Fostering Business Innovation with AI: Performance of Fine-Tuned Japanese Language Models under Resource Restrictions

Chau Nguyen[1], Thanh Tran[1], Son T. Luu[1], Nguyen-Khang Le[1], Dinh-Truong Do[1], Shintaro Kawamura[2], Shoichi Naito[2], Yuki Mogi[2], and Le-Minh Nguyen[1]

[1] Japan Advanced Institute of Science and Technology, Japan
[2] Ricoh Company, Ltd., Japan
chau.nguyen@jaist.ac.jp

Abstract. In recent times, the utility of large language models (LLMs) in a diverse range of practical applications has been well-established for both individual and organizational purposes. Enterprises may opt to build proprietary LLMs using their confidential datasets to align with their specific needs. Our research explores the feasibility of such enterprises, especially those with constrained GPU capabilities (e.g., micro-enterprises), in developing their own LLMs tailored to their private data. We examine the effectiveness of fine-tuning foundation LLMs for natural language processing tasks tailored to the Japanese language. Our empirical analysis covers three distinct tasks varying in complexity: text classification, machine reading comprehension, and summarization. The study involves a critical comparison across different model scales and focuses on language specialization. We specifically compare the performance of fine-tuned foundation LLMs with parameters scaling from 7 billion to 13 billion. Moreover, we contrast LLMs that are specialized in the Japanese language with those primarily designed for English but offer Japanese language support. Our results uncover a complex relationship between the size of the model, its language specialization, and the complexity of the given task. We observe that, while larger models tend to perform better, there are task-dependent intricacies and variations based on the foundational language focus of LLMs that can influence the outcomes distinctively. Our research is particularly informative for companies with limited GPU resources, as we address the challenges of fine-tuning and optimizing inference by providing practical comparative analyses. The insights derived from our study are intended to serve as a strategic guide for practitioners and organizations with constrained computational resources, enabling them to efficiently utilize foundational LLMs for processing the Japanese language.

Keywords: Business-specific AI Application · Japanese Natural Language Processing · Computational Resource Efficiency.

1 Introduction

The integration of foundation Large Language Models (LLMs) into the realm of natural language processing (NLP) has revolutionized the field, particularly in terms of language task automation and enhancement. As these models become more ubiquitous, a particular area of interest has arisen concerning their applicability to the Japanese language—a complex linguistic scenario that requires specialized attention. This paper is driven by the need to explore this under-investigated area, with a specific focus on assessing the adaptability of LLMs for small enterprises that face limited computational resources.

Our research is designed to test the feasibility of using foundation LLMs for three distinct but essential NLP tasks: text classification, machine reading comprehension, and text summarization, all within the context of the Japanese language. We take a two-pronged approach in our comparative analysis. Firstly, we evaluate the impact of model size on task performance by contrasting LLMs with 7 billion and 13 billion parameters. This comparison provides us with insights into whether and how increased model size translates into enhanced NLP task handling. Secondly, we examine the manner in which language specialization influences performance, comparing models primarily trained in English, but with Japanese support (English-focused Japanese-supported LLMs, or Japanese-supported LLMs for short), to those specifically designed for the Japanese language (Japanese-focused LLMs).

The intricacies uncovered through our study reveal a complex relationship between model size and task performance within the scope of language processing. While an increase in model size typically correlates with

better performance, the results are not consistent across all tasks, particularly when language specialization is considered. It is noteworthy that in specific contexts, Japanese-supported LLMs—though primarily trained on English—demonstrate an impressive ability to outperform their Japanese-focused counterparts, particularly in simpler tasks such as text classification. However, the overall trend suggests that these English-focused Japanese-supported LLMs, despite holding their ground in basic NLP tasks, encounter limitations when dealing with more complex assignments such as machine reading comprehension and text summarization. While they perform reasonably well, they seldom achieve the superior performance exhibited by the models tailored exclusively for the Japanese language.

Underscoring these points, the main findings of our study have highlighted that the choice of LLM for a given task is hardly straightforward, especially when balancing performance with available computational resources. Specifically, while English-focused Japanese-supported LLMs can be competitive with or sometimes even surpass the capabilities of Japanese-focused models in less demanding tasks, their efficacy is mitigated when faced with the intricacies of more advanced NLP challenges. In these scenarios, the Japanese-focused LLMs consistently demonstrate a superior ability to manage the complexities of the language, proving to be more suitable for navigating the nuanced linguistic features inherent to tasks like machine reading comprehension and text summarization. For small organizations looking to employ LLMs effectively, recognizing these nuances, as they must consider the benefits and limitations of each model against the context of their limited computational resources.

This exploration carries particular significance for micro-enterprises and other entities with limited GPU capacity. Through our investigation, we analyze the intricacies of fine-tuning and optimizing LLMs for efficient inference, despite resource constraints. Our research provides these organizations with a roadmap to utilize foundational LLMs effectively, thereby enabling them to leverage these powerful tools to enhance their NLP capabilities in the Japanese language. In summary, this paper offers a thorough examination of the robustness of foundation LLMs across varying tasks in the Japanese context, coupled with practical guidance for those with limited computational capacity seeking to maximize the benefits of these models.

2 Related works

Natural Language Understanding (NLU) and Natural Language Generation (NLG) are the two fundamental tasks in NLP that require the ability of language comprehension of LLMs [28]. To evaluate and compare the performances of various LLMs, we employ three tasks: text classification and machine reading comprehension (MRC) for NLU, and text summarization for NLG. In this paper, we focus on evaluating the ability of LLMs in Japanese texts and we investigate and choose available Japanese corpora that serve specific NLP tasks. The JGLUE [13] is one of the datasets constructed for NLU tasks in the Japanese language. JGLUE consists of several sub-tasks with corresponding datasets, including MARC-ja for binary sentiment analysis, JSQuAD for machine reading comprehension, JSTS for text similarity, and JNLI for textual entailment. Apart from JGLUE, there are other datasets that serve for the NLU task in Japanese, such as WRIME [23] and chABSA [16]. These two datasets are manually annotated and used for the aspect-based sentiment analysis task for Japanese texts. In addition, JaSQuAD [22] is a large-scale MRC dataset with nearly 39K question-answer pairs constructed based on Japanese Wikipedia articles. Besides, XL-SUM [8] and WikiLingua [14] are the two large-scale multilingual datasets for text summarization that support the Japanese language. However, WikiLingua is a cross-lingual text summarization dataset where the target and source text can be in different languages, while the languages between the target and source in XL-SUM are the same. Therefore, we choose the XL-SUM dataset with a Japanese text subset for the text summarization task and the JGLUE for the text classification and machine reading comprehension task.

Large Language Models (LLMs) have showcased their proficiency across a diverse array of natural language processing (NLP) tasks [3, 25]. Typically constructed upon the Transformer model [26], these LLMs undergo training on an extensive corpus of self-supervised text data [29]. Noteworthy examples of such LLMs encompass GPT-2 [19], GPT-3 [3], and the more recent GPT-4 [17], along with Llama-1 [24], Llama-2 [25], Falcons [18], etc. These LLMs possess some degree of multilingual capabilities, allowing them to comprehend and generate text in multiple languages. However, their primary linguistic focus remains on English. This linguistic bias is attributed to the fact that these models were predominantly trained on English-language data. Hence, researchers have focused on enhancing the capabilities of Large Language Models (LLMs) for specific languages or domains [21, 2]. Within the realm of the Japanese language, two notable models are

OpenCALM [5] and LLM-jp [10]. OpenCALM, functioning solely as a decoder model, is built upon the GPT-Neox model [1] and further fine-tuned using Japanese datasets. Its training encompasses the Japanese subset of Wikipedia [6] and Common Crawl[3] datasets, resulting in two versions, 7B and 13B. Similarly, LLM-jp is also a decoder-only LLM, utilizing the GPT-2 architecture [19]. It employs a new tokenizer with a mixed vocabulary of Japanese, English, and source code. In contrast to OpenCALM, LLM-jp is trained from scratch, incorporating datasets from Japanese, English, and code sources such as Wikipedia [6], mC4 [20], The Pile [7], and The Stack [12]. Subsequently, the pre-trained models undergo instruction fine-tuning based on human instructions derived from Japanese instructional datasets [10].

3 Methods

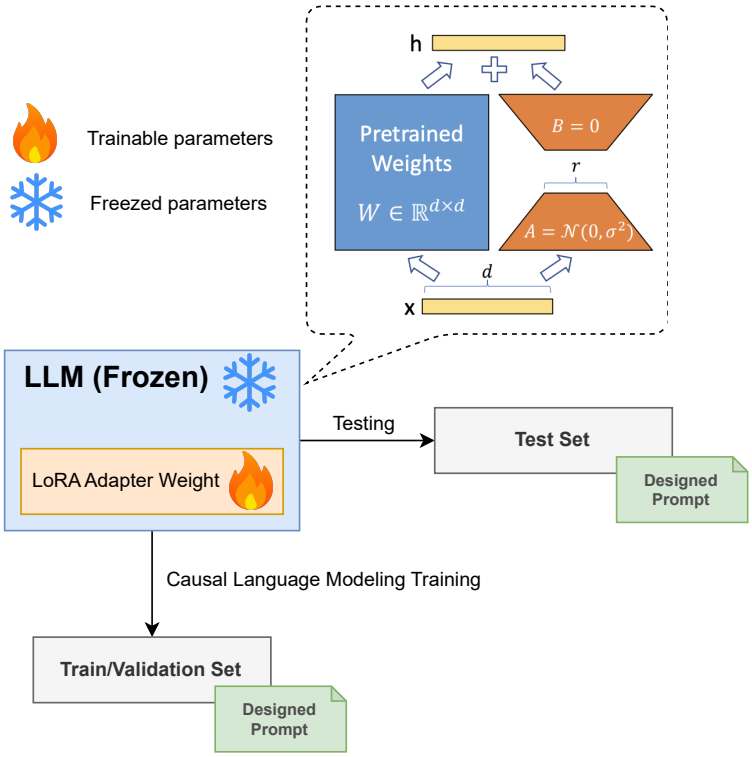

Fig. 1. Overview of the training and testing processes

The experiment methods aim to (1) explore the impact of different model sizes on the performance of various tasks and (2) compare Japanese-focused LLMs against Japanese-supported foundation LLMs on Japanese tasks. For all of our experiments, we set the maximum usage of VRAM in GPU to a fixed number, simulating the resource limitation constraint faced by individual companies. Figure 1 shows an overview of the training and testing processes in the experiments. Based on the purpose of each experiment, we use LLMs of different sizes (from 7 billion to 13 billion) and types (Japanese-focused and Japanese-supported LLMs). Due to the resource limitation constraint, we employ LoRA [9] (Low-Rank Adaptation), a popular Parameter-Efficient Fine-Tuning (PEFT) method, in the training process, where the pre-trained weights of the LLMs are frozen, and only the weights of the LoRA adapters are trainable. In the training process, we apply a pre-defined prompt template to each example in the training/validation set and use these examples to train the model in a causal language modeling fashion. A similar prompt template is used for each example

[3] https://commoncrawl.org/

in the testing phase, allowing the model to predict the answer generatively. Table 1 shows the prompts used in the method for each task.

Table 1. Prompts used in the method

Task	Prompt
Text classification	### Instruction: {text} ### Response: {response}
Machine reading comprehension	### Instruction: Context: {context} Question: {question} ### Response: {response}
Text summarization	以下のニュース記事をお読みください。 “‘見出し：{title} {text}“‘ このニュース記事を要約してください。 要約は以下の通りです： “‘{summarization}“‘

We choose LLMs of different sizes (from 7 billion to 13 billion) and types for the experiments. We aim to explore and compare the performance of foundation LLMs that support Japanese and LLMs that are specifically adapted to Japanese. Table 2 shows the models in the method with their corresponding number of pre-trained tokens and the percentages of English and Japanese in the pertaining data. Firstly, we choose Llama-2 (7B and 13B) [25] to represent the line of English-focused Japanese-supported foundation LLMs because its competitiveness has already been proven through various English and multilingual benchmarks. As we aim to compare models in different sizes, we choose the 7B version as the base size and the 13B version as the larger size. Secondly, we chose CALM2 and LLM-jp to represent the LLMs focused on the Japanese language. As CALM2 is only available in 7B and LLM-jp is only available in 13B, we choose to experiment on both models to compare the performance with Llama-2 7B and Llama-2 13B.

Table 2. Models used in the method.

Model	#Parameters	#Pre-trained tokens	% English in pre-train	% Japanese in pre-train
Llama-2	7 billion	2.0 trillion	89.70	0.10
Llama-2	13 billion	2.0 trillion	89.70	0.10
CALM2	7 billion	1.3 trillion	Not specified	Not specified
LLM-jp	13 billion	280 billion	48.69	47.83

4 Experiments

4.1 Base settings

We employed the huggingface checkpoints of the four selected LLMs: `cyberagent/calm2-7b`, `meta/Llama-2-7b-hf`, `meta/Llama-2-13b-hf`, and `llm-jp/llm-jp-13b-v1.0` to perform PEFT fine-tuning (LoRA method, particularly). We fine-tuned all models with maximum of 10 epochs with 4 different learning rates: 2×10^{-4}, 5×10^{-5}, 2×10^{-5}, and 1×10^{-5}. The maximum VRAM of GPU to be used is set to 46,000 MB.

4.2 Text classification

Dataset To benchmark the performance of PEFT-finetuned LLMs on the text classification task, we use the binary classification MARC-ja dataset, which is part of the JGLUE benchmark [13]. MARC-ja is the Japanese subset of the Multilingual Amazon Reviews Corpus, consisting of Amazon product reviews with 5-star ratings. To make it easier to interpret the meaning of the ratings, the creators of the dataset converted the star ratings into binary labels: 1/2-star is converted to "negative," 4/5-star is converted to "positive," and all the reviews with 3-star are removed. The ratings are then further refined through crowdsourcing to obtain the final dataset.

The MARC-ja dataset is divided into two splits: a training set, which contains roughly 188k rows, and a validation set, which contains approximately 5.65k rows. Each row consists of a review text and a binary label. Figure 2 depicts the length distribution of the review texts. We note that this dataset has unbalanced label distributions: the positive/negative ratio on the training set is close to 9/1 (88.2/11.8), and on the validation set is 8.5/1.5 (85.5/14.5).

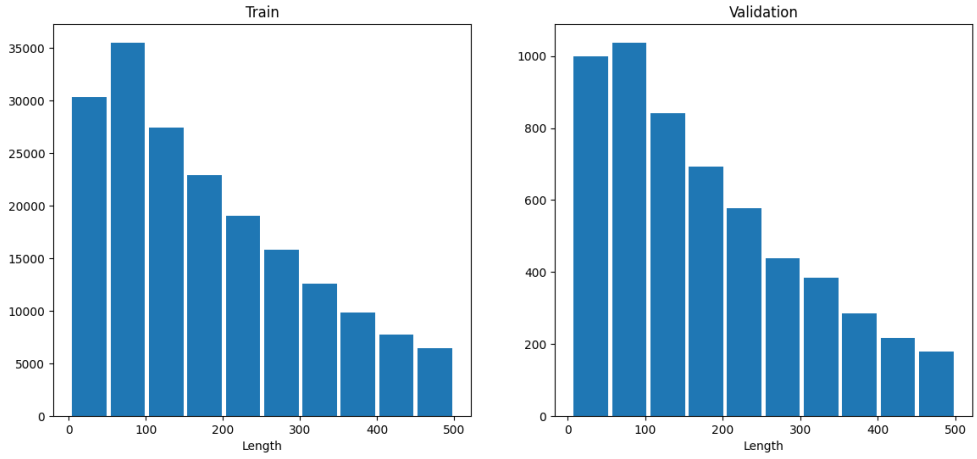

Fig. 2. MARC-ja dataset: sample length distribution

Experiment settings Four selected LLMs are fine-tuned under the base settings on the MARC-ja training dataset. For inference, we take the first word generated by LLMs after the input prompt as the predicted label. As the test dataset is unavailable, we measure the accuracy of the models on the validation set at the end of each epoch and report the highest accuracy of each model.

Table 3. Model performance on MARC-ja

Model	`lr=2e-4`	`lr=5e-5`	`lr=2e-5`	`lr=1e-5`
CALM2 7B	96.84	96.93	96.75	96.84
Llama-2 7B	96.58	96.45	96.27	96.19
Llama-2 13B	96.39	96.79	96.52	96.26
LLM-jp 13B	96.79	96.89	97.09	**97.11**

Experimental results We list the accuracy of each model on different learning rate settings in Table 3. Overall, all four LLMs perform equally well on the binary classification task presented on MARC-ja, with the LLM-jp 13B achieving the highest accuracy. The accuracy measure is stable among different learning rate settings, and all models reach a high accuracy score on the validation set. The trend in the loss value and accuracy metric is consistent between models, with the highest accuracy obtained after the second epoch in most cases with the best accuracy achieving epoch corresponding to the epoch with the lowest validation loss. Throughout the training process, the accuracy metric value fluctuates between 95-97% in all models in all training configurations. The two Japanese-focused LLMs, CALM2 7B, and LLM-jp 13B, only show minor improvement in accuracy over their Llama-2 counterpart despite being pre-trained on significantly more Japanese text. This could be due to the simplicity characteristic of the binary classification task and the fact that the Llama-2 models are pre-trained on a massive text dataset which includes Japanese.

4.3 Machine reading comprehension

Dataset JSQuAD [13] is an extractive machine reading comprehension dataset that is built on Japanese text. The dataset contains approximately 67,000 samples, each containing a question-answer pair and the corresponding reading passages from Wikipedia. The JSQuAD dataset is divided into two sets: the training dataset contains about 62,000 question-answer pairs based on more than 15,000 reading passages, while the development set contains about 4,000 question-answer pairs over approximately 1,000 reading passages. Table 4 describes the detailed information about both training and development sets of the JSQuAD. It can be seen that the average length of the answers is shorter than the length of the questions on both sets. Specifically, most of the length of answers are concentrated in about five words, fewer than the question with 25 words, and the reading passage (i.e., the context) with 169 words. Besides, as illustrated in Figure 3, the distribution of reading passages, questions, and answers on the JSQuAD is similar in both training and development sets.

Table 4. Overall statistics on the length of texts in the JSQuAD dataset

	Context	**Question**	**Answer**
TRAIN: 62,859			
# sample	15,651	62,859	62,859
max length	858,00	99,00	19,00
average length	190.19	28.07	6.11
median length	175.0	24.0	5.0
DEV: 4,442			
# sample	1,145	4,442	4,442
max length	906	98	19
average length	183.29	28.77	6.17
median length	169.0	25.0	5.0

Experiment settings We choose the four LLMs for the experiments as shown in Table 2 with the PEFT fine-tuning on the JSQuAD training set and use the development set to record the final performance results of the best checkpoint (as the test set is not available). To evaluate the impact of learning rate on the performance of LLMs on fine-tuning with LoRA, we use four different learning rates, including 2×10^{-4}, 5×10^{-5}, 2×10^{-5}, and 1×10^{-5}.

In the inference stage, we set the maximum generation of the response prompt (as described in Table 1) to 5 because most of the answers in the dataset fall into 5 words (as mentioned in Table 4). Besides, to reduce the redundant of the generated response, we integrate a post-processing method to remove the unexpected generated token such as *<EOD|LLM-jp>* and Japanese auxiliary words including は, が, と, に, で, を, へ, も, の, から, the comma (、) and the dot (。).

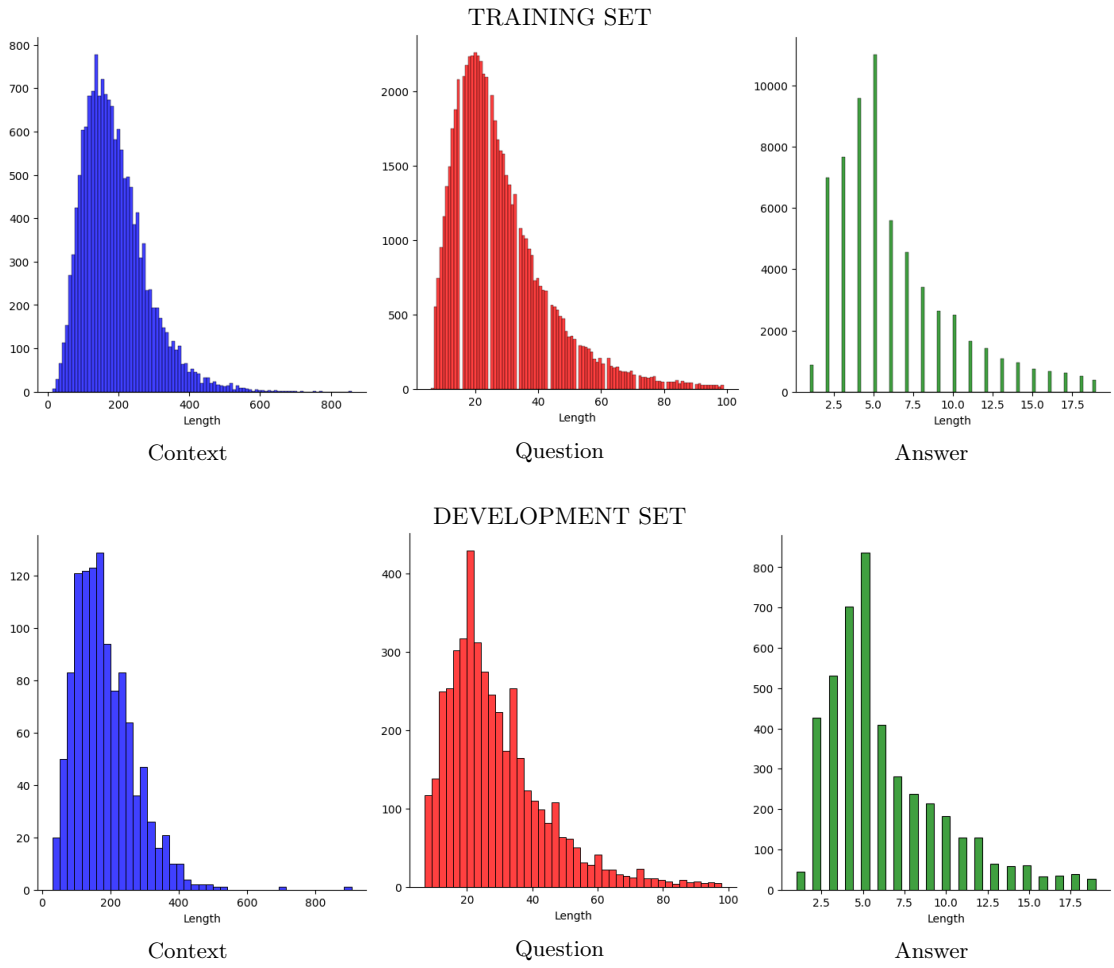

Fig. 3. Distribution of text length in the JSQuAD dataset

Table 5. Model performance on JSQuAD with results in EM/F1 (%)

Model	`lr=2e-4`	`lr=5e-5`	`lr=2e-5`	`lr=1e-5`
CALM2 7B	47.16/81.45	47.32/81.38	45.94/80.38	38.92/76.46
Llama-2 7B	56.19/85.24	56.93/85.65	58.68/86.46	58.32/85.57
Llama-2 13B	59.23/86.33	58.73/86.46	56.70/85.77	60.15/86.58
LLM-jp 13B	**87.55/93.89**	**87.55/93.89**	**87.55/93.89**	**87.55/93.89**

Experimental results Table 5 shows the empirical results of the four LLMs on the JSQuAD. We can see that the size of the model has significant effects on the performance. For Llama-2, the performance of the model with 13 billion parameters (Llama-2 13B) is better than the one with 7 billion parameters (Llama-2 7B) on both EM and F1 metrics. For this task, the LLM-jp 13B, which is a Japanese-focused LLM, shows the best performance on both EM and F1. However, it is observed that the Llama-2 7B model, which is an English-focused Japanese-supported LLM, performs better than a Japanese-focused LLM of the same size, which is the CALM2 7B model. It suggests that in practice, we should also consider English-focused Japanese-supported LLMs when adapting a new task, leveraging the effect of knowledge sharing between languages in multilingual LLMs. To summary, the experimental results indicates that in this task, the larger-

sized model generally has better performance than the smaller-sized model; and the adaptation capacity of English-focused Japanese-supported LLMs is considerable.

4.4 Text summarization

Dataset XL-Sum [8] is an extensive and varied collection of 1.35 million article-summary pairs from BBC, professionally annotated and extracted through meticulous heuristics. Encompassing 45 languages with varying resource levels, XL-Sum stands out for its highly abstractive, concise, and high-quality nature, as confirmed by both human and intrinsic evaluations. In our research, we selected the XL-Sum dataset, focusing on its Japanese text subset, for the Japanese text summarization task. This XL-Sum Japanese dataset contains 8,891 samples, divided into three splits: 7,113 samples for training, 889 samples for validation, and 889 samples for testing (train:validation:test = 8:1:1).

Each sample includes a title, a text, and a summary where the title and the text (content) are taken from the BBC Japanese [4], and the summary is annotated. Figure 4 demonstrates the length distribution of the test set. The data's length distribution in the training, validation, and test sets are similar. We can see that while the title does not exceed 50 words, the text (the content of the news) is often very long: it could be 5,000 words in length, but most of them are 3,000 words or less. For the summary, the max length of a gold summary is around 200 Japanese words.

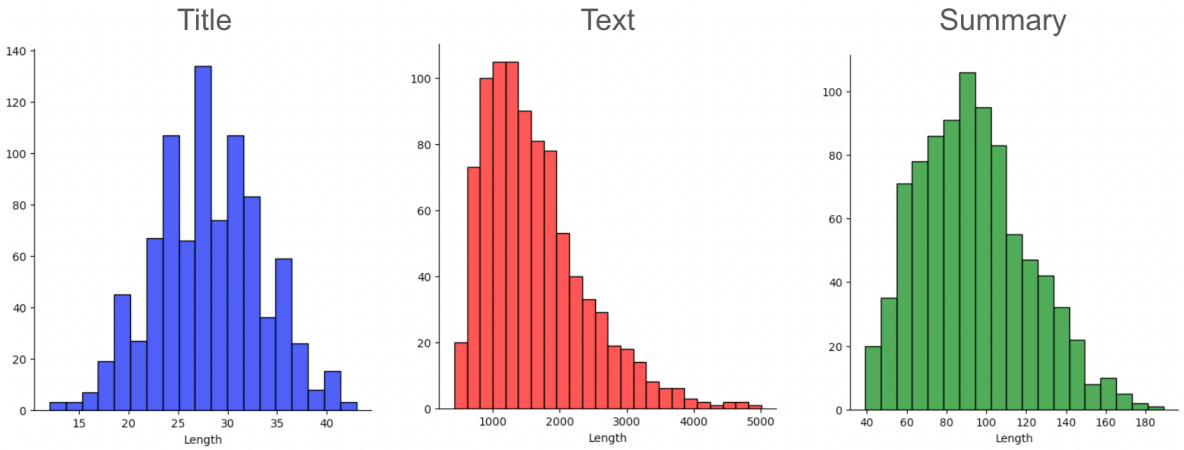

Fig. 4. Data length distribution of test set

Experimental settings We investigate the effectiveness of summarizing Japanese text sourced from the XL-Sum dataset using various Transformer-based models, namely CALM-2 7B, Llama-2 7B, Llama-2 13B, and LLM-jp 13B. The ideal case is to fine-tune these models with an optimal maximum sequence length that accommodates all data sample lengths. However, due to the limited VRAM size of GPU that we set, we are constrained to fine-tune with a maximum sequence length of 1,700 tokens for 7B models and 1,024 tokens for 13B models. In the case of 7B models, we also experiment with fine-tuning with the max sequence length of 1,024 tokens. In the inference phase, as a common practice, we set the length of input tokens to the maximum value possible, that is 4,096 for the 7B models and 1,500 for the 13B models.

Four different learning rates are considered to fine-tune the models: 2×10^{-4}, 5×10^{-5}, 2×10^{-5}, and 1×10^{-5}. The ROUGE-1 score, which measures the overlap of unigrams between the generated summary and reference summary, is employed as our evaluation metric to determine the summarization quality. We have omitted the inclusion of ROUGE-2 and ROUGE-L scores in our evaluation due to constraints on the length of our report. Nevertheless, it should be noted that the trends displayed by these metrics are consistent with those observed in the ROUGE-1 score.

[4] https://www.bbc.com/japanese

Table 6. Model Training Results

Model	train_max_len	infer_input_max_len	lr=2e-4	lr=5e-5	lr=2e-5	lr=1e-5
CALM2 7B	1,700	4,096	**0.3730**	0.3578	0.3572	0.3518
Llama-2 7B	1,700	4,096	0.3390	0.3150	0.3027	0.3015
CALM2 7B	1,024	4,096	**0.3679**	0.3649	0.3566	0.3541
Llama-2 7B	1,024	4,096	0.3095	0.2909	0.2732	0.2512
Llama-2 13B	1,024	1,500	0.2968	0.2966	0.2705	0.2387
LLM-jp 13B	1,024	1,500	0.2244	0.2480	0.2887	**0.3389**

Experimental results Based on the experimental results presented in Table 6, we can draw several insights and conclusions regarding the performance of Transformer-based models in summarizing Japanese text from the XL-Sum dataset.

Firstly, it is clear that CALM-2 7B outperforms the Llama-2 7B model across all learning rates when trained with a maximum sequence length of 1700 tokens. This suggests that the Japanese-focused foundation LLM (CALM-2 7B) is better suited for tasks involving Japanese text than the Japanese-supported LLM (Llama-2 7B). The best performance for CALM-2 7B with a training length of 1700 tokens was achieved with a learning rate of 2×10^{-4}, yielding a ROUGE-1 score of 0.3730, which is notably higher than the best score for Llama-2 7B at the same settings.

Secondly, when we consider the models with a reduced training maximum sequence length of 1024 tokens, the CALM-2 7B model again demonstrates superior performance compared to the Llama-2 7B at all learning rates. However, the performance of CALM-2 7B drops slightly with the smaller sequence length from 0.3730 to 0.3679, which is expected due to the limitation imposed on the model's input size.

When examining the larger 13B models, we see a different trend. The LLM-jp 13B achieved better performance as the learning rate decreased, with the best ROUGE-1 score of 0.3389 at the lowest learning rate of 1×10^{-5}. This improvement at lower learning rates suggests that the model may benefit from more nuanced weight adjustments during fine-tuning, potentially due to its larger capacity and pre-training on Japanese data. In contrast, the Llama-2 13B model peaking at a learning rate of 2×10^{-4} then declined as the learning rate was reduced further, indicating that this model might not require or benefit from the same level of learning rate fine-tuning.

Interestingly, the smaller 7B models, particularly CALM-2 7B, outperform their larger 13B counterparts under most conditions. This could be due to a variety of factors, such as the effectiveness of the models' pre-training on Japanese text (the CALM-2 model is pre-trained on 1.3 trillion tokens while the LLM-jp model is pre-trained on only 280 billion tokens), differences in fine-tuning sequence lengths or the capacity of the models to generalize from the training data.

In summary, the Japanese-focused foundation LLMs (CALM-2 7B and LLM-jp 13B) performed better overall in Japanese text summarization tasks compared to the Japanese-supported foundation LLMs (Llama-2 7B and Llama-2 13B). Additionally, the choice of learning rate is crucial for optimizing performance.

5 Discussion

Performance Gap Between Large and Small Models. In our experimental results, it is observed that larger language models tend to outperform their smaller counterparts in various natural language processing (NLP) tasks. However, the extent of this performance gap varies depending on the complexity and demands of the task at hand. For relatively straightforward tasks like text classification, the performance improvement offered by larger models is typically marginal, often amounting to only a few decimal points (Table 3). In contrast, for more complex and demanding tasks such as machine reading comprehension, the performance gap between large and small models can be substantial, with larger models demonstrating a significant advantage, sometimes improving the metric by several percentage points (Table 5). These observations align with the findings of previous research in this domain [4, 11].

Competitive Performance of English-Focused Japanese-Supported LLMs in Text Classification. For less demanding NLP tasks like text classification, English-focused language models that support the Japanese language can exhibit competitive performance compared to Japanese-focused models. In our experiments, these English-focused Japanese-supported models achieved impressive accuracy scores ranging from 96% to 97% on Japanese text classification tasks (Table 3). We attribute this competitive performance to the fact that for relatively simple tasks, even smaller English-focused models may possess sufficient capacity to effectively solve the problem, negating the need for specialized, larger Japanese-focused models.

Superior Performance of Japanese-Focused Models in Complex NLP Tasks. However, when it comes to more complex and demanding NLP tasks, such as machine reading comprehension or text summarization, Japanese-focused models that are specifically designed and optimized for the Japanese language demonstrate superior performance compared to their English-focused Japanese-supported counterparts (Table 5, 6). The reason for this performance advantage lies in the fact that Japanese-focused models are meticulously tailored to handle the unique intricacies of the Japanese language, including its grammar, writing system, and cultural nuances. These nuances are often challenging for generic, English-focused models to capture accurately, even when they support the Japanese language to some extent. Consequently, Japanese-focused models, with their specialized architecture and training, are better equipped to tackle the complexities of advanced Japanese NLP tasks, resulting in higher performance metrics.

The application of LLMs in Business Innovation. According to [27], the LLMs can be involved in Business Process Management (BPM) projects and can serve various tasks in BPMs with corresponding datasets. From the experiment in Section 4, it can be seen that the ability of LLMs outperforms other SOTA methods when it is fine-tuned for a specific task including text classification, machine reading comprehension, and text summarization. These tasks serve the BPM lifecycle, and the power of LLMs helps the BPM can be used in commercial products with a huge amount of users [27]. Instead, the usage of LLMs in Business also deals with several risks. One of the critical risks in LLMs usage is data privacy, which concerns the leakage of data when hosting LLMs on multi-party platforms and potential attack methods to the privacy of LLMs [15].

6 Conclusion

In conclusion, this paper has investigated the applicability and performance nuances of foundation Large Language Models (LLMs) across three diverse and complex tasks tailored to the Japanese language: text classification, machine reading comprehension, and text summarization. Through a methodical comparative analysis, we have provided empirical evidence on the relative influences of model size and language specialization within the constraints faced by organizations with limited computational resources, particularly micro-enterprises.

Our findings demonstrate a multifaceted relationship between model parameters and task performance. We have established that generally, models with a larger parameter count exhibit enhanced performance on benchmark tasks. However, this improvement is not uniform across tasks and must be considered in conjunction with language specialization. When comparing the performance of Japanese-focused Language Models (LLMs) with English-focused Japanese-supported LLMs, it is evident that the primary training language of an LLM has a considerable influence on its effectiveness in managing tasks within the Japanese language domain.

Specifically, while English-focused Japanese-supported models can exhibit commendable performance, our study suggests that Japanese-focused LLMs tend to perform better on more complex tasks. This finding stands as a pivotal insight for enterprises seeking to employ these models effectively, highlighting the trade-offs and decision points in selecting the appropriate model type for their specific NLP needs.

Furthermore, this paper has addressed the practical challenges associated with fine-tuning and optimizing the inference process in computationally constrained environments. We have offered guidance on how such entities can strategically assess and deploy foundation LLMs to achieve their objectives, enhancing their linguistic processing capabilities without overextending their limited resources.

In essence, our investigation fulfills two main objectives: it enhances the comprehension of LLM robustness when processing the Japanese language and it enables practitioners and organizations to make knowledgeable choices regarding the implementation of LLMs. It facilitates further research to investigate the complexities of foundational LLMs across a wider array of languages and tasks, to improve the guidance for NLP applications in environments with limited resources.

References

1. Andonian, A., Anthony, Q., Biderman, S., Black, S., Gali, P., Gao, L., Hallahan, E., Levy-Kramer, J., Leahy, C., Nestler, L., Parker, K., Pieler, M., Purohit, S., Songz, T., Phil, W., Weinbach, S.: GPT-NeoX: Large Scale Autoregressive Language Modeling in PyTorch (8 2021). https://doi.org/10.5281/zenodo.5879544, https://www.github.com/eleutherai/gpt-neox
2. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)
3. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
4. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., et al.: Palm: Scaling language modeling with pathways. Journal of Machine Learning Research **24**(240), 1–113 (2023)
5. cyberagent: Open-calm (2023), https://huggingface.co/cyberagent/open-calm-7b
6. Foundation, W.: Wikimedia downloads, https://dumps.wikimedia.org
7. Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., Presser, S., Leahy, C.: The Pile: An 800gb dataset of diverse text for language modeling. arXiv preprint arXiv:2101.00027 (2020)
8. Hasan, T., Bhattacharjee, A., Islam, M.S., Mubasshir, K., Li, Y.F., Kang, Y.B., Rahman, M.S., Shahriyar, R.: XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In: Zong, C., Xia, F., Li, W., Navigli, R. (eds.) Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. pp. 4693–4703. Association for Computational Linguistics, Online (Aug 2021). https://doi.org/10.18653/v1/2021.findings-acl.413, https://aclanthology.org/2021.findings-acl.413
9. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=nZeVKeeFYf9
10. llm jp: Llm-jp (2023), https://huggingface.co/llm-jp/llm-jp-13b-v1.0
11. Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., Amodei, D.: Scaling laws for neural language models. arXiv preprint arXiv:2001.08361 (2020)
12. Kocetkov, D., Li, R., Ben Allal, L., Li, J., Mou, C., Muñoz Ferrandis, C., Jernite, Y., Mitchell, M., Hughes, S., Wolf, T., Bahdanau, D., von Werra, L., de Vries, H.: The stack: 3 tb of permissively licensed source code. Preprint (2022)
13. Kurihara, K., Kawahara, D., Shibata, T.: JGLUE: Japanese general language understanding evaluation. In: Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., Piperidis, S. (eds.) Proceedings of the Thirteenth Language Resources and Evaluation Conference. pp. 2957–2966. European Language Resources Association, Marseille, France (Jun 2022), https://aclanthology.org/2022.lrec-1.317
14. Ladhak, F., Durmus, E., Cardie, C., McKeown, K.: WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In: Cohn, T., He, Y., Liu, Y. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2020. pp. 4034–4048. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.findings-emnlp.360, https://aclanthology.org/2020.findings-emnlp.360
15. Li, H., Chen, Y., Luo, J., Kang, Y., Zhang, X., Hu, Q., Chan, C., Song, Y.: Privacy in large language models: Attacks, defenses and future directions. arXiv preprint arXiv:2310.10383 (2023)
16. Nakayama, Y., Murakami, K., Kumar, G., Bhingardive, S., Hardaway, I.: A large-scale Japanese dataset for aspect-based sentiment analysis. In: Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., Piperidis, S. (eds.) Proceedings of the Thirteenth Language Resources and Evaluation Conference. pp. 7014–7021. European Language Resources Association, Marseille, France (Jun 2022), https://aclanthology.org/2022.lrec-1.758
17. OpenAI: Gpt-4 technical report (2023)
18. Penedo, G., Malartic, Q., Hesslow, D., Cojocaru, R., Cappelli, A., Alobeidli, H., Pannier, B., Almazrouei, E., Launay, J.: The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. arXiv preprint arXiv:2306.01116 (2023)

19. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
20. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv e-prints (2019)
21. Roziere, B., Gehring, J., Gloeckle, F., Sootla, S., Gat, I., Tan, X.E., Adi, Y., Liu, J., Remez, T., Rapin, J., et al.: Code llama: Open foundation models for code. arXiv preprint arXiv:2308.12950 (2023)
22. So, B., Byun, K., Kang, K., Cho, S.: Jaquad: Japanese question answering dataset for machine reading comprehension. arXiv preprint arXiv:2202.01764 (2022)
23. Suzuki, H., Miyauchi, Y., Akiyama, K., Kajiwara, T., Ninomiya, T., Takemura, N., Nakashima, Y., Nagahara, H.: A Japanese dataset for subjective and objective sentiment polarity classification in micro blog domain. In: Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., Piperidis, S. (eds.) Proceedings of the Thirteenth Language Resources and Evaluation Conference. pp. 7022–7028. European Language Resources Association, Marseille, France (Jun 2022), https://aclanthology.org/2022.lrec-1.759
24. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
25. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023), https://arxiv.org/abs/2307.09288
26. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
27. Vidgof, M., Bachhofner, S., Mendling, J.: Large language models for business process management: Opportunities and challenges. In: International Conference on Business Process Management. pp. 107–123. Springer (2023)
28. Yang, J., Jin, H., Tang, R., Han, X., Feng, Q., Jiang, H., Yin, B., Hu, X.: Harnessing the power of llms in practice: A survey on chatgpt and beyond. arXiv preprint arXiv:2304.13712 (2023)
29. Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al.: A survey of large language models. arXiv preprint arXiv:2303.18223 (2023)