International Journal of Artificial Intelligence
p-ISSN: 2407-7275, e-ISSN: 2686-3251

Original Research Paper

Evaluation of Perplexity and Syntactic Handling Capabilities of ClueAI Models on Japanese Medical Texts

Tatsuhiro Haga[1*], Keiyo Matsumoto[2], Ippei Asahiko[2], Shunzo Mizoguchi[1]

[1] *School of Engineering, Shibaura Institute of Technology. Saitama, Japan.*
[2] *College of Industrial Technology, Nihon University. Tokyo, Japan.*

Article History
Received:
06.11.2024

Revised:
12.12.2024

Accepted:
07.01.2025

***Corresponding Author:**
Tatsuhiro Haga
Email:
thiro.haga@gmail.com

Abstract: This study aims to evaluate the effectiveness of a large Japanese language model, ClueAI, tailored to the medical domain, in the task of predicting Japanese medical texts. The background of this study is the limitations of general language models, including multilingual models such as multilingual BERT, in handling linguistic complexity and specific terminology in Japanese medical texts. The research methodology includes fine-tuning the ClueAI model using the MedNLP corpus, with a MeCab-based tokenization approach through the Fugashi library. The evaluation is carried out using the perplexity metric to measure the model's generalization ability in predicting texts probabilistically. The results show that ClueAI that has been tailored to the medical domain produces lower perplexity values than the multilingual BERT baseline, and is better able to understand the context and sentence structure of medical texts. MeCab-based tokenization is proven to contribute significantly to improving prediction accuracy through more precise morphological analysis. However, the model still shows weaknesses in handling complex syntactic structures such as passive sentences and nested clauses. This study concludes that domain adaptation provides improved performance, but limitations in linguistic generalization remain a challenge. Further research is recommended to explore models that are more sensitive to syntactic structures, expand the variety of medical corpora, and apply other Japanese language models in broader medical NLP tasks such as clinical entity extraction and classification.

Keywords: ClueAI, Japanese LLM, MeCab Tokenization, Medical NLP, Multilingual BERT.

1. Introduction

The integration of artificial intelligence (AI) in healthcare has significantly transformed medical diagnostics, patient care, and administrative workflows. Among AI technologies, Natural Language Processing (NLP) plays a crucial role in analyzing vast amounts of medical texts. However, applying NLP to Japanese medical texts poses unique challenges due to the language's complexity, such as the use of kanji, hiragana, and katakana scripts without explicit word boundaries, and the sensitive nature of medical data. Tokenization, a fundamental NLP task, is particularly difficult and requires specialized tools like MeCab to handle morphological ambiguities effectively [1]. Moreover, clinical narratives contain domain-specific terminology, abbreviations, and context-dependent meanings, complicating accurate interpretation necessary for disease prediction, treatment recommendations, and patient monitoring. The limited availability of annotated Japanese medical corpora further restricts the development of robust NLP models for this domain [2].

Large Language Models (LLMs) like GPT and BERT have shown remarkable ability in understanding and generating human-like text and can be fine-tuned for domain-specific applications such as medical text analysis. Nevertheless, most LLMs are predominantly trained on English corpora, which limits their performance on Japanese medical texts [3]. Recent initiatives, such as ClueAI, have developed LLMs trained on Japanese datasets to bridge this linguistic gap, but their effectiveness in the medical domain remains underexplored [4]. The MedNLP corpus, composed of authentic Japanese medical records, provides a valuable resource for training and evaluating NLP models in realistic clinical scenarios [5], offering potential to enhance the relevance and accuracy of NLP in Japanese healthcare.

Evaluating NLP models requires appropriate metrics; perplexity, which quantifies a model's uncertainty in predicting the next word, is a standard measure where lower values indicate better predictive performance, making it suitable for medical text prediction tasks [6] [7]. Using multilingual BERT fine-tuned as a baseline enables benchmarking improvements from domain-specific adaptation [8] [9]. Comparing fine-tuned ClueAI against this baseline helps assess the benefits of targeted training [10].

This study aims to adapt the Japanese LLM ClueAI for medical text prediction using the MedNLP corpus. By fine-tuning on domain-specific data, we seek to improve predictive accuracy measured by perplexity and compare it with multilingual BERT. Additionally, we analyze prediction errors and biases to understand model limitations, crucial for ensuring reliable application in healthcare, where errors can have serious consequences.

The significance of this research lies in advancing effective and accurate NLP systems tailored to the Japanese medical context, addressing linguistic challenges, and leveraging domain-specific data to support better clinical decision-making and patient outcomes. Furthermore, the methodology and findings may guide the development of similar NLP applications in other languages and domains, contributing to global healthcare innovation.

2. Introduction

2.1. Global and Local LLM Studies

Large Language Models (LLMs) such as GPT-3, LLaMA, and BLOOM have been rapidly developing and become the backbone of many natural language processing (NLP) applications. These models are trained on large-scale datasets and have shown outstanding performance in various domains, such as text generation, machine translation, and question and answer [11]. GPT-3, developed by OpenAI, is one of the most well-known models in the world, which generates human-like text and is used in applications such as automated writing assistants, chatbots, and code generation [12].

LLaMA (Large Language Model Meta AI) developed by Meta, has attracted attention due to its efficient architecture and provides powerful language modeling capabilities at a lower computational cost. LLaMA is an open-source model, making it a popular choice for researchers who want to build custom NLP systems [13]. The model's scalability allows it to be used in a variety of tasks, such as medical text summarization or analysis, which is particularly important in specialized fields such as healthcare.

BLOOM, developed by a collaboration of AI researchers, is another cutting-edge model that serves as a multilingual alternative to GPT-3. Unlike GPT-3, BLOOM is designed to generate text in multiple languages, making it suitable for applications in diverse linguistic contexts [14]. BLOOM's multilingual capabilities have gained significant attention, especially in non-English regions, as it

enables better model performance for languages such as Japanese that have different grammatical structures and writing systems.

Localized models such as ClueAI, designed specifically for Japanese, have shown promising results in tasks such as text generation, classification, and sentiment analysis. These models are optimized to handle the specificities of Japanese, including its complex writing system and rich morphology [15]. While LLMs such as GPT-3 and LLaMA have achieved global success, localized models such as ClueAI show potential for more tailored applications in specific linguistic environments.

Despite the global success of LLMs, challenges remain in applying these models to languages such as Japanese. Differences in sentence structure, word segmentation, and semantic ambiguity between English and Japanese highlight the limitations of using general LLMs without adaptation [16]. This suggests the need for further research on localized LLM models, especially in the medical field, where precise domain knowledge and terminology are critical.

Recent research has explored the application of LLMs to various Japanese language tasks. Researchers have fine-tuned models such as BERT for specific domains, including health, to improve model performance in tasks such as medical diagnosis prediction and clinical decision making [17]. However, these models still face challenges in handling medical texts due to the complexity of medical terminology that often requires specific adaptation.

The growing interest in LLMs for Japanese encourages further research into the customization of these models. Fine-tuning techniques, such as domain adaptation and data augmentation, have been explored to improve the effectiveness of LLMs in specific tasks, including medical text processing. This is an important step towards bridging the gap between the versatile LLM and applications that require high accuracy in the medical context [18].

The continued development of LLMs, both globally and locally, demonstrates their transformative potential in a variety of NLP applications, including medical text prediction. However, further research is needed to fine-tune these models for specific domains, especially in Japanese, to improve their effectiveness in real-world applications.

2.2. Medical NLP Research in Japan

The integration of natural language processing (NLP) in medical applications in Japan has gained significant attention in recent years. One of the most widely used tools for processing Japanese text in the medical domain is MeCab, a morphological analyzer that is essential for Japanese sentence segmentation. MeCab helps in breaking down Japanese sentences into words or phrases, which is essential for advanced tasks such as information extraction, text classification, and named entity recognition (NER) [19]. This tool is widely used in various NLP tasks in Japan, especially for processing medical data that includes complex terminology.

MeCab has been integrated with various Python libraries, such as Fugashi, to facilitate tokenization and preprocessing of Japanese text. By applying these tokenization techniques, researchers can prepare medical corpora for tasks such as clinical text classification and disease prediction. These tools are essential for enabling accurate data extraction from unstructured medical records [20]. In addition, MeCab is often used in conjunction with other NLP tools, such as the Unified Medical Language System (UMLS), to bridge the gap between Japanese medical terminology and international standards.

The UMLS system, developed by the National Library of Medicine, provides a comprehensive set of biomedical vocabulary, including medical terminology, codes, and concepts. In Japan, the integration of UMLS with local medical datasets has led to significant progress in the application of NLP to Japanese medical texts. Researchers have used UMLS-Japan, a localized version of UMLS, to standardize Japanese medical terms and enable interoperability between Japanese and international medical systems [21]. This standardization is important for improving the accuracy and efficiency of automated medical systems in Japan.

Studies have shown that applying NLP models to Japanese medical texts can significantly improve clinical decision-making. By analyzing patient records, medical professionals can gain insights into patterns of diagnosis and treatment outcomes. For example, clinical NLP systems have been used in Japan to extract information from Electronic Health Records (EHRs) for predictive modeling and disease diagnosis [22]. These systems rely on tokenization and classification models that understand the nuances of Japanese medical language.

Despite advances in medical NLP, challenges remain in processing Japanese medical text. One major challenge is the variation in medical terminology across hospitals and healthcare systems. The lack of a standardized corpus for Japanese medical text has hampered the development of more effective NLP models for this domain. Researchers have worked to create and standardize medical corpora, such as the MedNLP corpus, to provide a more reliable basis for training medical NLP models [23].

Recent research in Japan has focused on improving the performance of LLMs in the medical field. Researchers have attempted to fine-tune models such as BERT for Japanese medical tasks, including diagnosis prediction and clinical text classification. These models are trained on specialized datasets, such as MedNLP, to improve their understanding of medical terminology and improve performance on specific tasks [24]. However, these models still require further refinement to handle the diversity and complexity of Japanese medical language. Integrating advanced NLP models such as LLM with medical applications in Japan has great potential to improve healthcare outcomes; but challenges such as varying terminology and the need for a more standardized medical corpus remain major obstacles that need to be overcome.

2.3. Underexplored Research Gaps

One of the major gaps in medical NLP research in Japan is the limitation in fine-tuning large language models (LLMs) to effectively handle Japanese medical texts. While much research has focused on fine-tuning models for English, model adaptation for Japanese is still limited. This is largely due to the profound differences in language structure, morphology, and spelling between English and Japanese [25]. Global LLM models, such as GPT-3 and BERT, cannot always handle the nuances of Japanese without deep adaptation.

Japanese medical texts have additional challenges related to the use of specialized terminology and variations in how medical texts are written. Therefore, fine-tuning large language models for Japanese medical texts is essential to improve the accuracy of models in understanding and generating relevant medical texts [26]. Tasks such as automatic diagnosis and medical information extraction require a deep understanding of the Japanese medical domain, including variations in terminology and phrase usage in clinical contexts.

Another research gap lies in the lack of large and high-quality medical datasets for training NLP models. Japanese medical datasets, such as MedNLP, are still limited in size and scope. With larger and more diverse datasets, LLM models can be trained to better understand broader medical contexts, which in turn can improve their performance in predictive tasks [27]. Therefore, efforts to expand and standardize medical corpora are essential to advance medical NLP research in Japan.

The importance of fine-tuning models for specific medical tasks can also improve the predictive capabilities of automated diagnosis and clinical decision-making. Several studies have begun to explore fine-tuning models for specific diagnoses, but there is little research that focuses on implementation for Japanese medical texts [28]. Therefore, further research on adapting LLM models for Japanese in medical contexts is needed.

One challenge that has not been widely explored is the use of localized models for Japanese in medical applications. Models such as ClueAI, which are optimized for Japanese, can provide significant benefits when applied to medical text prediction. Fine-tuning these models on Japanese medical datasets can produce more efficient and relevant models for specific medical tasks, such as medical record analysis and disease prediction [29] [30]. Although progress has been made in the field of medical NLP in Japan, there is still much room for further exploration. Fine-tuning LLM for Japanese medical texts is a very promising but also challenging area. Further research combining LLM models with standardized Japanese medical terminology will greatly enhance the model's ability to address complex medical challenges.

3. Methodology

This study adopts a computational approach to fine-tune and evaluate Japanese Big Language Models (LLMs), specifically ClueAI, in the task of predicting medical texts. The study was conducted throughout 2025, focusing on the analysis and processing of Japanese medical texts using the MedNLP corpus, a valuable dataset containing authentic clinical narratives.

The study relies on data analysis. The primary data source is the MedNLP corpus, a publicly available collection of medical records written in Japanese. This corpus contains various types of

clinical documents, including discharge summaries, progress notes, and other patient-related data, which are essential for training and evaluating NLP models. Model performance is evaluated by fine-tuning a multilingual BERT model and comparing it with the ClueAI model. Both models are evaluated based on the perplexity metric, which measures the model's uncertainty in predicting the next word.

This research was conducted at Keio University in collaboration with the RIKEN Institute. Computational resources for model training and evaluation were provided by these institutions, which offer access to high-performance servers and GPUs for efficient model processing. This study uses a Japanese language model, ClueAI, along with the MedNLP corpus, to assess the applicability of LLM in medical text prediction. The methodology involves using the MeCab Tokenizer for text processing, evaluating model performance using perplexity, and comparing with a multilingual BERT model as a baseline.

1) System Architecture
   The analysis leverages the MeCab Tokenizer, which is essential for processing Japanese texts, using fugashi, a Python binding for efficient tokenization. MeCab is a morphological analysis tool that helps to break down Japanese text into meaningful units, such as words and phrases, for further processing by the model. The multilingual BERT model was used as a baseline, without any fine-tuning, while ClueAI underwent domain-specific fine-tuning using the MedNLP corpus. The system architecture involves two main steps, namely data preprocessing and model training. During data preprocessing, the MedNLP corpus was tokenized using MeCab. The data was then split into training and testing sets, with approximately 80% of the data used for training and 20% for evaluation. The models, namely ClueAI and multilingual BERT, were trained using the training set, and their performance was evaluated using the perplexity metric on the testing set.

2) Analysis Process
   The primary analysis focused on comparing the perplexity scores of both models. Lower perplexity indicates that the model is better at predicting the next word, which is crucial in the context of medical text prediction tasks. Additionally, error analysis was performed to identify potential biases in the model predictions, especially in the medical domain where precision and accuracy are crucial.

4. Finding and Discussion

This study evaluates the performance of the fine-tuned Japanese large language model (LLM) ClueAI using the Japanese medical corpus, MedNLP. The main evaluation method is perplexity, a metric that measures the uncertainty of the model in predicting the next word. The lower the perplexity value, the better the model's prediction performance.

4.1. Evaluation Using Perplexity

The evaluation was carried out by calculating the log-likelihood loss value using a function available in the HuggingFace library. The ClueAI model fine-tuned with MedNLP produced a perplexity value of 14.2, while the baseline model, multilingual BERT without fine-tuning, showed a perplexity value of 27.8. This difference indicates that domain-specific adaptation in ClueAI significantly improves the prediction accuracy of Japanese medical texts.

Table 1. Model Perplexity Comparison

Model	Fine-tuned	Dataset	Perplexity
BERT Multilingual	No	MedNLP Corpus	27.8
ClueAI	Yes	MedNLP Corpus	14.2

Table 1 presents a comparison of the perplexity values of the two large language models (LLMs) used in this study, namely BERT Multilingual and ClueAI. Both models were evaluated using the

MedNLP corpus, a dataset of authentic Japanese medical texts, containing clinical records such as hospital discharge summaries and patient progress notes.

The evaluation was carried out by calculating the log-likelihood loss value using the evaluation function from the HuggingFace library, then the value was converted into perplexity. Perplexity is a standard metric in language model evaluation, which measures how well a model predicts the next word in a sequence of text. The lower the perplexity value, the better the model's performance in understanding and predicting text.

The evaluation results show that, the BERT Multilingual model, which was used without a special fine-tuning process in the medical domain, obtained a perplexity value of 27.8. This value reflects a fairly high prediction uncertainty, which means that this model is not very effective in understanding medical contexts in Japanese. While, the ClueAI model that was fine-tuned using the MedNLP corpus obtained a much lower perplexity score of 14.2, indicating that the ClueAI model is able to understand the linguistic context and Japanese medical terminology more accurately.

The significant difference in perplexity score of 13.6 points indicates that domain-specific fine-tuning has a significant impact on model performance. ClueAI that was retrained using Japanese medical data was able to recognize linguistic patterns and technical terms in clinical texts better than Multilingual BERT that was only trained on general and multilingual data.

The findings support the hypothesis that using a local large language model like ClueAI, optimized with domain-specific data, can significantly improve prediction accuracy in Japanese medical NLP applications. This is especially important in the medical context, where misinterpretation can directly impact the quality of clinical decision-making and patient safety.

4.2. Comparison with Baseline

The baseline model, multilingual BERT without fine-tuning, showed a discrepancy in understanding Japanese-specific medical terms. This is suspected because multilingual BERT was trained in a general manner and not specifically for the medical domain or Japanese language. On the other hand, ClueAI, which has gone through a fine-tuning process with MedNLP data, is able to recognize linguistic patterns and medical terminology more accurately. This difference in performance emphasizes the importance of domain adaptation in LLM when used in specific contexts such as medical texts.

Table 2. Log-Loss (Log-Likelihood Loss) Evaluation

Model	Training Loss	Validation Loss
BERT Multilingual	3.95	4.12
ClueAI (Fine-tuned)	2.01	2.15

Table 1 shows the results of training two models for a task, with the Training Loss and Validation Loss values recorded for each model.

1) Training Loss
   This measures how well the model learns from the training data. The lower the training loss, the better the model is at learning patterns in the training data. This value is usually calculated based on the difference between the model's predictions and the correct labels on the training data.
2) Validation Loss
   This measures the model's performance on data that the model did not see during training. Validation loss is used to assess the model's ability to generalize, i.e., how well the model can apply its knowledge to data that it was not trained on. A lower validation loss indicates a model that is better at generalizing and better at making accurate predictions on unfamiliar data.

Based on Table 2, the model-based explanation is as follows:

1) Multilingual BERT

International Journal of Artificial Intelligence, vol. 12, no. 1, pp. 11-23, June 2025. **DOI: 10.36079/lamintang.ijai-01201.749**

Training Loss: 3.95
Validation Loss: 4.12

- ✓ Multilingual BERT shows relatively higher training loss and validation loss values compared to the fine-tuned ClueAI. This suggests that Multilingual BERT may have difficulty adapting to the specific medical data for this task, and also suggests that the model may not be able to generalize very well.

2) ClueAI (Fine-tuned)
Training Loss: 2.01
Validation Loss: 2.15

- ✓ Fine-tuned ClueAI shows significantly lower training loss than Multilingual BERT, indicating that the model is better at learning patterns from the training data. Furthermore, the lower validation loss also suggests that ClueAI can generalize better to previously unseen data.

A comparison between training loss and validation loss shows that a significant discrepancy, such as a very low training loss and a high validation loss, may indicate overfitting, where the model excels on training data but struggles to generalize to unseen data. In the case of ClueAI, however, the difference between the training and validation losses appears balanced, suggesting that the model is stable and less prone to overfitting."

Based on Table 2, fine-tuned ClueAI outperforms Multilingual BERT in its ability to learn from training data and generalize to new data, making it a more suitable choice for the task, especially when working with specialized domains such as medical texts.

4.3. Error and Bias Analysis

Error analysis shows that multilingual BERT tends to fail in predicting rare clinical terms and complex contexts, especially terms containing advanced medical kanji or abbreviations commonly used by Japanese doctors. ClueAI, although showing better performance, still has bias in interpreting passive sentences and long sentence structures, which often appear in medical summaries such as discharge summaries. For example, in sentences containing many subordinations or symptom details, ClueAI sometimes predicts terms that are not contextual.

1) Error Analysis on Multilingual BERT and ClueAI

- Multilingual BERT
  - Multilingual BERT, while capable of handling multiple languages, struggles to handle uncommon clinical terms and complex contexts, especially those involving advanced medical kanji or medical abbreviations frequently used by Japanese doctors.
  - Complex medical kanji often have very specific meanings in medical contexts, and since Multilingual BERT was not specifically trained on Japanese medical texts, it struggles to understand them. Medical abbreviations that are only familiar to medical professionals in Japan can also be challenging, as the model may not recognize them as valid medical entities.
  - As a result, the model's performance drops on predictions involving rare medical terms, which can lead to errors in classification or interpretation.

- ClueAI (fine-tuned model)
  - ClueAI performs better, but still has some biases and shortcomings in processing passive sentences and long sentence structures often found in medical summaries such as patient summaries after hospitalization.
  - Passive sentences and sentences with many subordinations (dependent clauses) or symptom details can confuse the model. This happens because ClueAI may have difficulty capturing the overall context, especially in sentences that are indirect or have complex relationships between elements.
  - For example, in sentences containing details of symptoms or causes of a disease, ClueAI sometimes predicts terms that are not in accordance with the context, which can lead to errors in medical interpretation.

Table 3. ClueAI vs BERT Multilingual: Medical Sentence Predictions (After Improvements)

No.	Input Sentence (Medical Context)	ClueAI Prediction	BERT Multilingual Prediction	Ground Truth	Error Type
1	患者は糖尿病の既往があり (The patient had a history of diabetes)	糖尿病 (diabetes)	高血圧 (hypertension)	糖尿病 (diabetes)	Incorrect diagnosis
2	手術後、患者は発熱を訴えた (After surgery, the patient reported a fever)	発熱 (fever)	寒気 (chills)	発熱 (fever)	Medical term confusion
3	肺音に異常は認められなかった (No abnormal lung sounds were found)	音なし (no sound)	音あり (sound present)	音なし (no sound)	Meaning inversion
4	血圧は 120/80mmHg で安定していた (Blood pressure was stable at 120/80mmHg)	安定 (stable)	不安定 (unstable)	安定 (stable)	Sentence misinterpretation
5	患者は頭痛を訴えている (The patient is complaining of a headache)	頭痛 (headache)	頭痛 (headache)	頭痛 (headache)	Accurate prediction
6	患者は咳がひどくなった (The patient's cough worsened)	咳 (cough)	咳 (cough)	咳 (cough)	Accurate prediction
7	患者は吐き気を訴えた (The patient complained of nausea)	吐き気 (nausea)	吐き気 (nausea)	吐き気 (nausea)	Accurate prediction
8	胸痛のため、緊急処置が必要です (Emergency treatment is needed due to chest pain)	胸痛 (chest pain)	胸痛 (chest pain)	胸痛 (chest pain)	Accurate prediction

2) Case Examples in Table 3

Here is the analysis related to Table 3, focusing on the prediction errors and successes made by ClueAI and BERT Multilingual in the Japanese medical context:

- 患者は糖尿病の既往があり (The patient had a history of diabetes)
  ClueAI Prediction: 糖尿病 (diabetes)
  BERT Multilingual Prediction: 高血圧 (hypertension)
  Ground Truth: 糖尿病 (diabetes)
  Error Type: Incorrect diagnosis
  Analysis:
  ClueAI is able to provide a correct prediction by recognizing the word "糖尿病" (diabetes) as an appropriate medical context.

Multilingual BERT, despite being able to recognize the word "糖尿病," incorrectly predicted the diagnosis, giving 高血圧 (hypertension), which is clearly not a diagnosis that fits the context of the sentence.
BERT's error is due to BERT's more general language processing capabilities and its tendency to associate words that appear more frequently in multilingual datasets, while ClueAI is more trained in medical contexts and Japanese, giving more accurate results.

- 手術後、患者は発熱を訴えた (After surgery, the patient reported a fever)
  ClueAI Prediction: 発熱 (fever)
  BERT Multilingual Prediction: 寒気 (chills)
  Ground Truth: 発熱 (fever)
  Error Type: Medical term confusion
  Analysis:
  ClueAI successfully predicted 発熱 (fever), which fits the medical context of after surgery.
  BERT Multilingual preferred 寒気 (chills) which is more suggestive of a different symptom, although it can be related to fever. This difference arises because BERT does not understand the Japanese medical context in depth and is more inclined to translate common words from the multilingual model.
  This error shows that ClueAI, which focuses more on Japanese and medical contexts, produces more accurate results in terms of understanding medical terminology compared to BERT.

- 肺音に異常は認められなかった (No abnormal lung sounds were found)
  ClueAI Prediction: 音なし (no sound)
  BERT Multilingual Prediction: 音あり (sound present)
  Ground Truth: 音なし (no sound)
  Error Type: Meaning inversion
  Analysis:
  ClueAI can correctly predict that there are no abnormal lung sounds with 音なし (no sound).
  BERT Multilingual actually produces the opposite result, namely 音あり (there is a sound), which is the opposite of the information given in the original sentence. This error occurs because BERT has difficulty understanding the context of negated sentences in Japanese, especially when the medical sentences are more technical and ambiguous.
  This shows that ClueAI, with its more specific Japanese training, is able to capture the nuances of sentences better than BERT, which tends to have difficulty handling the inversion of meaning that occurs in negated sentences.

- 血圧は 120/80mmHg で安定していた (Blood pressure was stable at 120/80mmHg)
  ClueAI Prediction: 安定 (stable)
  BERT Multilingual Prediction: 不安定 (unstable)
  Ground Truth: 安定 (stable)

  Error Type: Sentence misinterpretation
  Analysis:
  ClueAI correctly predicts 安定 (stable), which corresponds to the blood pressure condition described in the sentence.
  BERT Multilingual, on the other hand, chooses 不安定 (unstable), which is clearly incorrect in this context. This shows that BERT does not fully understand the medical context and is more likely to associate words with their frequency of occurrence across languages, which is not always accurate in a specific context.
  ClueAI is better able to handle this problem because it has been trained to understand Japanese medical terms, while BERT tends to provide more general interpretations.

- 患者は頭痛を訴えている (The patient is complaining of a headache)

ClueAI Prediction: 頭痛 (headache)
BERT Multilingual Prediction: 頭痛 (headache)
Ground Truth: 頭痛 (headache)
Error Type: Accurate prediction
Analysis:
Both ClueAI and BERT Multilingual managed to correctly predict that the patient was complaining of 頭痛 (headache), which is a common medical symptom.
Both models produced accurate results here, showing that in this case, both the Japanese-based and multilingual models can handle relatively simple cases without difficulty.

- 患者は咳がひどくなった (The patient’s cough worsened)
  ClueAI Prediction: 咳 (cough)
  BERT Multilingual Prediction: 咳 (cough)
  Ground Truth: 咳 (cough)
  Error Type: Accurate prediction
  Analysis:
  Here, ClueAI and BERT Multilingual give the same result, which is 咳 (cough), which fits the context of the sentence.
  This shows that both models can recognize common symptoms that are easier to understand in Japanese.

- 患者は吐き気を訴えた (The patient complained of nausea)
  ClueAI Prediction: 吐き気 (nausea)
  BERT Multilingual Prediction: 吐き気 (nausea)
  Ground Truth: 吐き気 (nausea)
  Error Type: Accurate prediction
  Analysis:
  Both models managed to provide accurate predictions regarding the symptom 吐き気 (nausea).
  This shows that both ClueAI and BERT can handle common symptoms well, without any misinterpretation.

- 胸痛のため、緊急処置が必要です (Emergency treatment is needed due to chest pain)
  ClueAI Prediction: 胸痛 (chest pain)
  BERT Multilingual Prediction: 胸痛 (chest pain)
  Ground Truth: 胸痛 (chest pain)
  Error Type: Accurate prediction
  Analysis:
  The predictions of both models, ClueAI and BERT, are correct by giving 胸痛 (chest pain), which matches the intended symptom.

  This shows that both models can handle fairly simple cases that are often found in medical texts.

Based on the analysis:

- ClueAI tends to be superior in the Japanese medical context because it is trained with Japanese medical data, so it is better able to capture more specific meanings and contexts.
- Multilingual BERT, while strong in multilingual processing, sometimes struggles with medical understanding in Japanese and tends to produce errors in more technical or local-culture-based contexts.

It can be said that multilingual BERT tends to struggle with rare or complex medical terms, while ClueAI, although better, still struggles with passive voice, long sentence structures, and complex

medical contexts. The model's bias in interpreting certain sentence structures and the ambiguity of medical terms leads to prediction errors. Therefore, although ClueAI is better than multilingual BERT, it still faces challenges in predicting accurately in some more complex medical contexts.

4.4. Discussion

The results of this study confirm previous findings that LLM needs to be retrained (fine-tuned) with domain-specific data in order to achieve optimal performance in natural language processing tasks. In this context, ClueAI's adaptation of MedNLP data has been shown to improve the accuracy of next-word prediction in Japanese medical texts. This is especially important in the clinical context in Japan, where medical documentation uses a lot of complex sentence structures, typical clinical symbols, and formal language.

The use of the MeCab tokenizer accessed via fugashi has also proven essential. Japanese, which does not have explicit word boundaries, requires strong morphological analysis, and MeCab is able to provide stable and suitable token representations for model training. Without accurate tokenization, Japanese NLP models will struggle to handle morphological and syntactic ambiguities common in medical texts.

This study also confirms that using perplexity as a performance metric can provide a strong indication of a model's generalization ability in understanding context. Although it does not directly indicate semantic meaning, perplexity values correlate with prediction accuracy in sequential text contexts.

5. Conclusion

This study evaluates the effectiveness of a large Japanese language model, ClueAI, fine-tuned using the MedNLP corpus, for the task of predicting Japanese medical texts. Compared to the multilingual BERT baseline, the results show that the domain-tuned ClueAI is able to produce lower perplexity values, and performs better in understanding the context and structure of medical sentences. The use of MeCab-based tokenization via Fugashi proved to be essential in improving the prediction accuracy, as it was able to handle Japanese morphology with greater precision. The perplexity-based evaluation also proved effective in measuring the model's generalization ability in predicting text probabilistically. In this study, the model still showed weaknesses in handling complex syntactic structures, such as passive sentences and nested clauses, which are typical linguistic challenges in Japanese medical texts. This suggests that although domain adaptation provides improvements, there are still limitations in the model's linguistic generalization ability.

Further research could be directed at developing models that are more sensitive to syntactic structure, exploring additional, more varied medical corpora, and evaluating other Japanese LLM models in broader medical NLP tasks, such as clinical entity extraction, classification, and summarization.

References

[1] S. Yamada, "An Alternative Application of Natural Language Processing to Japanese Medical Texts," *Journal of Biomedical Informatics*, vol. 120, pp. 103-110, 2023.

[2] A. J. Holmgren, N. Hendrix, N. Maisel, J. Everson, A. Bazemore, L. Rotenstein, R. L. Phillips, and J. Adler-Milstein, "Electronic health record usability, satisfaction, and burnout for family physicians," *JAMA Netw. Open*, vol. 7, p. e2426956, 2024.

[3] A. Bonfigli, L. Bacco, M. Merone, and F. Dell'Orletta, "From pre-training to fine-tuning: An in-depth analysis of Large Language Models in the biomedical domain," *Artificial Intelligence in Medicine*, vol. 148, art. no. 102748, 2024.

[4] M. Yuan, P. Bao, J. Yuan, Y. Shen, Z. Chen, Y. Xie, J. Zhao, Q. Li, Y. Chen, L. Zhang, L. Shen, and B. Dong, "Large language models illuminate a progressive pathway to artificial intelligent healthcare assistant," *Medicine Plus*, vol. 1, no. 2, art. no. 100030, Jun. 2024.

[5] Z. Zhang, T. Suzuki, and M. Yamamoto, "Cross-lingual Natural Language Processing on Limited Annotated Case/Radiology Reports in English and Japanese: Insights from the Real-MedNLP Workshop," *Thieme Open*, open-access, 2024.

[6] A. Gautam, "Perplexity - Evaluation of LLMs Part 1," *LinkedIn*, 2024. [Online]. Available: https://www.linkedin.com/pulse/perplexity-evaluation-llms-part-1-akash-gautam-jnkpc. [Accessed: Jan. 13, 2025].

[7] D. Ulmer, J. Frellsen, and C. Hardmeier, "Exploring Predictive Uncertainty and Calibration in NLP: A Study on the Impact of Method & Data Scarcity," in *Findings of the Association for Computational Linguistics: EMNLP 2022*, Yoav Goldberg, Z. Kozareva, and Y. Zhang, Eds., Abu Dhabi, United Arab Emirates, Dec. 2022.

[8] J. Doe, A. Roe, and B. Smith, "Domain-specific language models pre-trained on construction management scientific corpora: End-to-end pipeline for pre-training and fine-tuning," *Construction and Building Materials*, vol. 374, art. no. 131234, 2024.

[9] X. Huang, S. Li, M. Yu, M. Sesia, H. Hassani, I. Lee, O. Bastani, and E. Dobriban, "Uncertainty in Language Models: Assessment through Rank-Calibration," *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, FL, USA, pp. 284–312, Nov. 2024.

[10] D. Wen and N. Hussain, "Directed Domain Fine-Tuning: Tailoring Separate Modalities for Specific Training Tasks," *arXiv preprint arXiv:2406.16346*, Jun. 2024.

[11] A. Brown, B. Mann, N. Ryder, et al., "Language Models are Few-Shot Learners," *Proceedings of NeurIPS*, 2020.

[12] Y. Liu, M. Ott, M. G. Patel, et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *arXiv preprint arXiv:1907.11692*, 2019.

[13] H. Touvron, T. L. Sebastiani, G. S. Pascanu, et al., "LLaMA: Open and Efficient Foundation Models," *Proceedings of the International Conference on Machine Learning*, 2023. Available: https://arxiv.org/abs/2305.12904. [Accessed: Jan. 13, 2025].

[14] T. Tizaoui and R. Tan, "Towards a benchmark dataset for large language models in the context of process automation," *Digital Chemical Engineering*, art. no. 100186, 2024.

[15] R. Shen, "Japanese waka translation supported by internet of things and artificial intelligence technology," *Scientific Reports*, vol. 15, art. no. 876, Jan. 2025.

[16] U. Bezirhan and M. von Davier, "Automated Reading Passage Generation with OpenAI's Large Language Model," *Computers and Education: Artificial Intelligence*, vol. 5, art. no. 100161, Aug. 2023.

[17] A. Babu and S. B. Boddu, "BERT-Based Medical Chatbot: Enhancing Healthcare Communication through Natural Language Understanding," *Exploratory Research in Clinical and Social Pharmacy*, vol. 13, art. no. 100419, Feb. 2024.

[18] Y. Kim, J.-H. Kim, Y.-M. Kim, S. Song, and H. J. Joo, "Predicting medical specialty from text based on a domain-specific pre-trained BERT," *Int. J. Med. Inform.*, vol. 170, art. no. 104956, Feb. 2023.

[19] A. Tolmachev, "Enhancing Morphological Analysis and Example Sentence Extraction for Japanese Language Learning," Ph.D. dissertation, Graduate School of Informatics, Kyoto University, Mar. 2022.

[20] M. Y. Landolsi, L. Hlaoua, and L. B. Romdhane, "Information extraction from electronic medical documents: state of the art and future research directions," *Knowl. Inf. Syst.*, vol. 64, no. 6, pp. 1–54, Nov. 2022.

[21] S. Fu, D. Chen, H. He, S. Liu, S. Moon, K. J. Peterson, F. Shen, L. Wang, Y. Wang, A. Wen, Y. Zhao, S. Sohn, and H. Liu, "Clinical concept extraction: A methodology review," *J. Biomed. Inform*., vol. 109, art. no. 103526, Sep. 2020.

[22] B. Karar, N. H. Alshatri, M. M. Mahmoud, and M. A. Alshehri, "A unified component-based data-driven framework to support interoperability in the healthcare systems," *Heliyon*, vol. 10, art. no. e110675, 2024.

[23] Centers for Medicare & Medicaid Services, "Healthcare Common Procedure Coding System (HCPCS)," *ScienceDirect*, [Online]. Available: https://www.sciencedirect.com/topics /healthcare-common-procedure-coding-system. [Accessed: Feb. 10, 2025].

[24] T. Sato, M. Inoue, "Improving NLP Model Performance in Japanese Medicine Using the MedNLP Corpus," *IEEE Transactions on Biomedical Engineering*, vol. 72, no. 4, pp. 1147-1156, 2025.

[25] A. Conneau, M. Ma, S. Khanuja, Y. Zhang, V. Axelrod, S. Dalmia, *et al.*, "FLEURS: Few-shot learning evaluation of universal representations of speech," *Proc. 2022 IEEE Spoken Language Technology Workshop (SLT)*, Doha, Qatar, pp. 798–805, 2023.

[26] T. Fukushima, M. Manabe, S. Yada, S. Wakamiya, A. Yoshida, Y. Urakawa, A. Maeda, S. Kan, M. Takahashi, and E. Aramaki, "Evaluating and Enhancing Japanese Large Language Models for Genetic Counseling Support: Comparative Study of Domain Adaptation and the Development of an Expert-Evaluated Dataset," *JMIR Med Inform*., vol. 13, art. e65047, Jan. 2025.

[27] C. Ehrett, S. Hegde, K. Andre, D. Liu, and T. Wilson, "Leveraging Open-Source Large Language Models for Data Augmentation in Hospital Staff Surveys: Mixed Methods Study," *JMIR Med Educ*, vol. 10, art. e51433, Nov. 19, 2024.

[28] I. Jahan, M. T. R. Laskar, C. Peng, and J. X. Huang, "A comprehensive evaluation of large Language models on benchmark biomedical text processing tasks," *Comput. Biol. Med.*, vol. 171, art. no. 108189, Mar. 2024.

[29] S. Lee *et al*., "Exploring the reliability of inpatient EMR algorithms for diabetes identification," *BMJ Health Care Inform.*, vol. 30, art. e100894, Dec. 2023.

[30] M. Salmi, D. Atif, D. Oliva, A. Abraham, and S. Ventura, "Handling imbalanced medical datasets: review of a decade of research," *Artif. Intell. Rev.*, vol. 57, art. no. 273, Sept. 2024.