Leveraging High-Resource English Corpora for Cross-lingual Domain Adaptation in Low-Resource Japanese Medicine via Continued Pre-training

Anonymous ACL submission

Abstract

Limited low-resource language corpora in professional domains like medicine hinder cross-lingual domain adaptation of pre-trained large language models (PLMs). While abundant English medical corpora could complement this scarcity, the effective mixture of English and target language, including machine-translated content, remains underexplored. We examined how corpus compositional statistics (e.g., token sizes and language proportions) affect performance on a Japanese–English medical knowledge benchmark. Through continued pre-training of a bilingual PLM on multilingual corpora with varying proportions of English and Japanese texts (both original and machine-translated), we analyzed correlations between corpus compositional statistics and fine-grained task performance. Our findings suggest a practical approach to optimizing multilingual corpora for cross-lingual domain adaptation, which requires leveraging specialized knowledge from English corpora while ensuring sufficient coverage of language-specific expressions in a target language (Japanese). Such insights will contribute to the development of multilingual models that effectively leverage English-language resources in various professional domains with low-resource languages.

1 Introduction

Imbalanced language resources pose a significant challenge for pre-trained large language models (PLMs) in achieving cross-lingual domain adaptation in specific target languages. This imbalance is especially pronounced in professional domains such as medicine, where general biomedical knowledge circulates globally in English, while available resources in the target language remain relatively limited. For example, PubMed hosts over 38 million biomedical papers globally[1], while J-STAGE, a comparable Japanese database, contains only around 5 million[2]. While abundant English medical corpora offer a promising avenue for augmenting scarce target-language data, the optimal continued pre-training strategy for acquiring knowledge from well-resourced source languages (often English) to support domain adaptation in less-resourced target languages has yet to be thoroughly explored.

Here, we investigate the optimal corpus composition for the *continued pre-training* of a bilingual (Japanese–English) PLM, with a particular focus on leveraging abundant English-language resources to enhance knowledge acquisition in Japanese medicine. Continued pre-training usually follows initial pre-training on general corpora, where large language models acquire foundational language abilities such as lexical, syntactic, and semantic patterns, as well as general factual knowledge (Petroni et al., 2019; AlKhamissi et al., 2022). Then, continued pre-training leverages additional corpora containing domain-specific or target-language texts, with its effectiveness for domain adaptation demonstrated across multiple studies (Gupta et al., 2023; Cui et al., 2024; Pires et al., 2023; Zhu et al., 2023; Zhao et al., 2024a; Fujii et al., 2024).

Nevertheless, several practical considerations have been overlooked for effective continued pre-training aimed at cross-lingual domain adaptation in low-resource professional domains. For example, the optimal mixing ratio of source and target languages for acquiring knowledge from English corpora remains unclear. While current machine translation systems provide reasonable quality, the balance between original and translated content is still not well understood. Furthermore, existing studies often lack detailed analyses of how *corpus compositional statistics* (e.g., token sizes and

[1] Statistics of PubMed: https://pubmed.ncbi.nlm.nih.gov/about/

[2] Statistics of J-STAGE: https://www.jstage.jst.go.jp/browse/-char/en

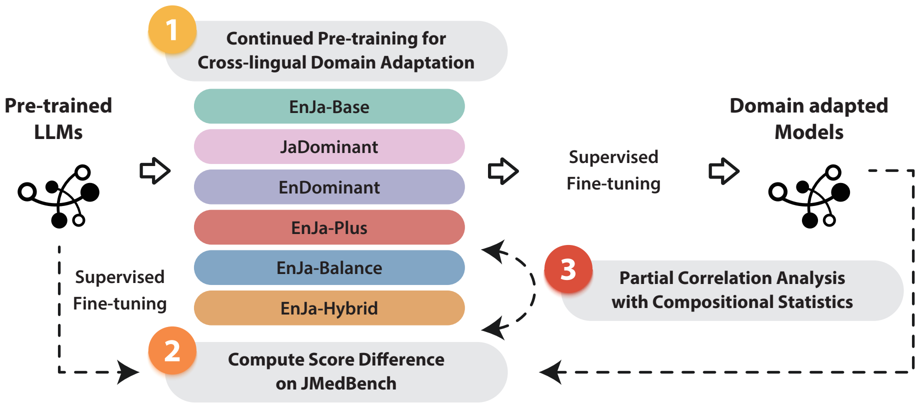

Fig. 1: Study Overview. This study comprises three steps. (1) First, we performed continued pre-training on pre-trained large language models using diverse multilingual corpora. (2) Next, we computed the difference in scores before and after the continued pre-training using the Japanese–English medical knowledge benchmark, JMedBench. (3) Finally, we conducted partial correlation analysis to identify task-wise language preferences, thereby revealing the optimal corpus composition for cross-lingual domain adaptation.

language proportions) representing corpus composition affect downstream performance across tasks and languages.

In this study, we address the following research questions (RQs):

RQ1: How do original English and machine-translated Japanese corpora help a bilingual (Japanese–English) PLM achieve domain adaptation in the Japanese medical domain?

RQ2: What is the optimal corpus configuration and proportion of English and Japanese texts for achieving the best performance in the medical domain?

RQ3: How do specific corpus compsitional statistics in multilingual corpora influence model performance across diverse medical tasks?

To investigate these questions, we systematically compared the impact of multilingual corpora containing varying proportions of Japanese and English medical content (see the study overview in **Fig. 1**). To characterize the language composition of these corpora, we defined seven *compositional statistics*: total token count, Japanese token count, English token count, parallel token count (paragraph-aligned bilingual medical texts), and the ratios of Japanese, English, and parallel tokens. Then, we employed 13-billion-parameter bilingual (Japanese–English) PLMs and computed the difference in model performance on a comprehensive Japanese–English medical knowledge benchmark, *JMedBench* (Jiang et al., 2025), before and after continued pre-training. JMedBench comprises 20 Japanese and 7 English tasks, including multiple-choice question answering (MCQA), machine translation (MT), named entity recognition (NER), document classification (DC), and semantic textual similarity (STS) (see **Appendix A**). Finally, we applied *partial correlation analysis*, which estimates the strength and direction of a relationship between two variables while controlling for other covariates. This enabled us to isolate the unique contribution of each compositional statistic to task performance despite inherent mutual correlations—for example, more Japanese tokens automatically raise the total token count. Our findings underscore the need to optimize corpus composition so that high-resource English texts can be leveraged effectively for cross-lingual domain adaptation in the low-resource Japanese medical domain.

Our contributions, which correspond to the RQs, can be summarized as follows:

- We systematically evaluate multilingual corpora featuring varying proportions of Japanese and English medical texts, identifying the potential benefits of both original English and machine-translated Japanese texts.
- We demonstrate that a well-balanced multilingual corpus can enhance knowledge acquisition in both Japanese and English medical domains, achieving the best performance on JMedBench.
- Our partial correlation analysis quantifies how specific compositional statistics in multilingual corpora influence task-specific performance across various medical tasks, providing insights into the optimal configuration of the corpus for cross-lingual domain adaptation.

2 Related Work

Cross-lingual Domain Adaptation. Techniques aimed at enhancing multilingual language models' understanding of low-resource languages have attracted considerable attention (Xu et al., 2024), leading to broadly recognized concepts such as cross-lingual alignment and cross-lingual transfer (Hämmerl et al., 2024). Typically, assuming the presence of high-resource (source) and low-resource (target) languages, the objectives of these approaches fall into two main categories: (1) promoting knowledge transfer from source to target languages (Castellucci et al., 2021; Rathore et al., 2023; Tanwar et al., 2023; Awasthi et al., 2023; Singh et al., 2024; Zhang et al., 2024; Yong et al., 2023); and (2) acquiring new domain-specific knowledge within the target language (Zhao et al., 2024a; Wan et al., 2024; Fujii et al., 2024). Furthermore, these approaches can be classified based on whether cross-lingual representations require explicit alignment within embedding spaces (Zhao et al., 2024b). In this study, we define *cross-lingual domain adaptation* as an approach that specifically facilitates knowledge acquisition from a high-resource English medical corpus to complement a low-resource Japanese corpus, without explicitly aligning cross-lingual embedding spaces.

Techniques for the Cross-lingual Domain Adaptation. Algorithms for the cross-lingual domain adaptation can be categorized along two dimensions: (1) the training stage at which the method is applied, and (2) the types of signals used for alignment.

Multilingual pre-training has been explored (Chi et al., 2021); however, effectively capturing nuanced semantics and specialized terminology, particularly in low-resource languages, remains challenging (Wu et al., 2022). Continued pre-training, typically performed after initial pre-training, leverages additional corpora containing domain-specific or target-language texts. While its effectiveness has been demonstrated in various studies (Gupta et al., 2023; Cui et al., 2024; Pires et al., 2023; Zhu et al., 2023; Zhao et al., 2024a; Fujii et al., 2024), detailed analyses of how the language composition of corpus influences specific task performance—particularly from the perspective of leveraging high-resourced language corpus—are still lacking. Additionally, supervised fine-tuning performed after (continued) pre-training plays a pivotal role in enhancing cross-lingual performance, especially when substantial instruction datasets in the target domain are available (Mecklenburg et al., 2024; Razumovskaia et al., 2024; Shaham et al., 2024).

There are several types of signals used for alignment. A *multilingual corpus*, as employed in this study, contains texts from both the source and target languages (Qin et al., 2025; ImaniGooghari et al., 2023; Shaham et al., 2024). A *parallel corpus* is a specialized type of multilingual corpus consisting of explicitly aligned sentences or paragraphs across the source and target languages. While parallel corpora have demonstrated clear positive effects on specific tasks such as machine translation (Chi et al., 2022; Hu et al., 2020; Feng et al., 2022; Yang et al., 2023; Lin et al., 2025), their effectiveness in a broader range of tasks, especially within professional domains, remains controversial. To address this, we conduct a detailed analysis of parallel corpora, examining their advantages and disadvantages specifically for Japanese–English medical domain adaptation. Other alignment signals include *transliteration*, which leverages the romanized forms of text to enhance alignment through shared tokens with English (Husain et al., 2024), and *code-switching*, which augments original data by explicitly introducing cross-lingual supervision (Yamada and Ri, 2024; Hong et al., 2025).

3 Method

This study comprises three steps (see **Fig. 1**): (1) continued pre-training of a bilingual (Japanese–English) PLM on diverse multilingual corpora with various language compositions; (2) computation of task-wise score differences on JMedBench before and after continued pre-training; and (3) partial correlation analysis to examine task-wise correlations with corpus compositional statistics.

3.1 Multilingual Corpora

As shown in **Fig. 2**, we constructed six medical corpora with varying Japanese–English compositions:

- **EnJa-Base**: Contains basic medical content from textbooks, clinical guidelines, paper abstracts, and web-crawled data in Japanese and English, as well as a certain amount of parallel corpus. The parallel subcorpus refers to text containing aligned English and Japanese sentences or paragraphs presented in randomized order.
- **JaDominant**: Adds a machine-translated

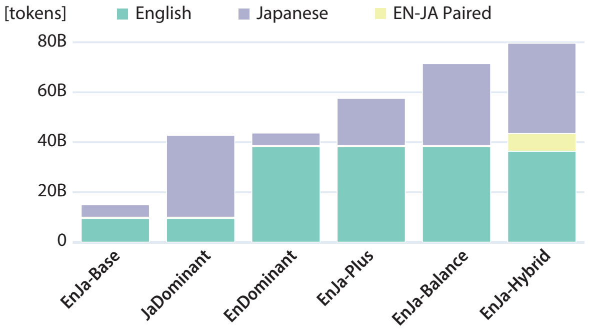

Fig. 2: Multilingual Corpora. Six multilingual medical corpora with varying Japanese–English compositions were constructed. Notably, the token distributions across the corpora show that the total number of tokens increases from EnJa-Base to EnJa-Hybrid.

Japanese version of the PubMed Central (PMC) full-text subcorpus[3] to EnJa-Base, resulting in a Japanese-dominant corpus. Refer to **Appendix B** regarding the accuracy of the machine translation used in this research.

- **EnDominant**: Adds the original English PMC subcorpus to EnJa-Base, resulting in an English-dominant corpus. Note that between JaDominant and EnDominant, the Japanese translation of the PMC full-text subcorpus is replaced with the original English.
- **EnJa-Plus**: Extends the EnDominant corpus by adding half of the translated PMC subcorpus.
- **EnJa-Balance**: Builds on EnDominant by adding full the size of the translated Japanese PMC subcorpus. Note that in EnDominant, EnJa-Plus, and EnJa-Balance, the English corpus remains constant, and these variants respectively contain none, half, or all of the Japanese translation of the PMC full-text subcorpus.
- **EnJa-Hybrid**: Further extends EnJa-Balance with additional medical textbooks and clinical guidelines. Besides, this contains a large amount of parallel corpus that was created by translating PubMed paper abstracts[4].

Notably, we defined seven compositional statistics to characterize each corpus. One group pertains to the number of tokens in each language, including *Japanese token count*, *English token count*, *parallel token count*, and *total token count*. Another group of statistics represents the proportion of each language within a corpus, including *Japanese token ratio*, *English token ratio*, and *parallel token ratio*.

[3]We used Commercial Use Allowed articles from the PMC Open Access Subset.

[4]https://pubmed.ncbi.nlm.nih.gov/download/

Multilingual Corpora	Japanese Tasks	English Tasks
EnJa-Base	0.447	0.429
JaDominant	0.453	0.455
EnDominant	0.468	0.467
EnJa-Plus	0.461	0.469
EnJa-Balance	**0.475**	**0.473**
EnJa-Hybrid	0.467	0.466

Table 1: Average Scores on JMedBench. The model trained with EnJa-Balance achieved the highest performance on both Japanese tasks (0.475 average score across all 20 tasks) and English tasks (0.473 average score across all 7 tasks), outperforming models trained with other corpus compositions.

Refer to **Appendix C** for detailed values on the compositional statistics of each corpus.

3.2 Continued Pre-training on the Multilingual Corpora

Using multilingual corpora, we performed continued pre-training on bilingual (Japanese–English) PLMs, namely `llm-jp/llm-jp-3-13b`[5] (LLM-jp et al., 2024) (see Step 1 in **Fig. 1**). Since JMedBench requires basic instruction-following capability, we applied instruction tuning to both models *before* and *after* continued pre-training. By computing the score difference between the two training states, we can evaluate the performance gain attributable to the specific pre-training corpus. See **Appendix D** for the detailed model architecture, training hyperparameters, and instruction tuning dataset.

3.3 Performance Evaluation on JMedBench

We evaluated model performance in both the Japanese and English medical domains using JMedBench (see Step 2 in **Fig. 1**), which comprises 27 tasks in total (20 in Japanese and 7 in English). We tested the models before and after the continued pre-training, resulting in 12 models overall (6 corpora $\times$ 2 training states). We then computed a *score difference* for each corpus, defined as: (performance after mid-training + SFT) $-$ (performance before mid-training + SFT), where SFT stands for supervised fine-tuning described in **Appendix D**.

Note that since the multilingual corpora are constructed in an additive or ablative manner (see **Section 3.1**), comparing score differences between models trained on them effectively constitutes an *additive* or *ablation* study. For instance, the com-

[5]https://huggingface.co/llm-jp/llm-jp-3-13b

parison between JaDominant and EnDominant offers insights into whether the PMC subcorpus should be translated into Japanese or used in its original English form when added individually. Differences among EnDominant, EnJa-Plus, and EnJa-Balance help clarify the optimal mixing ratio (none, half, or full) of translated data. Lastly, the contrast between EnJa-Balance and EnJa-Hybrid highlights the utility of enriched text sources, such as parallel corpora.

3.4 Partial Correlation Analysis

Since mutual correlations exist among compositional statistics and task-wise score differences (see Appendix E), we applied partial correlation analysis to isolate the unique impact of each variable. This approach allowed us to assess the direct association between a predictor (e.g., a compositional statistic) and an outcome variable (e.g., a task-wise score difference) while controlling for other covariates (see Step 3 in **Fig. 1**). First, we used ordinary least squares regression to regress the predictor on the covariates, extracting residuals to remove the covariates' linear effects. We then applied the same procedure to the outcome variable and computed the Pearson correlation between these two sets of residuals. This method yields the partial correlation coefficient r, indicating how strongly the predictor is related to the outcome when shared variance with the covariates is accounted for. The associated p-value tests the significance of this unique relationship. Hereinafter, significance levels are denoted as follows: *** for $p < 0.001$, ** for $p < 0.01$, and * for $p < 0.05$. A statistically significant correlation is considered "strong" when $p < 0.01$ in this study. We use abbreviations such as `Ja/MCQA` to indicate Japanese MCQA tasks.

4 Results

4.1 Model Performance on JMedBench

We evaluated the task performance of the continued pre-trained models on JMedBench. **Table 1** shows the average score across the 20 Japanese and 7 English tasks. Overall, three key observations emerge from these results, particularly from the aspect of the benefit of the machine-translation data.

First, even for Japanese tasks, using the original PMC subcorpus in English yielded a greater performance gain than the machine-translated one, as indicated by the average score of EnDominant (0.468) versus JaDominant (0.453). This suggests

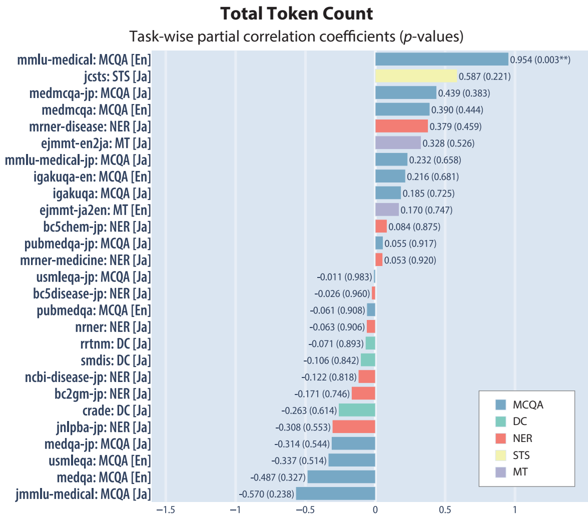

Fig. 3: Task-wise Correlation with Total Token Count. Partial correlation analysis showed the strongest positive correlation with MMLU-Medical. The x-axis shows partial correlation coefficients with p-values in parentheses.

that the machine-translated data might be of suboptimal quality, limiting its impact on model performance.

Second, despite the above limitation, there can be an additive effect from the translated data. By comparing EnDominant, EnJa-Plus, and EnJa-Balance, we see how adding none, half, or the full amount of the Japanese-translated PMC subcorpus affects performance. Notably, only incorporating the full amount of translation raises the average score from 0.468 (EnDominant) to 0.475 (EnJa-Balance). The same benefit can also be observed in English tasks (see EnJa-Balance in **Table 1**).

Finally, the model using EnJa-Balance achieves the highest score for both Japanese and English tasks, outperforming EnJa-Hybrid despite the latter using a larger corpus (EnJa-Balance = 71.44B tokens, EnJa-Hybrid = 79.62B tokens). This indicates that simply adding more tokens does not necessarily improve performance, highlighting the importance of balancing corpus composition, which we further analyze in the following sections.

4.2 Task-wise Correlation with Corpus Compositional Statistics

4.2.1 Total Token Count

As shown in **Fig. 3**, a strong positive correlation with total token count was observed in MMLU-Medical (`En/MCQA`, $r = 0.954$, $p = 0.003$), suggesting that a larger corpus—regardless of language specificity for either Japanese or English—

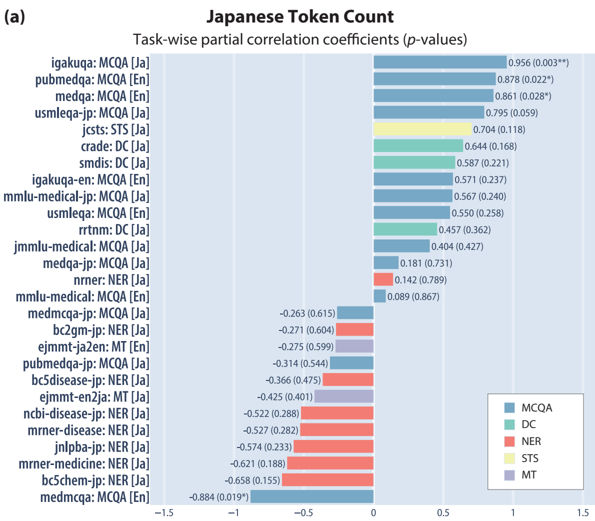

Fig. 4: Task-wise Correlation with Japanese Tokens. **(a)** Japanese token count was positively correlated with IgakuQA, PubMedQA, and MedQA, but negatively with MedMCQA. **(b)** Japanese token ratio showed broader positive correlations, including IgakuQA, MRNER-Disease, MMLU-Medical-Jp, JMMLU-Medical, and MedMCQA-Jp. The x-axis shows partial correlation coefficients with p-values in parentheses.

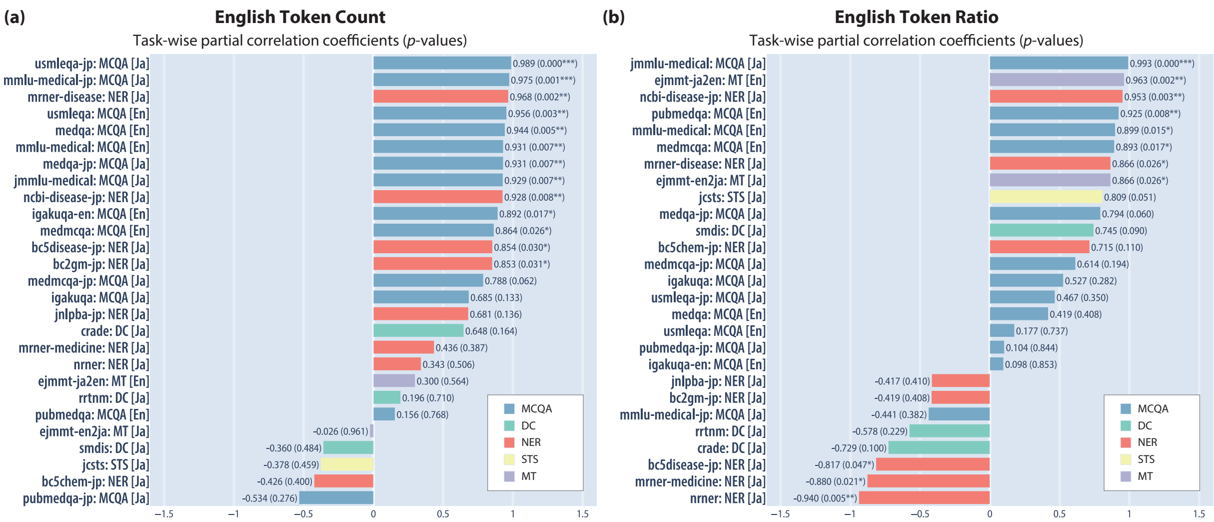

Fig. 5: Task-wise Correlation with English Tokens. **(a)** English token count showed the most consistent and strongest correlations across multiple tasks. **(b)** English token ratio exhibited strong correlations with several tasks but negatively affected specialized Japanese NER tasks (e.g., NRNER). The x-axis shows partial correlation coefficients with p-values in parentheses.

significantly benefited this complex English medical MCQA task. Notably, no other task exhibited significant correlations with total token count, which contrasts with the patterns observed for other language-specific statistics, as presented below.

4.2.2 Japanese Tokens (Count and Ratio)

Fig. 4a shows that the Japanese token count exhibited a strong positive correlation with IgakuQA (`Ja/MCQA`, $r = 0.956$, $p = 0.003$), a representative Japanese medical MCQA task for specialized expertise in the Japanese medical system (Kasai et al., 2023). Surprisingly, certain English MCQA tasks including PubMedQA and MedQA also showed positive correlations with the Japanese token count. This suggests that exposure to diverse linguistic representations, including machine-translated Japanese medical texts and original ones, may have enhanced the model's generalization ability in English medical tasks. In contrast, MedMCQA (`En/MCQA`) exhibited a significant negative correlation with the Japanese token count, suggesting an adverse impact of Japanese token representation. Additionally, as shown in **Fig. 4b**, the Japanese token ratio demonstrated strong positive correlations with some Japanese tasks, such as IgakuQA

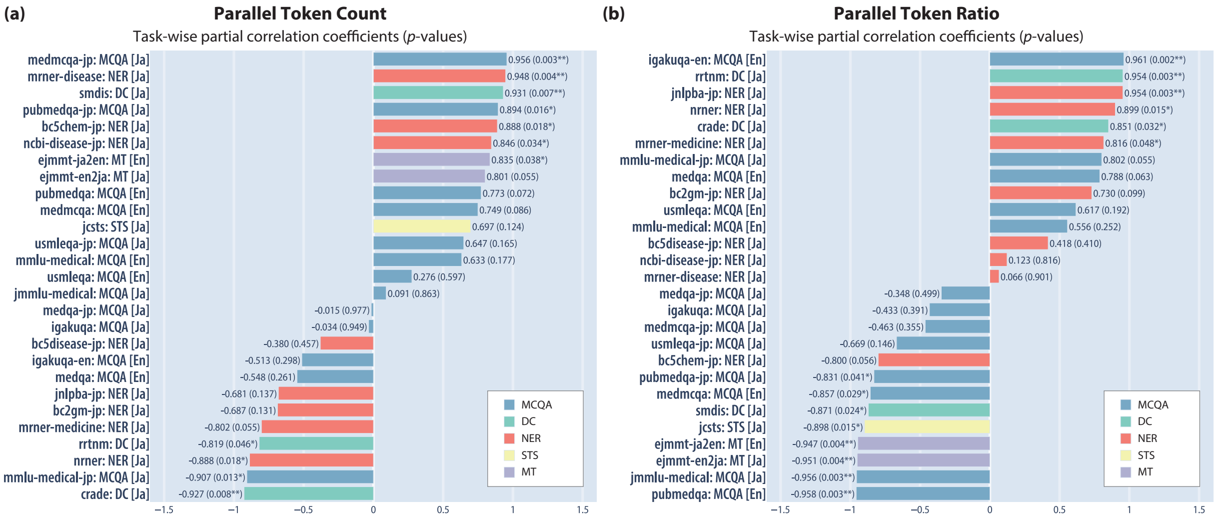

Fig. 6: Task-wise Correlation with Parallel Tokens. (a) Parallel token count showed both positive and negative correlations, with strong positive effects in MedMCQA-Jp, MRNER-Disease, and SMDIS and strong negative effects in CRADE. **(b)** Parallel token ratio exhibited similar trends. The x-axis shows partial correlation coefficients with p-values in parentheses.

(Ja/MCQA) and MRNER-Disease (Ja/NER).

4.2.3 English Tokens (Count and Ratio)

Fig. 5a illustrates that the English token count exhibited the most consistent and strongest correlations across multiple tasks, with 9 tasks showing correlations above 0.9 ($p < 0.01$). The most notable were USMLEQA-Jp (Ja/MCQA, $r = 0.989$, $p < 0.001$) and MMLU-Medical-Jp (Ja/MCQA, $r = 0.975$, $p < 0.001$), suggesting that English token representation plays a critical role in enhancing performance across both Japanese and English medical tasks. This indicates that an English corpus can help the model to acquire medical knowledge that can be exploited even when the task is primarily in Japanese. Similarly, as presented in **Fig. 5b**, the English token ratio demonstrated strong correlations with several tasks, including JMMLU-Medical (Ja/MCQA), EJMMT-Ja2En (En/MT), NCBI-Disease-Jp (Ja/NER), and PubMedQA (En/MCQA). Notably, it also negatively impacted specialized Japanese NER tasks (i.e., NRNER, MRNER-Medicine, and BC5Disease-Jp).

4.2.4 Parallel Tokens (Count and Ratio)

Fig. 6a illustrates that the parallel corpus exhibits both positive and negative correlations across various tasks. In particular, the parallel token count showed strong positive correlations with MedMCQA-Jp (Ja/MCQA, $r = 0.956$), $p = 0.003$), MRNER-Disease (Ja/NER, $r = 0.948$, $p = 0.004$), and SMDIS (Ja/DC, $r = 0.931$, $p = 0.007$), while demonstrating a strong negative correlation with CRADE (Ja/DC, $r = -0.927$, $p = 0.008$). Moreover, **Fig. 6b** indicates that the parallel token ratio positively impacted IgakuQA-En (En/MCQA), RRTNM (Ja/DC), and JNLPBA-Jp (Ja/NER), but exhibited strong negative correlations with PubMedQA (En/MCQA), JMMLU-Medical (Ja/MCQA), EJMMT-En2Ja (Ja/MT), and EJMMT-Ja2En (En/MT). The latter two tasks fall under the category of MT. While parallel corpora are widely regarded as effective for MT tasks (Chi et al., 2022; Hu et al., 2020; Feng et al., 2022; Yang et al., 2023; Lin et al., 2025), these findings suggest that an excessive amount may hinder the learning of language-specific patterns, potentially limiting overall MT performance.

5 Analysis

Here, we analyze the optimal corpus composition for continued pre-training.

RQ1: How do original English and machine-translated Japanese corpora help a bilingual (Japanese–English) PLM achieve domain adaptation in the Japanese medical domain?

In terms of the corpus-alone effect, incorporating *original* English texts (as PMC full-text) is generally more beneficial even for Japanese-domain tasks than using *machine-translated* data, as the model using EnDominant outperformed that using JaDominant (see **Table 1**). This suggests that translation quality can limit its effectiveness in conveying medical knowledge. Nonetheless, machine-translated texts still offer additive gains when used alongside the original English texts, as adding the full machine-translated subcorpus (i.e., EnJa-Balance) leads to an additional performance gain. Thus, the balanced use of machine-translated data with original English texts can be essential for cross-lingual domain adaptation.

RQ2: What is the optimal corpus composition of English and Japanese medical texts for effective continued pre-training of PLMs in a multilingual medical domain?

Partial correlation analysis indicated that each task is differentially sensitive to certain corpus compositional statistics, including the specific ratio of Japanese to English content (see **Fig. 3–6**). Therefore, tailoring a corpus composition for particular downstream tasks by balancing language components facilitates effective knowledge acquisition in practice. Indeed, in our case, the highest average score across both Japanese (20 tasks) and English (7 tasks) was achieved by the model continued pre-trained on EnJa-Balance, even surpassing the EnJa-Hybrid model, which was trained on a larger corpus (see **Table 1**).

RQ3: How do specific compositional statistics within multilingual corpora influence model performance across diverse evaluation tasks?

Effect of the Japanese Corpus: Only the Japanese corpus positively correlated with IgakuQA (see **Fig. 4**), a unique MCQA benchmark requiring specialized Japanese medical system expertise. This underscores the importance of incorporating language-specific resources with localized knowledge alongside translated general knowledge. It also benefits some English MCQA tasks like PubMedQA and MedQA. We hypothesize that exposure to diverse linguistic representations enhances the model's generalization in English medical tasks. However, excessive Japanese corpus may impede certain English-specific tasks, as shown by its negative correlation with MedMCQA.

Effect of the English Corpus: The size of the English corpus exhibited a strong correlation with score improvements not only in English QA tasks (e.g., USMLEQA, MedQA, and MMLU-Medical) but also in select Japanese tasks (e.g., USMLEQA-Jp and MMLU-Medical-Jp) (see **Fig. 5**). This suggests that an English corpus can effectively transfer medical knowledge to Japanese tasks, improving performance even when the task is primarily in Japanese. However, an excessive proportion of English tokens may degrade performance in Japanese-specific tasks, particularly those related to NER.

Effect of the Parallel Corpus: The parallel corpus exhibited both positive and negative correlations depending on the task type (see **Fig. 6**). On one hand, the size of the parallel corpus showed strong positive correlations with several tasks, suggesting that bilingual alignment facilitates cross-lingual knowledge transfer between English and Japanese. On the other hand, an excessive proportion of parallel data negatively impacted some tasks, even including MT tasks. This might be because parallel corpora switch languages at the paragraph level, which is unnatural as a language-specific pattern and negatively affects the performance of certain tasks (see **Appendix F** for an example).

6 Conclusions

We systematically examined how continued pre-training on Japanese and English medical domain corpora—at varying proportions—affects task performance to seek optimal corpus composition for the comprehensive Japanese–English medical benchmark. The results suggest that effective cross-lingual domain adaptation requires (1) leveraging specialized knowledge from well-resourced corpora, (2) ensuring sufficient coverage of language-specific expressions in the target language, and (3) using parallel corpora in moderation. These findings highlight the importance of balanced corpus design that accounts for both linguistic diversity and domain-specific terminology, particularly in settings involving a well-resourced source language and a low-resource target language. While grounded in the Japanese–English medical context, these insights are broadly applicable to multilingual adaptation of PLMs across diverse domains.

Limitations

One limitation of this study is the small sample size; however, the strong effect sizes, reflected in large correlation coefficients and low p-values, reinforce the reliability of the key findings. Besides, this study primarily identifies correlations between corpus compositional statistics and task performance without directly addressing causal interpretations. However, the additive and ablative design of the corpus composition allows for certain causal inferences rather than merely reflecting statistical correlations (see **Section 4.1**). Further controlled experiments and deeper analyses are needed to establish definitive causal relationships.

References

Badr AlKhamissi, Millicent Li, Asli Celikyilmaz, Mona Diab, and Marjan Ghazvininejad. 2022. A Review on Language Models as Knowledge Bases. *arXiv preprint arXiv:2204.06031*.

Abhijeet Awasthi, Nitish Gupta, Bidisha Samanta, Shachi Dave, Sunita Sarawagi, and Partha Talukdar. 2023. Bootstrapping Multilingual Semantic Parsers using Large Language Models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2455–2467.

Giuseppe Castellucci, Simone Filice, Danilo Croce, and Roberto Basili. 2021. Learning to Solve NLP Tasks in an Incremental Number of Languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 837–847.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An Information-Theoretic Framework for Cross-Lingual Language Model Pre-Training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588.

Zewen Chi, Shaohan Huang, Li Dong, Shuming Ma, Bo Zheng, Saksham Singhal, Payal Bajaj, Xia Song, Xian-Ling Mao, Heyan Huang, and Furu Wei. 2022. XLM-E: Cross-lingual Language Model Pre-training via ELECTRA. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6170–6182.

Nigel Collier, Tomoko Ohta, Yoshimasa Tsuruoka, Yuka Tateisi, and Jin-Dong Kim. 2004. Introduction to the Bio-entity Recognition Task at JNLPBA. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP)*, pages 73–78.

Yiming Cui, Ziqing Yang, and Xin Yao. 2024. Efficient and Effective Text Encoding for Chinese LLaMA and Alpaca. *arXiv preprint arXiv:2304.08177*.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. NCBI Disease Corpus: A Resource for Disease Name Recognition and Concept Normalization. *Journal of biomedical informatics*, 47:1.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT Sentence Embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891.

Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. 2024. Continual Pre-Training for Cross-Lingual LLM Adaptation: Enhancing Japanese Language Capabilities. In *First Conference on Language Modeling (COLM)*.

Kshitij Gupta, Benjamin Thérien, Adam Ibrahim, Mats Leon Richter, Quentin Gregory Anthony, Eugene Belilovsky, Irina Rish, and Timothée Lesort. 2023. Continual Pre-Training of Large Language Models: How to re-warm your model? In *Workshop on Efficient Systems for Foundation Models @ ICML2023*.

Katharina Hämmerl, Jindřich Libovický, and Alexander Fraser. 2024. Understanding Cross-Lingual Alignment—A Survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10922–10943.

Takeshi Hayakawa and Yuki Arase. 2020. Fine-Grained Error Analysis on English-to-Japanese Machine Translation in the Medical Domain. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 155–164.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Seongtae Hong, Seungyoon Lee, Hyeonseok Moon, and Heuiseok Lim. 2025. MIGRATE: Cross-Lingual Adaptation of Domain-Specific LLMs through Code-Switching and Embedding Transfer. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9184–9193.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421.

Jaavid Aktar Husain, Raj Dabre, Aswanth Kumar, Jay Gala, Thanmay Jayakumar, Ratish Puduppully, and Anoop Kunchukuttan. 2024. RomanSetu: Efficiently unlocking multilingual capabilities of Large Language Models via Romanization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15593–15615.

Ayyoob ImaniGooghari, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling Multilingual Corpora and Language Models to 500 Languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117.

Junfeng Jiang, Jiahao Huang, and Akiko Aizawa. 2025. JMedBench: A Benchmark for Evaluating Japanese Biomedical Large Language Models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5918–5935.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What Disease Does This Patient Have? A Large-Scale Open Domain Question Answering Dataset from Medical Exams. *Applied Sciences*, 11(14).

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. 2019. PubMedQA: A Dataset for Biomedical Research Question Answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577.

Jungo Kasai, Yuhei Kasai, Keisuke Sakaguchi, Yutaro Yamada, and Dragomir Radev. 2023. Evaluating gpt-4 and chatgpt on japanese medical licensing examinations. *arXiv preprint arXiv:2303.18027*.

Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. 2016. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database : the journal of biological databases and curation*, 2016.

Peiqin Lin, André F. T. Martins, and Hinrich Schütze. 2025. A Recipe of Parallel Corpora Exploitation for Multilingual Large Language Models. *arXiv preprint arXiv:2407.00436*.

LLM-jp, Akiko Aizawa, Eiji Aramaki, Bowen Chen, Fei Cheng, Hiroyuki Deguchi, Rintaro Enomoto, Kazuki Fujii, Kensuke Fukumoto, Takuya Fukushima, Namgi Han, Yuto Harada, Chikara Hashimoto, Tatsuya Hiraoka, Shohei Hisada, Sosuke Hosokawa, Lu Jie, Keisuke Kamata, Teruhito Kanazawa, and 63 others. 2024. LLM-jp: A Cross-organizational Project for the Research and Development of Fully Open Japanese LLMs. *arXiv preprint arXiv:2407.03963*.

Nick Mecklenburg, Yiyou Lin, Xiaoxiao Li, Daniel Holstein, Leonardo Nunes, Sara Malvar, Bruno Silva, Ranveer Chandra, Vijay Aski, Pavan Kumar Reddy Yannam, Tolga Aktas, and Todd Hendry. 2024. Injecting New Knowledge into Large Language Models via Supervised Fine-Tuning. *arXiv preprint arXiv:2404.00213*.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. MedMCQA: A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174, pages 248–260.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language Models as Knowledge Bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473.

Ramon Pires, Hugo Abonizio, Thales Sales Almeida, and Rodrigo Nogueira. 2023. Sabiá: Portuguese Large Language Models. In *Intelligent Systems*, pages 226–240.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S. Yu. 2025. A survey of multilingual large language models. *Patterns*, 6(1):101118.

Vipul Rathore, Rajdeep Dhingra, Parag Singla, and Mausam. 2023. ZGUL: Zero-shot Generalization to Unseen Languages using Multi-source Ensembling of Language Adapters. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6969–6987.

Evgeniia Razumovskaia, Ivan Vulić, and Anna Korhonen. 2024. Analyzing and Adapting Large Language Models for Few-Shot Multilingual NLU: Are We There Yet? *arXiv preprint arXiv:2403.01929*.

Uri Shaham, Jonathan Herzig, Roee Aharoni, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. 2024. Multilingual Instruction Tuning With Just a Pinch of Multilinguality. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2304–2317.

Vaibhav Singh, Amrith Krishna, Karthika NJ, and Ganesh Ramakrishnan. 2024. A Three-Pronged Approach to Cross-Lingual Adaptation with Multilingual LLMs. *arXiv preprint arXiv:2406.17377*.

Larry Smith, Lorraine K. Tanabe, Rie Ando, Cheng Ju Kuo, I. Fang Chung, Chun Nan Hsu, Yu Shi Lin, Roman Klinger, Christoph M. Friedrich, Kuzman Ganchev, Manabu Torii, Hongfang Liu, Barry Haddow, Craig A. Struble, Richard J. Povinelli, Andreas Vlachos, William A. Baumgartner, Lawrence Hunter, Bob Carpenter, and 15 others. 2008. Overview of BioCreative II gene mention recognition. *Genome biology*, 9 Suppl 2.

Eshaan Tanwar, Subhabrata Dutta, Manish Borthakur, and Tanmoy Chakraborty. 2023. Multilingual LLMs are Better Cross-lingual In-context Learners with Alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6292–6307.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv:2307.09288*.

Zhen Wan, Yating Zhang, Yexiang Wang, Fei Cheng, and Sadao Kurohashi. 2024. Reformulating Domain Adaptation of Large Language Models as Adapt-Retrieve-Revise: A Case Study on Chinese Legal Domain. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5030–5041.

Shijie Wu, Benjamin Van Durme, and Mark Dredze. 2022. Zero-shot Cross-lingual Transfer is Under-specified Optimization. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 236–248.

Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Kexin XU, Yuqi Ye, and Hanwen Gu. 2024. A Survey on Multilingual Large Language Models: Corpora, Alignment, and Bias. *Frontiers of Computer Science*.

Ikuya Yamada and Ryokan Ri. 2024. LEIA: Facilitating Cross-lingual Knowledge Transfer in Language Models with Entity-based Data Augmentation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7029–7039.

Wen Yang, Chong Li, Jiajun Zhang, and Chengqing Zong. 2023. BigTranslate: Augmenting Large Language Models with Multilingual Translation Capability over 100 Languages. *arXiv preprint arXiv:2305.18098*.

Zheng Xin Yong, Hailey Schoelkopf, Niklas Muennighoff, Alham Fikri Aji, David Ifeoluwa Adelani, Khalid Almubarak, M Saiful Bari, Lintang Sutawika, Jungo Kasai, Ahmed Baruwa, Genta Winata, Stella Biderman, Edward Raff, Dragomir Radev, and Vassilina Nikoulina. 2023. BLOOM+1: Adding Language Support to BLOOM for Zero-Shot Prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11682–11703.

Yuanchi Zhang, Yile Wang, Zijun Liu, Shuo Wang, Xiaolong Wang, Peng Li, Maosong Sun, and Yang Liu. 2024. Enhancing Multilingual Capabilities of Large Language Models through Self-Distillation from Resource-Rich Languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11189–11204.

Jun Zhao, Zhihao Zhang, Luhui Gao, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024a. LLaMA Beyond English: An Empirical Study on Language Capability Transfer. *arXiv preprint arXiv:2401.01055*.

Weixiang Zhao, Yulin Hu, Jiahe Guo, Xingyu Sui, Tongtong Wu, Yang Deng, Yanyan Zhao, Bing Qin, Wanxiang Che, and Ting Liu. 2024b. Lens: Rethinking Multilingual Enhancement for Large Language Models. *arXiv preprint arXiv:2410.04407*.

Wenhao Zhu, Yunzhe Lv, Qingxiu Dong, Fei Yuan, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Extrapolating Large Language Models to Non-English by Aligning Languages. *arXiv preprint arXiv:2308.04948*.

Appendix A Overview of JMedBench

Appendix A.1 Multi-Choice Question Answering (MCQA)

MedMCQA/MedMCQA-Jp MedMCQA is a large-scale, MCQA dataset designed to address real-world medical entrance exam questions, covering 2.4 thousand health topics and 21 medical subjects sampled from medical entrance exams across India (Pal et al., 2022). This contains 4,183 test samples. MedMCQA-Jp is a Japanese translation of MedMCQA.

USMLEQA/USMLEQA-Jp USMLEQA is a large-scale, MCQA dataset with 1,273 test samples with 4 options, which are sampled from United States Medical Licensing Examinations (Jin et al., 2021). USMLEQA-Jp is a Japanese translation of USMLEQA, containing the same number of test samples.

MedQA/MedQA-Jp MedQA is a 5-option version of USMLEQA, known as a representative benchmark for medical large language models in the assessment of medical knowledge sufficient for medical licensure (Jin et al., 2021). MedQA-Jp is a Japanese translation of MedQA, containing the same number of test samples.

MMLU-Medical/MMLU-Medical-Jp MMLU-Medical contains 1,871 biomedical questions at the college level as test samples, which is extracted as a subset of a large-scale, multi-topics benchmark, MMLU (Hendrycks et al., 2021). MMLU-Medical-Jp is a Japanese translation of MMLU-Medical.

JMMLU-Medical While the MMLU-Medical-Jp is a machine-translated version of MMLU-Medical, JMMLU-Medical consists of human-translated Japanese version of MMLU-Medical comprising 1,271 test samples[6].

IgakuQA/IgakuQA-En IgakuQA contains 989 Japanese questions based on Japanese medical licensing examinations from 2018 to 2022 (Kasai et al., 2023). This uniquely reflects Japanese-specific medical practices, healthcare systems, and epidemiological profiles. IgakuQA-En is an English translation of IgakuQA.

[6] https://huggingface.co/datasets/nlp-waseda/JMMLU

PubMedQA/PubMedQA-Jp PubMedQA contains 1,000 test samples focusing on the biomedical field collected from PubMed Abstracts (Jin et al., 2019). The task of PubMedQA is to answer research questions with yes/no/maybe. PubMedQA-JP is a Japanese translation of PubMedQA.

Appendix A.2 Machine Translation (MT)

EJMMT-Ja/EJMMT-En EJMMT is a Japanese–English medical machine-translation dataset with fine-grained annotation of error spans and error types (Hayakawa and Arase, 2020). EJMMT-Ja indicates the translation accuracy in the direction of English to Japanese, while EJMMT-En indicates the Japanese to English direction. These include 2,400 test samples.

Appendix A.3 Named Entity Recognition (NER)

MRNER-Medicine MRNER-Medicine (Medical Report Named Entity Recognition for medicine) contains 90 test samples for extracting medication-related information from case reports in Japanese[7].
MRNER-Disease MRNER-Disease (Medical Report Named Entity Recognition for positive disease) contains 90 test samples for extracting symptoms actually observed in patients from case reports and radiology reports in Japanese[7].
NRNER NRNER (Nursing Record Named Entity Recognition) contains 90 test samples, involving extracting information about symptoms actually observed in patients and medication from simulated nursing records in Japanese[7].
BC2GM-Jp BC2GM-Jp is a Japanese translation of BC2GM (BioCreative II Gene Mention Recognition) (Smith et al., 2008), which contains 5,037 test samples to identify a gene mention in a sentence.
BC5Chem-Jp BC5Chem-Jp is a Japanese translation of BC5Chem (Li et al., 2016), which contains 4,801 test samples to identify disease, chemical entities and their relations from biomedical texts.
BC5Disease-Jp BC5Disease-Jp is a Japanese translation of BC5Disease (Li et al., 2016), which contains 4,797 test samples to identify disease, chemical entities and their relations from biomedical texts.
JNLPBA-Jp JNLPBA-Jp is a Japanese translation of JNLPBA (Collier et al., 2004), which features 4,260 test samples for bio-entity recognition, identifying and classifying technical terms in the domain of molecular biology.
NCBI-Disease-Jp NCBI-Disease-Jp is a Japanese translation of NCBI-Disease (Doğan et al., 2014), which contains 940 test samples to identify the disease name on the NCBI disease corpus.

Appendix A.4 Document Classification (DC)

CRADE CRADE (Case Report Adverse Drug Event) contains 92 test samples, which involves classifying the possibility of adverse events from medications and symptoms in case reports in Japanese[7].
RRTNM RRTNM (Radiology Report Tumor Nodes Metastasis) contains 89 test samples, which involves predicting TNM classification of cancer from radiology reports of lung cancer patients in Japanese[7].
SMDIS SMDIS (Social Media Disease) comprises 84 test samples, which involve classifying the presence or absence of diseases or symptoms of the poster or people around them from simulated Tweets in Japanese[7].

Appendix A.5 Semantic Text Similarity (STS)

JCSTS JCSTS (Japanese Clinical Semantic Textual Similarity) has 3,500 test samples in Japanese. This is a medical version of the semantic textual similarity task that determines the semantic similarity between two sentences, dealing with case reports[7].

Appendix B Translation Performance of the Machine-Translation Models

The English–to–Japanese translation performance of the machine translation models—including our model[8], which was used to translate the PMC subcorpus and PubMed abstracts—as well as comparative models, is evaluated on EJMMT. As shown in **Table B.1**, the model used in this research demonstrates relatively high performance. “Baseline in EJMMT” refers to the baseline performance reported in Hayakawa and Arase (2020). BLEU was used to measure the degree of agreement with the ground truth, employing the SacreBLEU library[9]

[7]This benchmark is originally included in JMED-LLM (Japanese Medical Evaluation Dataset for Large Language Models): https://github.com/sociocom/jmed-llm

[8]We used the science translation engine provided courtesy of the National Institute of Information and Communications Technology (NICT).

[9]https://github.com/mjpost/sacrebleu

Translation Model	**BLEU**	**COMET-22**	**COMET-23**
Ours	**37.71**	**80.78**	65.64
Baseline in EJMMT	26.77	77.86	64.93
gpt-4o-2024-08-06	27.23	79.86	**68.16**

Table B.1: Translation Performance of the Machine-Translation Models

with the MeCab tokenizer[10]. COMET-22[11] and COMET-23[12] were used as neural frameworks for machine translation evaluation.

Appendix C Compositional Statistics of Multilingual Corpora

Compositional statistics of multilingual corpora are shown in **Table C.1**.

Appendix D Training Details

The bilingual (Japanese–English) PLMs, namely `llm-jp/llm-jp-3-13b` and its equivalent model, were pre-trained from scratch on 2.1 trillion tokens using a general corpus containing both English and Japanese text.[13] Their architectures, including the hidden size, number of attention heads, number of layers, and context length, are identical to those of Llama 2 (Touvron et al., 2023). For continued pre-training, we used a global batch size of 1024, employing the Adam optimizer with a cosine scheduler. The hyperparameters were as follows: $\beta_1 = 0.9$, $\beta_2 = 0.95$, $\epsilon = 1.0 \times 10^{-8}$, learning rate = 1×10^{-4}, minimum learning rate = 1×10^{-5}, warm-up fraction = 0.03, and weight decay = 0.1.

Then, we performed supervised fine-tuning on both models, before and after continued pre-training. We used the first version of the general-domain instruction tuning dataset published by `llm-jp`[14], along with the original training datasets from MedQA (Jin et al., 2021), PubMedQA (Jin et al., 2019), and MedMCQA (Pal et al., 2022), as well as Japanese translations of the MedQA and PubMedQA training datasets. Additionally, we incorporated past questions from the Japanese National Medical Examination spanning 12 years, excluding any portions overlapping with IgakuQA (Kasai et al., 2023). Such instruction tuning is necessary because JMedBench requires a basic instruction-following capability. As for the training settings, we used a global batch size of 64, employing the Adam optimizer with a cosine scheduler over two epochs. The other hyperparameters were as follows: $\beta_1 = 0.9$, $\beta_2 = 0.98$, learning rate = 2×10^{-5}, minimum learning rate = 2×10^{-6}, warm-up steps = 20, and weight decay = 0.1.

The computational budget used in this study is as follows: For continued pre-training of the 13B models using approximately 80B tokens of EnJa-Hybrid corpus, we required a computational cluster consisting of 32 nodes, each equipped with 8 NVIDIA H100 GPUs (total GPU count: 8×32 = 256 GPUs), with a computation time of about 24 hours. Continued pre-training using other corpora required computation time proportional to their token count. Additionally, for supervised fine-tuning, we used 8 nodes, each equipped with 8 NVIDIA H100 GPUs (total GPU count: 8×8 = 64 GPUs), requiring approximately 2 hours of computation time. For evaluation based on JMedBench, we used only a single node equipped with 8 NVIDIA H100 GPUs, requiring about 1 hour.

Appendix E Mutual Correlation Between Covariates

Here, we demonstrate the necessity of partial correlation analysis as employed in this study. Compositional statistics of corpora may exhibit correlations with one another, such as the relationship where an increased Japanese token count naturally leads to an increase in the total token count. Similarly, JMedBench includes some related tasks, for example, both MMLU-Medical-Jp and JMMLU-Medical originate from MMLU-Medical as their English source; therefore, it is essential to account for correlations between task scores.

To illustrate these interdependencies, **Fig. E.1** presents mutual correlation coefficients among compositional statistics in multilingual corpora,

[10] https://pypi.org/project/mecab-python3/

[11] https://huggingface.co/Unbabel/wmt22-cometkiwi-da

[12] https://huggingface.co/Unbabel/wmt23-cometkiwi-da-xl

[13] We used two functionally equivalent base models, both pre-trained from scratch on a total of 2.1T tokens, differing only in the composition of the final 0.3T tokens.

[14] https://huggingface.co/llm-jp/llm-jp-13b-v1.0

Corpus Name	Total token count (B)	Japanese token count (B)	English token count (B)	Parallel token count	Japanese token ratio (%)	English token ratio (%)	Parallel token ratio (%)
EnJa-Base	15.00	5.00	9.50	0.48	33.38	63.36	3.26
JaDominant	42.76	32.76	9.50	0.48	76.62	22.24	1.14
EnDominant	43.68	5.00	38.18	0.48	11.47	87.41	1.12
EnJa-Plus	57.56	18.88	38.18	0.48	32.81	66.34	0.85
EnJa-Balance	71.44	32.76	38.18	0.48	45.86	53.45	0.68
EnJa-Hybrid	79.62	36.11	36.42	7.07	45.36	45.75	8.89

Table C.1: Compositional Statistics of Multilingual Corpora

while **Fig.** E.2 shows mutual correlation coefficients of task-wise score differences among continued pre-trained models using multilingual corpora.

Moreover, **Fig.** E.3 illustrates the difference between regular correlation analysis and partial correlation analysis, using the score difference in IgakuQA (`Ja/MCQA`) as an example. For instance, while total token count exhibited a significant correlation with score difference in the regular correlation analysis ($r = 0.915$, $p = 0.011$), this effect disappeared in the partial correlation analysis ($r = 0.185$, $p = 0.725$). Instead, the effect of Japanese token count turned out to be significant ($r = 0.956$, $p = 0.003$), which is more intuitive when considering the specific expertise tested in this particular benchmark.

Thus, by adjusting for the effects of covariates through partial correlation analysis, we can better distinguish the correlations between task-wise score differences and corpus compositional statistics.

Appendix F Example of Parallel Corpus

An example of the parallel corpus is shown in **Fig.** F.1. An original PubMed abstract in English and its machine-translated Japanese version are concatenated at the paragraph level. Notably, the machine-translated data is of reasonable quality, accurately rendering biomedical terminology even in specialized contexts. This observation is well-aligned with the quantitative comparison of the translation quality (see **Table** B.1).

However, because the parallel corpus is artificially constructed to switch languages at the paragraph level, it deviates from natural language patterns and may potentially hinder certain learning tasks.

Fig. E.1: Mutual Correlation Coefficients of Corpus Compositional Statistics. Mutual correlation coefficients among compositional statistics in multilingual corpora—including total token count, Japanese token count, English token count, parallel token count, Japanese token ratio, English token ratio, and parallel token ratio—were computed. The results reveal several strong correlations between specific statistics.

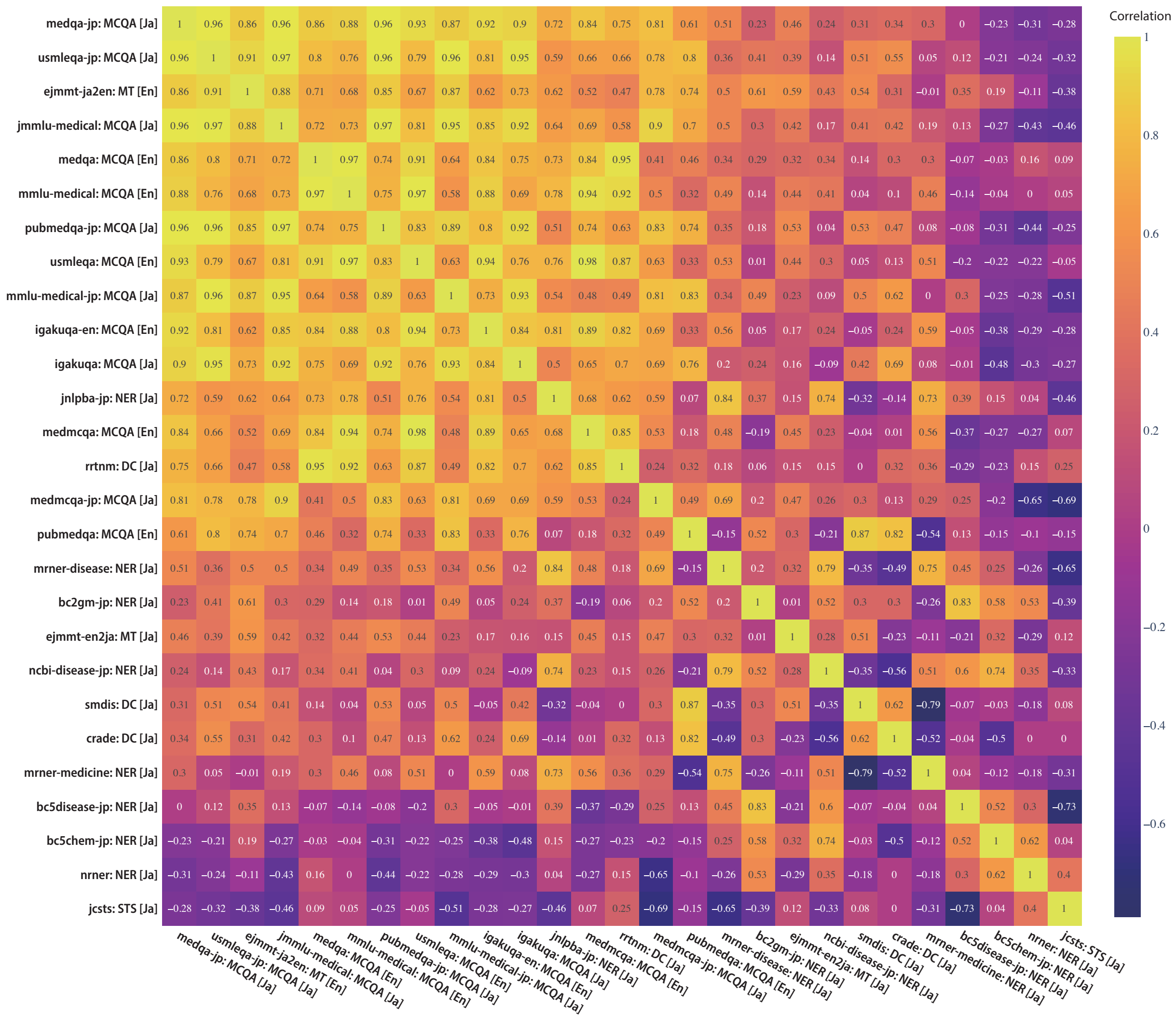

Fig. E.2: Mutual Correlation Coefficients of Task-wise Score Differences. Mutual correlation coefficients of task-wise score differences among continued pre-trained models using various corpora—including EnJa-Base, JaDominant, EnDominant, EnJa-Plus, EnJa-Balance, and EnJa-Hybrid—were computed. The results reveal several strong correlations between specific tasks.

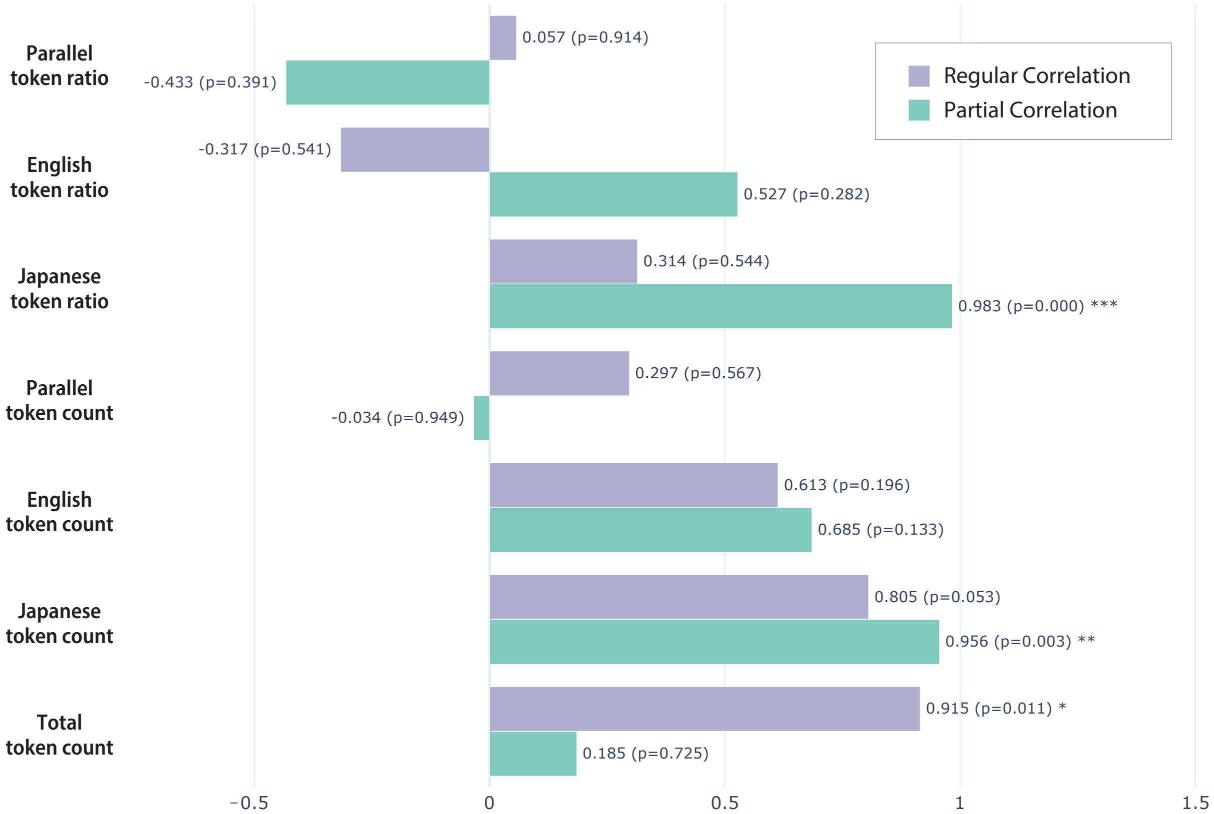

Fig. E.3: Difference between Partial Correlation and Regular Correlation. Comparison between regular correlation and partial correlation analyses for the IgakuQA (`Ja/MCQA`) task. Notably, total token count exhibited a significant correlation with score difference in the regular correlation analysis ($r = 0.915$, $p = 0.011$), but this effect disappeared in the partial correlation analysis ($r = 0.185$, $p = 0.725$). Instead, the effect of Japanese token count turned out to be significant ($r = 0.956$, $p = 0.003$). This demonstrates that partial correlation analysis can reveal differential relationships between task-wise score differences and corpus compositional statistics by adjusting for the effects of covariates.

Parvovirus B19 is the causative agent of erythema infectiosum in children, but the virus is associated with an increasing range of different diseases. These include acute and chronic arthritis, hydrops fetalis in pregnant women, aplastic anemia, and thrombocytopenia. The host's immune response is directed against the viral structural proteins VP1 and VP2. This study investigated the presence of IgG against the viral nonstructural protein NS1 using Western blot. Serum panels from healthy individuals, B19-infected pregnant women, and various disease groups were tested. The disease groups included patients with symptoms that may be linked to parvovirus B19 infection. The results showed that IgG against the NS1 protein was present in 22% of healthy individuals with past B19 infection. In cases of persistent or prolonged B19 infections, the prevalence of NS1-specific antibodies was as high as 80%. It is concluded that NS1-specific IgG may be used as an indicator of chronic or more severe courses of parvovirus B19 infections. パルボウイルスB19は小児の伝染性紅斑の原因ウイルスであるが、このウイルスは様々な疾患の増加と関連している。これらには、急性および慢性関節炎、妊婦の胎児水腫、再生不良性貧血、血小板減少などがある。宿主の免疫反応はウイルス構造タンパク質VP1とVP2に向けられる。本研究では、ウイルス非構造タンパク質NS1に対するIgGの存在をウェスタンブロットを用いて調べた。健常者、B19感染妊婦、および種々の疾患群の血清中パネルを検査した。疾患群には、パルボウイルスB19感染に関連する可能性のある症状を有する患者が含まれていた。その結果、NS1蛋白に対するIgGは、過去にB19に感染した健常者の22%に存在していた。B19感染が持続または遷延した症例では、NS1特異的抗体の保有率は80%と高かった。NS1特異的IgGは、パルボウイルスB19感染の慢性あるいはより重症な経過の指標として使用できると結論した。

Fig. F.1: An Example of a Parallel Corpus. The parallel corpus is constructed by arranging machine-translated Japanese paragraphs alongside their original English counterparts in random sequences. As a result, the text exhibits random language switching between English and Japanese at the paragraph level, creating an artificial linguistic environment that differs from language-specific textual patterns.