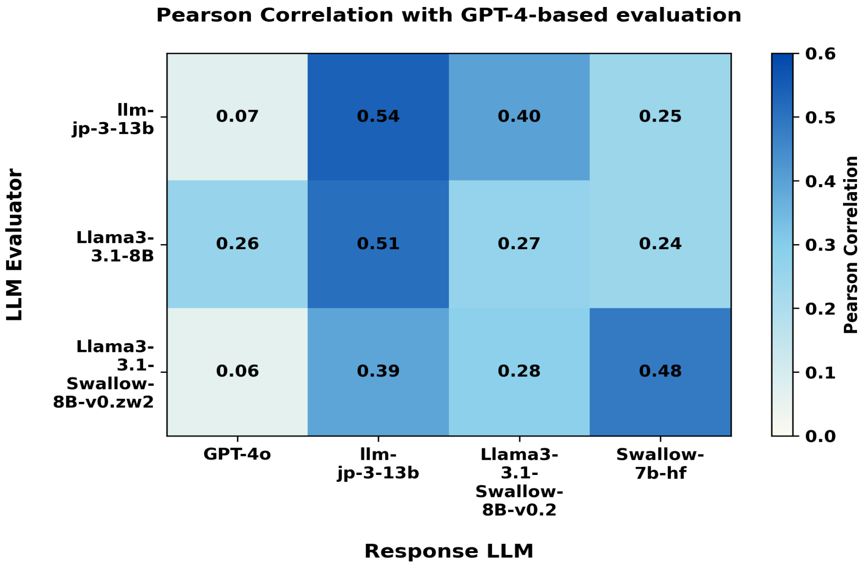

Figure 2 Pearson Correlation between our LLM evaluator and GPT-4-based evaluation.

4 Result Analysis

4.1 LLM Evaluator Evaluation

As shown in Figure 2, we can observe the Pearson Correlation between various base model LLM evaluators evaluation results and GPT-4-based evaluation. The Japanese LLM evaluator demonstrates high Pearson Correlation results for responses generated by Llama-3.1-Swallow-8B-v0.2, llm-jp-3-13b, and Swallow-7b-hf.

Regarding GPT-4o responses, only the LLAMA3 base LLM evaluator shows a high Pearson Correlation. However, our additional analysis of GPT-4o average scores across different LLM evaluators retails all of them is approximately 4.3 score, demonstrating consistency in evaluation capabilities. You can find the detail results of LLMs responses average scores in Appendix B.

4.2 Score Rubrics Generation Evaluation

To evaluate the instruction with diverse criteria, we trained the Japanese rubrics generator. We evaluate the score rubrics generation capability of the Japanese rubrics generator by calculating the ROUGE-L scores of score rubrics generated by rubrics generator with specific instructions. All score rubrics newly generated for evaluation will be compared with the score rubrics created handcrafted.

As results in Table 1, we can see the average ROUGE-L scores of score rubrics generated by three Japanese LLM evaluators compared to handcrafted score rubrics. We observe that all generated score rubrics have an average F-measure higher than 0.2. This shows that the Japanese LLM evaluator provides instructive and diverse criteria for specific instruction.

Generator	Llama-3.1-8b	llm-jp-3-13b	Swallow
ROUGE Score	0.243	0.211	0.270

Table 1 Average ROUGE-L F-measure Score

5 Conclusion

This paper presented a method for developing an open-source Japanese LLM evaluator. With the human translation of a small number of seed rubrics and utilizing GPT-4o, we constructed a comprehensive dataset for training the Japanese LLM evaluator. Subsequently, we performed SFT on the Japanese LLM evaluator using this training dataset, implementing the ability of LLM evaluation and score rubrics generation. For LLM evaluation, the Japanese LLM evaluator demonstrates reliable consistency with GPT-4-based evaluation. The score rubrics generation provides an automatic method for generating instructive and diverse score rubrics by referencing specific instructions, guiding the LLM evaluation. For the future work, we plan to explore the consistency between Japanese LLM evaluation and human preference consistency. Moreover, we plan to make our Japanese LLM evaluator capable for pairwise evaluation.

6 Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP23K28144 and the “R&D Hub Aimed at Ensuring Transparency and Reliability of Generative AI Models” project of the Ministry of Education, Culture, Sports, Science and Technology.

References

[1] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena, 2023.

[2] Mingqi Gao, Xinyu Hu, Jie Ruan, Xiao Pu, and Xiaojun Wan. LLM-based NLG Evaluation: Current Status and Challenges, 2024.

[3] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality, March 2023.

[4] Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. Prometheus: Inducing Fine-grained Evaluation Capability in Language Models, 2024.

[5] Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. Prometheus 2: An Open Source Language Model Specialized in Evaluating Other Language Models, 2024.

[6] Yikun Sun, Zhen Wan, Nobuhiro Ueda, Sakiko Yahata, Fei Cheng, Chenhui Chu, and Sadao Kurohashi. Rapidly Developing High-quality Instruction Data and Evaluation Benchmark for Large Language Models with Minimal Human Effort: A Case Study on Japanese. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, **Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)**, pp. 13537–13547, Torino, Italia, May 2024. ELRA and ICCL.

[7] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-Instruct: Aligning Language Models with Self-Generated Instructions, 2023.

[8] Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. Unnatural Instructions: Tuning Language Models with (Almost) No Human Labor, 2022.

[9] Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. Baize: An Open-Source Chat Model with Parameter-Efficient Tuning on Self-Chat Data, 2023.

[10] OpenAI. GPT-4 Technical Report. **arXiv preprint arXiv:2303.08774**, 2023.

[11] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models, 2021.

[12] LLM-jp: A Cross-organizational Project for the Research and Development of Fully Open Japanese LLMs, author=LLM-jp and : and Akiko Aizawa and Eiji Aramaki and Bowen Chen and Fei Cheng and Hiroyuki Deguchi and Rintaro Enomoto and Kazuki Fujii and Kensuke Fukumoto and Takuya Fukushima and Namgi Han and Yuto Harada and Chikara Hashimoto and Tatsuya Hiraoka and Shohei Hisada and Sosuke Hosokawa and Lu Jie and Keisuke Kamata and Teruhito Kanazawa and Hiroki Kanezashi and Hiroshi Kataoka and Satoru Katsumata and Daisuke Kawahara and Seiya Kawano and Atsushi Keyaki and Keisuke Kiryu and Hirokazu Kiyomaru and Takashi Kodama and Takahiro Kubo and Yohei Kuga and Ryoma Kumon and Shuhei Kurita and Sadao Kurohashi and Conglong Li and Taiki Maekawa and Hiroshi Matsuda and Yusuke Miyao and Kentaro Mizuki and Sakae Mizuki and Yugo Murawaki and Akim Mousterou and Ryo Nakamura and Taishi Nakamura and Kouta Nakayama and Tomoka Nakazato and Takuro Niitsuma and Jiro Nishitoba and Yusuke Oda and Hayato Ogawa and Takumi Okamoto and Naoaki Okazaki and Yohei Oseki and Shintaro Ozaki and Koki Ryu and Rafal Rzepka and Keisuke Sakaguchi and Shota Sasaki and Satoshi Sekine and Kohei Suda and Saku Sugawara and Issa Sugiura and Hiroaki Sugiyama and Hisami Suzuki and Jun Suzuki and Toyotaro Suzumura and Kensuke Tachibana and Yu Takagi and Kyosuke Takami and Koichi Takeda and Masashi Takeshita and Masahiro Tanaka and Kenjiro Taura and Arseny Tolmachev and Nobuhiro Ueda and Zhen Wan and Shuntaro Yada and Sakiko Yahata and Yuya Yamamoto and Yusuke Yamauchi and Hitomi Yanaka and Rio Yokota and Koichiro Yoshino, 2024.

[13] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, and Angela Fan et al. The Llama 3 Herd of Models, 2024.

[14] Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. Continual Pre-Training for Cross-Lingual LLM Adaptation: Enhancing Japanese Language Capabilities. In **Proceedings of the First Conference on Language Modeling**, COLM, p. (to appear), University of Pennsylvania, USA, October 2024.

[15] Naoaki Okazaki, Kakeru Hattori, Hirai Shota, Hiroki Iida, Masanari Ohi, Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Rio Yokota, and Sakae Mizuki. Building a Large Japanese Web Corpus for Large Language Models. In **Proceedings of the First Conference on Language Modeling**, COLM, p. (to appear), University of Pennsylvania, USA, October 2024.

[16] Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In **Text Summarization Branches Out**, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

A Japanese LLM Evaluator Training Dataset Construction Process

A.1 Seed Translation

The first step involves constructing a high-quality Japanese seed rubrics. We referenced the handcrafted seed rubrics from PROMETHEUS and translated 50 of seed rubrics into Japanese through human translation to achieve native-level quality. Subsequently, we can utilize GPT-4o to directly generate high-quality Japanese score rubric data given the seeds as few-shot examples.

A.2 Score Rubrics Augment

In this step, we augment the high-quality Japanese seed rubrics. We construct a comprehensive prompt with specific requirements to guide GPT-4o in automatically generating new high-performance score rubrics. In the prompt, we emphasize that these score rubrics should demonstrate creative and diverse evaluation capabilities, and assess problems from multiple perspectives. Finally, the rubrics specify answer criteria for each score level from 1 to 5. During the self-instruct generation process, to ensure the diversity of generated score rubrics, we implement various diversity optimization methods to filter and optimize the generated score rubrics. For each generation round, we compare the newly generated score rubrics with those in the rubrics pool using two primary methods:

- We employ ROUGE-L [16] to evaluate the similarity between newly generated instructions and all previously generated data in the instruction pool. If the generated data exhibit a ROUGE-L score exceeding 0.7 with any existing rubric in the rubrics pool, we subject these data to a second optimization phase.
- The second optimization phase primarily involves paraphrasing to restructure the generated score rubrics. Through constructing guiding prompts, we direct GPT-4o to paraphrase the score rubrics. This process focuses on modifying word choice and sentence structure while preserving the inherent meaning of the generated score rubrics.

Through GPT-4o batch generation, we successfully generated 1K diverse score rubrics.

A.3 Instruction and Reference Answer Generation

In this step, we construct corresponding instructions and response answers for the generated score rubrics. For each generated score rubric, we guide GPT-4o to generate, through zero-shot learning, 20 instructions that can be evaluated using the score rubric, along with corresponding score 5 reference answers. Each score rubric can be used to evaluate various questions, while the reference answers serve as examples of optimal responses earning a score of 5 under the respective score rubric. Ultimately, we obtained 20K diverse instruction-score rubric pairs.

A.4 Response and Feedback Generation

The final step focuses on generating responses scoring from scores 1 to 5 and corresponding evaluation feedback for the instruction-score rubric pairs. We establish instructive prompts to guide GPT-4o in generating corresponding responses and feedback for each instruction and rubrics pair. Each reference answer serves as a score 5 reference, acting as a upper-bound for generating response levels.

Finally, we obtained a dataset of 100K sets, each set containing unique score rubrics, instructions, reference responses, responses, and feedback. This data set will be used for future Japanese LLM evaluator training.

B Average Scores of LLMs Responses

In this part, we give the average scores of LLMs responses. It also show the score consistency with GPT-4o based evaluation.

Table 2 Average Scores of LLMs Responses

LLM Evaluator	GPT-4o	llmjp-3-13b	Llama3-Swallow-8B-v0.2	Swallow-7b-hf
LLaMA3	4.43	4.52	3.65	1.58
llmjp	4.32	4.50	3.35	1.43
swallow	4.37	4.55	3.45	1.40
GPT-4o	4.38	3.91	3.78	2.25