Exploring Open Large Language Models for the Japanese Language: A Practical Guide

Kaito Sugimoto
Technology Division, Innovation Center, NTT Communications Corporation
Granpark Tower 3-4-1 Shibaura, Minato-ku, Tokyo 108-8118, Japan
kaito.sugimoto@ntt.com

Abstract

While large language models (LLMs) have demonstrated remarkable capabilities in handling Japanese, they are conventionally trained on English-centric corpora, which may cause a deficiency in understanding and generating Japanese texts. In response, researchers have been actively developing LLMs with a specific focus on Japanese, many of which have been made publicly available. This rapid growth has made it challenging to obtain a comprehensive overview of the developments. To address this issue, this report reviews open LLMs for Japanese, including instruction-tuned models and multimodal models. We also introduce existing LLM evaluation benchmarks for Japanese, aiming to offer a practical guide to choosing the most suitable model. We continually update our work at `https://github.com/llm-jp/awesome-japanese-llm`.

Keywords: Large Language Models, Japanese Language

1. Introduction

Large language models (LLMs) have exhibited vast knowledge and strong reasoning abilities, demonstrating their potential to support human activities (Wang et al., 2023a; Xi et al., 2023). Their influence has now extended far beyond the confines of the traditional NLP community. In Japan, the success of LLMs has encouraged the exploration of their applications to various industries (METI, 2024). Several studies in the medical field (Kunitsu, 2023; Tanaka et al., 2024; Nakao et al., 2024) have already investigated the effectiveness of proprietary LLMs such as GPT-4 (OpenAI, 2023). As Japan suffers from a severe labor shortage (Shinkawa, 2012), LLMs are expected to provide assistance and improve productivity in Japanese society.

One major drawback of widespread LLMs is that they are (presumably) trained on web-scale corpora, which include a considerably low ratio of Japanese documents. Indeed, Llama 2 (Touvron et al., 2023) reveals that only 0.1% of its pre-training data is in Japanese. This imbalance could make such models less optimal for Japanese speakers in the following aspects. First, it has been reported that the factual knowledge of LLMs correlates with its frequency in their pre-training data (Kandpal et al., 2023; Chang et al., 2023a). This suggests that English-centric LLMs might not accurately reflect the shared values, beliefs, and customs of Japanese people. Another practical issue arises from the tokenizers of LLMs. Popular tokenization methods, such as BPE (Sennrich et al., 2016) and Unigram (Kudo, 2018), regard a frequent pattern of characters (or bytes) as a unit. Therefore, infrequent Japanese words tend to be split into smaller units (see Figure 1). This not only slows inference speed but also imposes higher costs because proprietary LLMs usually charge by the token count.

As an alternative to English-centric LLMs, researchers have been developing a "Japanese LLM"—a large language model trained primarily on the Japanese language (Hornyak, 2023). Many research groups have released Japanese LLMs to demonstrate their technological prowess and promote further research (Sawada et al., 2024; Levine et al., 2024; Takahashi et al., 2024; Akiba et al., 2024; Inoue et al., 2024, *inter alia*). This development race has become so intense that it is difficult to get the whole picture. To tackle this issue, we in this study provide a concise overview of the recent advancements in Japanese LLMs. Specifically, we try to answer the following questions:

Q1: *What kind of Japanese LLMs are out there that anyone can use?* (§2.)

Q2: *How can we measure the performance of LLMs in Japanese?* (§3.)

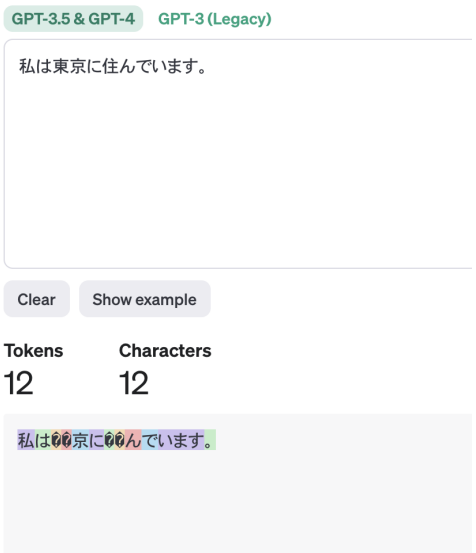

Figure 1: Illustration of how GPT-4 (OpenAI, 2023) tokenizes a Japanese text. In this example, the sentence "私は東京に住んでいます。 (I live in Tokyo.)" is tokenized into twelve tokens. Kanjis such as "東" and "住" are not included in the vocabulary and thus mapped to multiple tokens. Note that OpenAI has announced that they will release a custom model optimized for Japanese (OpenAI, 2024).

Model	Type	Release Date	Model ID
LLM-jp-13B	FS	10/2023	`llm-jp/llm-jp-13b-v1.0`
PLaMo-13B	FS	09/2023	`pfnet/plamo-13b`
Stockmark-13B	FS	10/2023	`stockmark/stockmark-13b`
Weblab-10B	FS	08/2023	`matsuo-lab/weblab-10b`
JSLM-Alpha-7B	FS	08/2023	`stabilityai/japanese-stablelm-base-alpha-7b`
CALM2-7B	FS	11/2023	`cyberagent/calm2-7b`
Swallow-70B	CP	12/2023	`tokyotech-llm/Swallow-70b-hf`
KARAKURI-LM-70B	CP	01/2024	`karakuri-ai/karakuri-lm-70b-v0.1`
JSLM-Beta-70B	CP	11/2023	`stabilityai/japanese-stablelm-base-beta-70b`
Swallow-MX-8x7B	CP	03/2024	`tokyotech-llm/Swallow-MX-8x7b-NVE-v0.1`
ABEJA-8x7B	CP	04/2024	`abeja/Mixtral-8x7B-Instruct-v0.1-japanese`
Nekomata-14B	CP	12/2023	`rinna/nekomata-14b`
ELYZA-13B	CP	12/2023	`elyza/ELYZA-japanese-Llama-2-13b`
Youri-7B	CP	10/2023	`rinna/youri-7b`
SambaLingo-Japanese-7B	CP	02/2024	`sambanovasystems/SambaLingo-Japanese-Base`
Swallow-MS-7B	CP	03/2024	`tokyotech-llm/Swallow-MS-7b-v0.1`
RakutenAI-7B	CP	03/2024	`Rakuten/RakutenAI-7B`
JSLM-Gamma-7B	CP	10/2023	`stabilityai/japanese-stablelm-base-gamma-7b`

Table 1: A selected list of base LLMs for Japanese as of the end of April 2024. In the **Type** column, FS denotes models built from scratch, while CP denotes models built out of other LLMs via continual pre-training. **Model ID** indicates the corresponding repository id on Hugging Face Transformers library (Wolf et al., 2020).

2. Development Trends

2.1. Scaling-up of the Model Size

To our knowledge, the largest Japanese LLM in terms of the number of parameters was ABEJA-2.7B (ABEJA, 2022) before the advent of ChatGPT (OpenAI, 2022). After ChatGPT has been released, development has progressed on an even larger scale, with several models exceeding 10 billion parameters. Table 1 presents a selected list of Japanese LLMs that are publicly available.

As the number of parameters increases, escalating computational costs associated with pre-training from scratch arise. Therefore, there has been a growing trend towards a method known as continual pre-training, wherein an existing LLM serves as a base, and additional Japanese corpora are used to train the model further to enhance its capabilities in Japanese. Swallow (Okazaki et al., 2023), ELYZA (Sasaki et al., 2023b), JSLM-Beta (Lee et al., 2023a), and Youri (Sawada et al., 2024) are developed based on Llama 2 (Touvron et al., 2023), while Swallow-MS (Okazaki et al., 2024), JSLM-Gamma (Lee et al., 2023b), and RakutenAI (Levine et al., 2024) are on Mistral (Jiang et al., 2023).

An intriguing approach to continual pre-training is to adapt Chinese-centric LLMs, such as Qwen (Bai et al., 2023) and Baichuan 2 (Yang et al., 2023), to Japanese LLMs. Japanese and Chinese people share the characteristic of using thousands of Kanjis (or Chinese characters) in everyday communication. Therefore, the knowledge of Kanji in a Chinese LLM could be beneficial. Nekomata (Sawada et al., 2024) is a pioneering model in this direction, which is constructed upon Qwen.

2.2. Human Alignment

Aligning models with human expectations is an integral part of building LLMs (Wang et al., 2023b). Instruction-tuning (Wei et al., 2022) is widely applied to Japanese LLMs, as shown in Table 2.

Some works explore more advanced methods to pursue better alignment. LLM-jp-13B-DPO (Kiyomaru et al., 2024), CALM2-7B-DPO (Jinnai, 2024), Shisa-7B (Lin et al., 2023), and SambaLingo-Japanese-7B-Chat (Csaki et al., 2024) adopt DPO (Rafailov et al., 2023), while KARAKURI-LM-70B-Chat (KARAKURI, 2024) employ SteerLM (Dong et al., 2023) to reflect the human preferences.

A challenging problem in LLM alignment is how to construct labeled training data. Early Japanese LLMs rely on machine-translated datasets such as kunishou/databricks-dolly-15k-ja (Kuniyoshi, 2023a) and kunishou/oasst1-89k-ja (Kuniyoshi, 2023b). More recently, there has been a shift toward building datasets without the help of machine translation. The llm-japanese-dataset (Suzuki et al., 2023) has been devised by reusing Japanese NLP datasets for instruction-tuning. The ichikara-instruction (RIKEN, 2023) is an initiative to build Japanese instruction-tuning datasets annotated by Japanese experts.

2.3. Domain Adaptation

LLMs have significant applications in domains that require specialized knowledge, such as medicine, chemistry, and finance (Zhou et al., 2023; Ho et al., 2024; Li et al., 2023). Various studies have been done on infusing domain-specific knowledge into Japanese LLMs.

Takahashi et al. (2024) demonstrate that including business-related web pages and patents in the pre-training stage can enhance the performance of business-related question-answering tasks. Similarly, Hirano and Imajo (2024) show the effectiveness of domain-specific continual pre-training in the financial domain. Sukeda et al. (2023) investigate domain-specific instruction-tuning in the medical domain. Taking a completely different approach, Akiba et al. (2024) claim that their proposed method, Evolutionary Model Merge, can successfully merge models from dif-

Model	Release Date	Model ID
LLM-jp-13B-DPO	02/2024	`llm-jp/llm-jp-13b-dpo-lora-hh_rlhf_ja-v1.1`
PLaMo-13B-Instruct	11/2023	`pfnet/plamo-13b-instruct`
Stockmark-13B-Instruct	11/2023	`stockmark/stockmark-13b-instruct`
Weblab-10B-Instruct	08/2023	`matsuo-lab/weblab-10b-instruction-sft`
JSLM-Alpha-7B-Instruct	10/2023	`stabilityai/japanese-stablelm-instruct-alpha-7b-v2`
CALM2-7B-DPO	01/2024	`cyberagent/calm2-7b-chat-dpo-experimental`
ao-Karasu-72B	03/2024	`lightblue/ao-karasu-72B`
Swallow-70B-Instruct	12/2023	`tokyotech-llm/Swallow-70b-instruct-hf`
KARAKURI-LM-70B-Chat	01/2024	`karakuri-ai/karakuri-lm-70b-chat-v0.1`
JSLM-Beta-70B-Instruct	11/2023	`stabilityai/japanese-stablelm-instruct-beta-70b`
Nekomata-14B-Instruct	12/2023	`rinna/nekomata-14b-instruction`
Qarasu-14B	12/2023	`lightblue/qarasu-14B-chat-plus-unleashed`
ELYZA-13B-Instruct	12/2023	`elyza/ELYZA-japanese-Llama-2-13b-instruct`
Youri-7B-Chat	10/2023	`rinna/youri-7b-chat`
Houou-7B	12/2023	`moneyforward/houou-instruction-7b-v2`
SambaLingo-Japanese-7B-Chat	02/2024	`sambanovasystems/SambaLingo-Japanese-Chat`
Deepreneur-blue-lizard-7B	02/2024	`Deepreneur/blue-lizard`
RakutenAI-7B-Chat	03/2024	`Rakuten/RakutenAI-7B-chat`
JSLM-Gamma-7B-Instruct	10/2023	`stabilityai/japanese-stablelm-instruct-gamma-7b`
ChatNTQ-JA-7B	12/2023	`NTQAI/chatntq-ja-7b-v1.0`
Shisa-Gamma-7B	12/2023	`augmxnt/shisa-gamma-7b-v1`
Shisa-7B	12/2023	`augmxnt/shisa-7b-v1`
Karasu-7B	12/2023	`lightblue/karasu-7B-chat-plus-unleashed`

Table 2: A selected list of instruction-tuned LLMs for Japanese as of the end of April 2024.

Model	Release Date	Model ID
EvoVLM-JP-7B	03/2024	`SakanaAI/EvoVLM-JP-v1-7B`
Heron-GIT-7B	04/2024	`turing-motors/heron-chat-git-ja-stablelm-base-7b-v1`
Heron-BLIP-7B	03/2024	`turing-motors/heron-chat-blip-ja-stablelm-base-7b-v1-llava-620k`
JSVLM-7B	11/2023	`stabilityai/japanese-stable-vlm`

Table 3: A selected list of LVLMs for Japanese as of the end of April 2024.

ferent domains (e.g., a Japanese LLM and an English Math LLM).

2.4. Multimodality

It has been established that the combination of an LLM and a visual encoder can effectively produce a vision-and-language model (VLM) (Zhu et al., 2023; Liu et al., 2023a). VLMs based on LLMs are referred to as LVLMs. Table 3 shows a list of Japanese LVLMs that are publicly available.

It is also worthwhile to investigate the applications of LLMs to other modalities. Hono et al. (2023) demonstrate the effectiveness of combining a Japanese LLM with an audio encoder on automatic speech recognition in Japanese.

2.5. Dense Retrieval Models

Retrieval-augmented generation (RAG) is a common technique for incorporating external knowledge into LLMs (Gao et al., 2023b). Retrieval, a process of collecting relevant documents to the query, is often employed by dense retrieval models (Zhao et al., 2022). Therefore, it is highly beneficial to optimize such models for Japanese.

Chen et al. (2023b) and Tsukagoshi et al. (2023) present Japanese sentence embedding models based on SimCSE (Gao et al., 2021). Clavié (2023) introduces a Japanese version of ColBERT (Khattab and Zaharia, 2020).

3. Evaluation Trends

3.1. Traditional NLP Benchmarks

The Japanese NLP community has a longstanding history of constructing various datasets for specific tasks, such as natural language inference (Kawazoe et al., 2015; Hayashibe, 2020; Yanaka and Mineshima, 2022) and named entity recognition (Iwakura et al., 2016; Yada et al., 2022). More recently, JGLUE (Kurihara et al., 2022) has been developed to evaluate the performance of pre-trained language models comprehensively. These datasets are widely used to assess the capabilities of Japanese LLMs in a few-shot or zero-shot manner.

Benchmarks that holistically assess Japanese capabilities across multiple datasets are also emerging. As of this writing, there are two dominant frameworks:

JP Language Model Evaluation Harness (StabilityAI, 2023b) is constructed as the Japanese version of Language Model Evaluation Harness (Gao et al., 2023a), which is a unifying framework for few-shot evaluation of language models.

llm-jp-eval (Han et al., 2024) is a framework developed with a focus on the generative abilities of language models. Unlike JP Language Model Evaluation Harness, which employs the log-likelihood of output labels for evaluation,

llm-jp-eval uses the results of generated texts, presenting a more challenging evaluation method for the models.

3.2. Assessing Real-world Capabilities

Conventional NLP benchmarks have often been criticized for being unsuitable for measuring the ability to handle real-world user interactions (Zheng et al., 2023; Liu et al., 2023b). Benchmarks employing open-ended questions have been proposed to address these concerns and provide a more realistic assessment of Japanese LLMs.

Japanese MT-bench (StabilityAI, 2023a) is built as the Japanese version of MT-bench (Zheng et al., 2023). It includes 80 questions from eight categories: writing, role-play, reasoning, math, coding, extraction, STEM, and humanities. Some questions have been modified to fit with Japanese culture.

Japanese Vicuna QA Benchmark (Sun et al., 2024) is the Japanese version of Vicuna benchmark (Chiang et al., 2023), which is the predecessor of MT-Bench.

Rakuda Benchmark (Passaglia and Yu, 2023) is an evaluation framework based on model answers to 40 open-ended questions on Japanese geography, history, politics, and society.

ELYZA-tasks-100 (Sasaki et al., 2023a) is an evaluation framework based on model answers to 100 complex and diverse tasks, including tasks testing summarization, correction, abstraction, induction, and other skills.

3.3. Other Benchmarks

LLM evaluation can be conducted from various perspectives (Chang et al., 2023b; Chen et al., 2023a). Here we highlight some of the attempts specific to the Japanese language or culture.

JMMLU (Yin et al., 2024) is constructed as the Japanese version of MMLU benchmark (Hendrycks et al., 2021). Unlike the original MMLU, JMMLU includes questions about Japanese history, geography, civics, and idioms.

JFLD (Morishita et al., 2024) is a dataset for evaluating the deductive reasoning capabilities of LLMs in Japanese. It aims to measure pure reasoning skills isolated from knowledge.

Japanese Language Model Financial Evaluation Harness (Hirano, 2024) is an evaluation framework on the Japanese financial domain, consisting of five subtasks.

JA-VLM-Bench-In-the-Wild (Akiba et al., 2024) is a dataset for evaluating Japanese VLMs. It collects 42 images with 50 questions. Notably, the images are characterized by the Japanese culture.

Heron-Bench (Inoue et al., 2024) is a VLM benchmark for Japanese. It compiles 21 images with 102 questions. Each image features Japanese objects, such as food, landscape, and anime.

3.4. Leaderboards

LLM leaderboards are convenient tools for quickly understanding which is the best-performing model. Some notable examples include Chatbot Arena Leaderboard (Chiang et al., 2024), C-Eval Leaderboard (Huang et al., 2023), and Open Ko-LLM Leaderboard (Park et al., 2023).

As for Japanese, the most up-to-date leaderboard is Nejumi LLM Leaderboard (Kamata, 2023), which compiles the results of llm-jp-eval and Japanese MT-bench. As of this writing, the scores of Japanese LLMs are lower than proprietary LLMs like GPT-4. This gap underscores an opportunity for further advancements in developing open LLMs capable of understanding and generating Japanese texts.

4. Conclusion

In this report, we summarize the trends in the development and evaluation of Japanese LLMs. We hope this study will provide a better understanding of what is going on.

We acknowledge that the contents of this report could be quickly outdated. Therefore, we strongly recommend keeping up with the latest information on our GitHub repository `https://github.com/llm-jp/awesome-japanese-llm`.

Acknowledgments

We would like to thank all the people who have committed to the research and development of Japanese large language models. We are particularly grateful to Sam Passaglia and Akim Mousterou for translating the contents of our GitHub repository into English and French, respectively.

References

ABEJA. (2022). gpt-neox-japanese-2.7b. `https://huggingface.co/abeja/gpt-neox-japanese-2.7b`.

Akiba, T., Shing, M., Tang, Y., Sun, Q., and Ha, D. (2024). Evolutionary optimization of model merging recipes. *CoRR*, abs/2403.13187.

Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., Hui, B., Ji, L., Li, M., Lin, J., Lin, R., Liu, D., Liu, G., Lu, C., Lu, K., Ma, J., Men, R., Ren, X., Ren, X., Tan, C., Tan, S., Tu, J., Wang, P., Wang, S., Wang, W., Wu, S., Xu, B., Xu, J., Yang, A., Yang, H., Yang, J., Yang, S., Yao, Y., Yu, B., Yuan, H., Yuan, Z., Zhang, J., Zhang, X., Zhang, Y., Zhang, Z., Zhou, C., Zhou, J., Zhou, X., and Zhu, T. (2023). Qwen technical report. *CoRR*, abs/2309.16609.

Chang, K., Cramer, M., Soni, S., and Bamman, D. (2023a). Speak, memory: An archaeology of books known to ChatGPT/GPT-4. In Houda Bouamor, et al., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7312–7327, Singapore, December. Association for Computational Linguistics.

Chang, Y., Wang, X., Wang, J., Wu, Y., Zhu, K., Chen, H., Yang, L., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q., and Xie, X. (2023b). A survey on evaluation of large language models. *CoRR*, abs/2307.03109.

Chen, H., Jiao, F., Li, X., Qin, C., Ravaut, M., Zhao, R., Xiong, C., and Joty, S. (2023a). Chatgpt's one-year anniversary: Are open-source large language models catching up? *CoRR*, abs/2311.16989.

Chen, Z., Handa, H., and Shirahama, K. (2023b). JCSE: contrastive learning of japanese sentence embeddings and its applications. *CoRR*, abs/2301.08193.

Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., Stoica, I., and Xing, E. P. (2023). Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March.

Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhang, H., Zhu, B., Jordan, M., Gonzalez, J. E., and Stoica, I. (2024). Chatbot arena: An open platform for evaluating llms by human preference.

Clavié, B. (2023). Jacolbert and hard negatives, towards better japanese-first embeddings for retrieval: Early technical report. *CoRR*, abs/2312.16144.

Csaki, Z., Li, B., Li, J., Xu, Q., Pawakapan, P., Zhang, L., Du, Y., Zhao, H., Hu, C., and Thakker, U. (2024). Sambalingo: Teaching large language models new languages.

Dong, Y., Wang, Z., Sreedhar, M., Wu, X., and Kuchaiev, O. (2023). SteerLM: Attribute conditioned SFT as an (user-steerable) alternative to RLHF. In Houda Bouamor, et al., editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11275–11288, Singapore, December. Association for Computational Linguistics.

Gao, T., Yao, X., and Chen, D. (2021). SimCSE: Simple contrastive learning of sentence embeddings. In Marie-Francine Moens, et al., editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Gao, L., Tow, J., Abbasi, B., Biderman, S., Black, S., DiPofi, A., Foster, C., Golding, L., Hsu, J., Le Noac'h, A., Li, H., McDonell, K., Muennighoff, N., Ociepa, C., Phang, J., Reynolds, L., Schoelkopf, H., Skowron, A., Sutawika, L., Tang, E., Thite, A., Wang, B., Wang, K., and Zou, A. (2023a). A framework for few-shot language model evaluation, 12.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Guo, Q., Wang, M., and Wang, H. (2023b). Retrieval-augmented generation for large language models: A survey. *CoRR*, abs/2312.10997.

Han, N., Ueda, N., Otake, M., Katsumata, S., Kamata, K., Kiyomaru, H., Kodama, T., Sugawara, S., Chen, B., Matsuda, H., Miyao, Y., Murawaki, Y., and Ryu, K. (2024). llm-jp-eval. `https://github.com/llm-jp/llm-jp-eval`.

Hayashibe, Y. (2020). Japanese realistic textual entailment corpus. In Nicoletta Calzolari, et al., editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6827–6834, Marseille, France, May. European Language Resources Association.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. (2021). Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Hirano, M. and Imajo, K. (2024). Construction of domain-specified japanese large language model for finance through continual pre-training.

Hirano, M. (2024). Construction of a japanese financial benchmark for large language models. *CoRR*, abs/2403.15062.

Ho, X., Nguyen, A. D., Dao, T., Jiang, J., Chida, Y., Sugimoto, K., To, H. Q., Boudin, F., and Aizawa, A. (2024). A survey of pre-trained language models for processing scientific text. *CoRR*, abs/2401.17824.

Hono, Y., Mitsuda, K., Zhao, T., Mitsui, K., Wakatsuki, T., and Sawada, K. (2023). An integration of pre-trained speech and language models for end-to-end speech recognition. *CoRR*, abs/2312.03668.

Hornyak, T. (2023). Why japan is building its own version of chatgpt. *Nature*.

Huang, Y., Bai, Y., Zhu, Z., Zhang, J., Zhang, J., Su, T., Liu, J., Lv, C., Zhang, Y., Lei, J., Fu, Y., Sun, M., and He, J. (2023). C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. In Alice Oh, et al., editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Inoue, Y., Sasaki, K., Ochi, Y., Fujii, K., Tanahashi, K., and Yamaguchi, Y. (2024). Heron-bench: A benchmark for evaluating vision language models in japanese.

Iwakura, T., Komiya, K., and Tachibana, R. (2016). Constructing a Japanese basic named entity corpus of various genres. In Xiangyu Duan, et al., editors, *Proceedings of the Sixth Named Entity Workshop*, pages 41–46, Berlin, Germany, August. Association for Computational Linguistics.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de Las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. (2023). Mistral 7b. *CoRR*, abs/2310.06825.

Jinnai, Y. (2024). calm2-7b-chat-dpo-experimental. `https://huggingface.co/cyberagent/calm2-7b-chat-dpo-experimental`.

Kamata, K. (2023). Nejumi LLM Leaderboard. `https://api.wandb.ai/links/wandb-japan/xm2pju5m`.

Kandpal, N., Deng, H., Roberts, A., Wallace, E., and Raffel, C. (2023). Large language models struggle to learn long-tail knowledge. In Andreas Krause, et al., editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 15696–15707. PMLR.

KARAKURI. (2024). KARAKURI LM. `https://huggingface.co/karakuri-ai/karakuri-lm-70b-chat-v0.1`.

Kawazoe, A., Tanaka, R., Mineshima, K., and Bekki, D. (2015). An inference problem set for evaluating semantic theories and semantic processing systems for japanese. In Mihoko Otake, et al., editors, *New Frontiers in Artificial Intelligence - JSAI-isAI 2015 Workshops,*

LENLS, JURISIN, AAA, HAT-MASH, TSDAA, ASD-HR, and SKL, Kanagawa, Japan, November 16-18, 2015, Revised Selected Papers, volume 10091 of *Lecture Notes in Computer Science*, pages 58–65.

Khattab, O. and Zaharia, M. (2020). Colbert: Efficient and effective passage search via contextualized late interaction over BERT. In Jimmy X. Huang, et al., editors, *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 39–48. ACM.

Kiyomaru, H., Matsuda, H., Suzuki, J., Han, N., Sugawara, S., Sasaki, S., Kurita, S., Nakamura, T., Kodama, T., and Okamoto, T. (2024). llm-jp-13b-dpo-lora-hh_rlhf_ja-v1.1. https://huggingface.co/llm-jp/llm-jp-13b-dpo-lora-hh_rlhf_ja-v1.1.

Kudo, T. (2018). Subword regularization: Improving neural network translation models with multiple subword candidates. In Iryna Gurevych et al., editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia, July. Association for Computational Linguistics.

Kunitsu, Y. (2023). The potential of gpt-4 as a support tool for pharmacists: analytical study using the japanese national examination for pharmacists. *JMIR Medical Education*, 9:e48452.

Kuniyoshi, S. (2023a). kunishou/databricks-dolly-15k-ja. https://huggingface.co/datasets/kunishou/databricks-dolly-15k-ja.

Kuniyoshi, S. (2023b). kunishou/oasst1-89k-ja. https://huggingface.co/datasets/kunishou/oasst1-89k-ja.

Kurihara, K., Kawahara, D., and Shibata, T. (2022). JGLUE: Japanese general language understanding evaluation. In Nicoletta Calzolari, et al., editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2957–2966, Marseille, France, June. European Language Resources Association.

Lee, M., Nakamura, F., Shing, M., McCann, P., Akiba, T., and Orii, N. (2023a). Japanese StableLM Base Beta 70B. https://huggingface.co/stabilityai/japanese-stablelm-base-beta-70b.

Lee, M., Nakamura, F., Shing, M., McCann, P., Akiba, T., and Orii, N. (2023b). Japanese StableLM Base Gamma 7B. https://huggingface.co/stabilityai/japanese-stablelm-base-gamma-7b.

Levine, A., Huang, C., Wang, C., Batista, E., Szymanska, E., Ding, H., Chou, H. W., Pessiot, J., Effendi, J., Chiu, J., Ohlhus, K. T., Chopra, K., Shinzato, K., Murakami, K., Xiong, L., Chen, L., Kubota, M., Tkachenko, M., Lee, M., Takahashi, N., Jwalapuram, P., Tatsushima, R., Jain, S., Yadav, S. K., Cai, T., Chen, W., Xia, Y., Nakayama, Y., and Higashiyama, Y. (2024). Rakutenai-7b: Extending large language models for japanese. *CoRR*, abs/2403.15484.

Li, Y., Wang, S., Ding, H., and Chen, H. (2023). Large language models in finance: A survey. In *4th ACM International Conference on AI in Finance, ICAIF 2023, Brooklyn, NY, USA, November 27-29, 2023*, pages 374–382. ACM.

Lin, L., Durbin, J., Sato, M., and von Bock, F. (2023). Shisa 7B. https://huggingface.co/augmxnt/shisa-7b-v1.

Liu, H., Li, C., Li, Y., and Lee, Y. J. (2023a). Improved baselines with visual instruction tuning. *CoRR*, abs/2310.03744.

Liu, X., Lei, X., Wang, S., Huang, Y., Feng, Z., Wen, B., Cheng, J., Ke, P., Xu, Y., Tam, W. L., Zhang, X., Sun, L., Wang, H., Zhang, J., Huang, M., Dong, Y., and Tang, J. (2023b). Alignbench: Benchmarking chinese alignment of large language models. *CoRR*, abs/2311.18743.

METI. (2024). GENIAC. https://www.meti.go.jp/english/policy/mono_info_service/geniac/index.html.

Morishita, T., Yamaguchi, A., Morio, G., Hikaru, T., Imaichi, O., and Sogawa, Y. (2024). Jfld: A japanese benchmark for deductive reasoning based on formal logic. In *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation*.

Nakao, T., Miki, S., Nakamura, Y., Kikuchi, T., Nomura, Y., Hanaoka, S., Yoshikawa, T., Abe, O., et al. (2024). Capability of gpt-4v (ision) in the japanese national medical licensing examination: Evaluation study. *JMIR Medical Education*, 10(1):e54393.

Okazaki, N., Mizuki, S., Iida, H., Loem, M., Hirai, S., Hattori, K., Ohi, M., Yokota, R., Fujii, K., and Nakamura, T. (2023). Swallow. https://huggingface.co/tokyotech-llm/Swallow-70b-hf.

Okazaki, N., Mizuki, S., Iida, H., Loem, M., Hirai, S., Hattori, K., Ohi, M., Yokota, R., Fujii, K., and Nakamura, T. (2024). Swallow-MS-7b-v0.1. https://huggingface.co/tokyotech-llm/Swallow-MS-7b-v0.1.

OpenAI. (2022). ChatGPT. https://openai.com/chatgpt.

OpenAI. (2023). GPT-4 technical report. *CoRR*, abs/2303.08774.

OpenAI. (2024). Introducing OpenAI Japan. https://openai.com/blog/introducing-openai-japan.

Park, C., Lee, H., Park, H., Kim, H., Kim, S., Cho, S., Kim, S., and Lee, S. (2023). Open ko-llm leaderboard. https://huggingface.co/spaces/upstage/open-ko-llm-leaderboard.

Passaglia, S. and Yu, S. (2023). Rakuda Benchmark. https://github.com/yuzu-ai/japanese-llm-ranking.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. (2023). Direct preference optimization: Your language model is secretly a reward model. In Alice Oh, et al., editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023,*

NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023.

RIKEN. (2023). Release of ichikara-instruction (in Japanese). https://liat-aip.sakura.ne.jp/wp/llm%E3%81%AE%E3%81%9F%E3%82%81%E3%81%AE%E6%97%A5%E6%9C%AC%E8%AA%9E%E3%82%A4%E3%83%B3%E3%82%B9%E3%83%88%E3%83%A9%E3%82%AF%E3%82%B7%E3%83%A7%E3%83%B3%E3%83%87%E3%83%BC%E3%82%BF%E4%BD%9C%E6%88%90/llm%E3%81%AE%E3%81%9F%E3%82%81%E3%81%AE%E6%97%A5%E6%9C%AC%E8%AA%9E%E3%82%A4%E3%83%B3%E3%82%B9%E3%83%88%E3%83%A9%E3%82%AF%E3%82%B7%E3%83%A7%E3%83%B3%E3%83%87%E3%83%BC%E3%82%BF-%E5%85%AC%E9%96%8B/.

Sasaki, A., Hirakawa, M., Horie, S., and Nakamura, T. (2023a). Elyza-tasks-100. https://huggingface.co/datasets/elyza/ELYZA-tasks-100.

Sasaki, A., Hirakawa, M., Horie, S., Nakamura, T., Passaglia, S., and Oba, D. (2023b). Elyza-japanese-llama-2-13b. https://huggingface.co/elyza/ELYZA-japanese-Llama-2-13b.

Sawada, K., Zhao, T., Shing, M., Mitsui, K., Kaga, A., Hono, Y., Wakatsuki, T., and Mitsuda, K. (2024). Release of pre-trained models for the Japanese language. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 5.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In Katrin Erk et al., editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August. Association for Computational Linguistics.

Shinkawa, T. (2012). Substitutes for immigrants? social policy responses to population decreases in japan. *American behavioral scientist*, 56(8):1123–1138.

StabilityAI. (2023a). Japanese MT-Bench. https://github.com/Stability-AI/FastChat.

StabilityAI. (2023b). JP Language Model Evaluation Harness. https://github.com/Stability-AI/lm-evaluation-harness.

Sukeda, I., Suzuki, M., Sakaji, H., and Kodera, S. (2023). Jmedlora: Medical domain adaptation on japanese large language models using instruction-tuning. *CoRR*, abs/2310.10083.

Sun, Y., Wan, Z., Ueda, N., Yahata, S., Cheng, F., Chu, C., and Kurohashi, S. (2024). Rapidly developing high-quality instruction data and evaluation benchmark for large language models with minimal human effort: A case study on japanese. In *The 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*.

Suzuki, M., Hirano, M., and Sakaji, H. (2023). From base to conversational: Japanese instruction dataset and tuning large language models. In Jingrui He, et al., editors, *IEEE International Conference on Big Data, BigData 2023, Sorrento, Italy, December 15-18, 2023*, pages 5684–5693. IEEE.

Takahashi, K., Omi, T., Arima, K., and Ishigaki, T. (2024). Pretraining and updating language- and domain-specific large language model: A case study in japanese business domain.

Tanaka, Y., Nakata, T., Aiga, K., Etani, T., Muramatsu, R., Katagiri, S., Kawai, H., Higashino, F., Enomoto, M., Noda, M., et al. (2024). Performance of generative pretrained transformer on the national medical licensing examination in japan. *PLOS Digital Health*, 3(1):e0000433.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Canton-Ferrer, C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Tsukagoshi, H., Sasano, R., and Takeda, K. (2023). Japanese simcse technical report. *CoRR*, abs/2310.19349.

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W. X., Wei, Z., and Wen, J. (2023a). A survey on large language model based autonomous agents. *CoRR*, abs/2308.11432.

Wang, Y., Zhong, W., Li, L., Mi, F., Zeng, X., Huang, W., Shang, L., Jiang, X., and Liu, Q. (2023b). Aligning large language models with human: A survey. *CoRR*, abs/2307.12966.

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V. (2022). Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Transformers: State-of-the-art natural language processing. In Qun Liu et al., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October. Association for Computational Linguistics.

Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., Zheng, R., Fan, X., Wang, X., Xiong, L., Zhou, Y., Wang, W., Jiang, C.,

Zou, Y., Liu, X., Yin, Z., Dou, S., Weng, R., Cheng, W., Zhang, Q., Qin, W., Zheng, Y., Qiu, X., Huan, X., and Gui, T. (2023). The rise and potential of large language model based agents: A survey. *CoRR*, abs/2309.07864.

Yada, S., Nakamura, Y., Wakamiya, S., and Aramaki, E. (2022). Real-mednlp: Overview of real document-based medical natural language processing task. In *Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies*, pages 285–296.

Yanaka, H. and Mineshima, K. (2022). Compositional evaluation on Japanese textual entailment and similarity. *Transactions of the Association for Computational Linguistics*, 10:1266–1284.

Yang, A., Xiao, B., Wang, B., Zhang, B., Bian, C., Yin, C., Lv, C., Pan, D., Wang, D., Yan, D., Yang, F., Deng, F., Wang, F., Liu, F., Ai, G., Dong, G., Zhao, H., Xu, H., Sun, H., Zhang, H., Liu, H., Ji, J., Xie, J., Dai, J., Fang, K., Su, L., Song, L., Liu, L., Ru, L., Ma, L., Wang, M., Liu, M., Lin, M., Nie, N., Guo, P., Sun, R., Zhang, T., Li, T., Li, T., Cheng, W., Chen, W., Zeng, X., Wang, X., Chen, X., Men, X., Yu, X., Pan, X., Shen, Y., Wang, Y., Li, Y., Jiang, Y., Gao, Y., Zhang, Y., Zhou, Z., and Wu, Z. (2023). Baichuan 2: Open large-scale language models. *CoRR*, abs/2309.10305.

Yin, Z., Wang, H., Horio, K., Kawahara, D., and Sekine, S. (2024). Should we respect llms? A cross-lingual study on the influence of prompt politeness on LLM performance. *CoRR*, abs/2402.14531.

Zhao, W. X., Liu, J., Ren, R., and Wen, J. (2022). Dense text retrieval based on pretrained language models: A survey. *CoRR*, abs/2211.14876.

Zheng, L., Chiang, W., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., and Stoica, I. (2023). Judging llm-as-a-judge with mt-bench and chatbot arena. In Alice Oh, et al., editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Zhou, H., Gu, B., Zou, X., Li, Y., Chen, S. S., Zhou, P., Liu, J., Hua, Y., Mao, C., Wu, X., Li, Z., and Liu, F. (2023). A survey of large language models in medicine: Progress, application, and challenge. *CoRR*, abs/2311.05112.

Zhu, D., Chen, J., Shen, X., Li, X., and Elhoseiny, M. (2023). Minigpt-4: Enhancing vision-language understanding with advanced large language models. *CoRR*, abs/2304.10592.