ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

ORIGINAL RESEARCH ARTICLE

EpilepsyLLM: Fine-tuning large language models for Japanese epilepsy knowledge representation

Xuyang Zhao[1,2,3], Qibin Zhao[4], and Toshihisa Tanaka[5]*

[1]Medical Science Data-driven Mathematics Team, RIKEN Center for Interdisciplinary Theoretical and Mathematical Sciences, Yokohama, Kanagawa, Japan

[2]Medical Data Mathematical Reasoning Special Team, RIKEN Center for Integrative Medical Sciences, Yokohama, Kanagawa, Japan

[3]Department of Artificial Intelligence Medicine, Chiba University, Chiba, Japan

[4]Tensor Learning Team, RIKEN Center for Advanced Intelligence Project, Chuo, Tokyo, Japan

[5]Department of Electrical Engineering and Computer Science, Tokyo University of Agriculture and Technology, Koganei, Tokyo, Japan

***Corresponding author:**
Toshihisa Tanaka
(tanakat@cc.tuat.ac.jp)

Citation: Zhao X, Zhao Q, Tanaka T. EpilepsyLLM: Fine-tuning large language models for Japanese epilepsy knowledge representation. *Artif Intell Health*. 2026;3(1):104-115. doi: 10.36922/AIH025180042

Received: May 3, 2025

Revised: August 5, 2025

Accepted: August 14, 2025

Published online: September 8, 2025

Abstract

With massive training data and sufficient computing resources, large language models (LLMs) have demonstrated impressive capabilities. These models can rapidly respond to questions in almost all domains and are capable of retrieving, synthesizing, and summarizing information. The capabilities demonstrated by LLMs can enhance our livelihood and foster innovation. Nonetheless, in some professional domains, the focus is not only on response speed but also on higher requirements for response reliability. For example, in the medical domain, the reliability of information provided by the model poses a great risk to subsequent diagnosis and treatment, especially when the language is not English. In specific domains, domain-specific knowledge can be used to refine pre-trained LLMs to improve their performance in specific tasks. In this study, we aimed to build an LLM for epilepsy, called EpilepsyLLM. We constructed an epilepsy knowledge dataset in Japanese for LLM fine-tuning, and the dataset contained basic information on epilepsy, common treatment methods and drugs, and important notes on patients' lives. Using the constructed dataset, we refined several different pre-trained models with supervised learning. In the evaluation process, we applied multiple metrics to measure the reliability of the LLMs' output. The experimental results highlighted that the fine-tuned EpilepsyLLM can provide more reliable and specialized epilepsy responses.

Keywords: Epilepsy; Large language models; Domain-specific; Fine-tuning

1. Introduction

In recent years, large language models (LLMs) have demonstrated remarkable advances across a broad range of natural language processing (NLP) tasks. Their ability to understand instructions, generate human-like responses, and generalize across domains has significantly transformed the landscape of artificial intelligence (AI). These models have consistently set new benchmarks in widely recognized NLP evaluation tasks, showcasing their superior reasoning, comprehension, and generative capabilities. A majority of high-performing LLMs are built upon the transformer architecture.[1] This design enables efficient parallel processing and the capture of long-range dependencies in text. The success of LLMs is attributed not only to architectural innovations but also to the availability of massive training corpora and extensive computational resources, enabling these models to scale to hundreds of billions of parameters. One of the most influential series in this domain is OpenAI's Generative Pre-trained Transformer (GPT) line, which includes GPT-1,[2] a foundational model that explored unsupervised learning on large-scale corpora; GPT-2,[3] known for its fluency and coherence; and GPT-3,[4] which introduced few-shot and zero-shot learning capabilities with its 175 billion parameters. InstructGPT[5] further refined the inherent capabilities by incorporating reinforcement learning from human feedback, allowing it to align better with human instructions. The most recent iteration, GPT-4,[6] pushes the boundaries further with enhanced multimodal capabilities and improved alignment with human preferences. However, the architectural details, training datasets, and optimization strategies for these models remain proprietary, limiting transparency and reproducibility in the academic community. In contrast, Meta's Large Language Model Meta AI (LLaMA)[7] adopts a more open approach. Released specifically for research purposes, the LLaMA family includes models with 7B, 13B, 30B, and 65B parameters. Despite having fewer parameters than some commercial models, LLaMA exhibits strong performance across standard benchmarks. For example, LLaMA-13B surpasses GPT-3 on several tasks, while being significantly more parameter-efficient. The largest model, LLaMA-65B, achieves performance on par with other models, such as Chinchilla[8] and PaLM-540B,[9] both of which are far larger in size. Advancements in LLMs demonstrated that with efficient training, massive amounts of data can yield higher performance. The growing availability of LLMs has opened new possibilities for real-world applications, including education, healthcare, creative writing, and human–computer interaction. The rapid progress and accessibility of LLMs continue to inspire innovations, suggesting a transformative impact on society and daily life.

LLMs have found increasingly diverse applications in the medical field, showcasing their ability to assist in a wide range of healthcare-related tasks. Among the most prominent applications are medical licensing examination,[10-12] diagnostic support,[13,14] patient communication,[15,16] and medical education,[17-19] and their impact extends well beyond these areas. However, alongside these promising developments, the growing reliance on LLMs also introduces a range of risks and ethical concerns that must be addressed.[20-22] One major issue is the risk of misinformation, as LLMs can generate plausible-sounding but factually incorrect responses, which could lead to misdiagnosis or inappropriate clinical decisions if not properly supervised. Furthermore, the black-box nature of most models makes it difficult to understand how conclusions are reached, raising issues of accountability and trust in clinical environments.

To further advance the performance and accessibility of LLMs, Stanford Alpaca[23,24] was introduced, demonstrating an efficient, scalable, and low-cost method for fine-tuning LLMs to follow instructions more effectively. The Alpaca is built upon the LLaMA-7B model, and through a carefully curated fine-tuning process, it achieved performance that is qualitatively comparable to GPT-3.5 (text-davinci-003). In addition, the Alpaca collects a small dataset of 175 human-written instruction-output pairs. These examples represent a diverse set of tasks designed to probe a model's ability to understand and carry out various instructions, ranging from summarization and translation to reasoning and creative writing. Rather than relying on massive manual annotation, Alpaca uses GPT-3.5 (text-davinci-003) to automatically generate additional instruction-following examples, scaling the dataset to 52,000 unique prompts and responses.

LLMs have demonstrated remarkable capabilities across a wide range of general NLP tasks, achieving state-of-the-art performance in areas such as question answering, summarization, translation, and dialogue generation. However, when these models are applied to highly specialized domains, such as medicine, law, and finance, their performance often diminishes. The reduced accuracy and reliability stem from the fact that general-purpose LLMs are typically trained on broad, heterogeneous corpora that lack the depth and nuance required for domain-specific expertise. To address this limitation, one effective method is domain-specific fine-tuning, where a general LLM is further trained or adapted using targeted corpora relevant to a particular professional field. In the medical domain, this approach has yielded significant improvements in performance across various medical NLP benchmarks and real-world clinical tasks. In the medical field, some of these medical-specific LLMs include PubMedBERT,[25] BioLinkBert,[26] BioMedLM,[27] BioGPT,[28] Med-PlaM,[29] ClinicalGPT,[30] PMC-LLaMA,[31] and ChatDoctor.[32] In addition, these LLMs typically incorporate various types of medical knowledge, including patient-physician dialogue transcripts, PubMed abstracts, full-text articles from PubMed Central (PMC), and deidentified electronic health records, to further enhance model understanding in clinical settings. These models demonstrate the effectiveness of incorporating structured and unstructured domain-specific data into the training pipeline. With the help of medical knowledge, fine-tuned LLMs achieve better performance in medical tasks. For instance, in the United States Medical Licensing Examination (USMLE), MedPaLM 2 achieved the highest score of 86.5.

While existing medical LLMs have demonstrated impressive capabilities in handling a broad range of medical tasks, their success has largely been confined to English and generalized medical knowledge. These models focus on English-based medical reasoning, question answering, and literature comprehension, and they often fall short when applied to non-English languages or specialized subdomains of medicine. In this study, we present the EpilepsyLLM, a domain-specific LLM designed to focus exclusively on epilepsy and operate primarily in the Japanese language. Epilepsy is one of the most prevalent neurological disorders worldwide,[33] affecting millions of individuals across age groups. It manifests through a variety of seizure types, including tonic rigidity, myoclonic jerks, and atonic seizures, each of which can significantly impair a patient's quality of life.[33-36] The burden of epilepsy is particularly acute in pediatric patients, where the condition may hinder cognitive development, behavioral stability, and social integration.[37] Treatment for epilepsy generally begins with medication, which helps control seizures in many patients. However, a substantial subset of individuals suffer from refractory (drug-resistant) epilepsy, for which surgical intervention, such as resective surgery or neurostimulation, is considered. Despite medical or surgical treatment, epilepsy patients often live with numerous daily life restrictions, including avoiding seizure triggers, adhering to medication schedules, navigating driving limitations, and managing social stigma.[38] Given these challenges, EpilepsyLLM aims to serve as a specialized medical assistant capable of understanding, generating, and interpreting Japanese texts related to epilepsy. Its applications may span from assisting clinicians with diagnosis and treatment planning to supporting patients and caregivers in understanding disease management and improving communication. By concentrating on a single neurological disease and embracing linguistic diversity, EpilepsyLLM offers a promising direction in the field of medical AI, highlighting the potential for more precise, localized, and equitable healthcare solutions.

The application of LLMs to the medical field requires a high level of professionalism and domain accuracy. Unlike general-purpose use cases, medical applications demand high precision, factual correctness, and a nuanced understanding of clinical terminology and practices. Errors in model outputs can have serious consequences, which makes the incorporation of expert-level medical knowledge essential for safe and effective deployment. To address these needs in the context of epilepsy, we constructed a fine-tuning dataset composed of high-quality, domain-specific knowledge collected from publicly available Japanese resources on the internet. These sources are reliable epilepsy-focused content. The collected data were transformed into instruction-following demonstrations. As a base model, we employed two pre-trained models, LLaMA[7] and LLM-jp[39] (a Japanese language foundation model known for its linguistic alignment with Japanese text). LLM-jp provided a strong foundation for our study, given that its vocabulary, tokenizer, and pre-training corpus are optimized for the Japanese language, making it highly suitable for our targeted application. Our experiments demonstrated that the LLMs fine-tuned with our curated epilepsy-specific dataset significantly outperformed other baseline models. EpilepsyLLM presents more professional and reliable answers when faced with epilepsy knowledge. The experimental results also confirmed that by using more domain-specific knowledge to fine-tune the LLMs, the performance of the model in the particular domain can be significantly enhanced.

The proposed EpilepsyLLM holds potential for application across several clinically relevant areas. Clinical decision support is one promising area where the model could assist healthcare providers by synthesizing guidelines, summarizing treatment options, or generating initial responses to structured clinical queries. In the context of patient education, the system could help produce tailored, accessible explanations of medical information—particularly valuable for underserved or linguistically diverse populations. In addition, the model could serve in medical scribe and documentation assistance, helping clinicians to structure or translate free-text notes into more formal documentation based on high-level prompts or queries. Despite the model's potential, some issues should also be noted in its prospective clinical applications, such as misinformation, overreliance, and the lack of transparency, thereby warranting further research in these areas.

2. Methods

In this study, we leveraged domain-specific knowledge related to epilepsy to fine-tune LLMs to enhance their performance in epilepsy-focused medical applications. Moreover, we focused on optimizing the model's performance in Japanese, ensuring that the resulting language model can offer high-quality, linguistically and culturally tailored support for Japanese-speaking medical professionals, researchers, and patients. The overview of EpilepsyLLM is displayed in Figure 1.

2.1. Epilepsy knowledge dataset

To construct a high-quality dataset suitable for fine-tuning a domain-specific model, we systematically collected

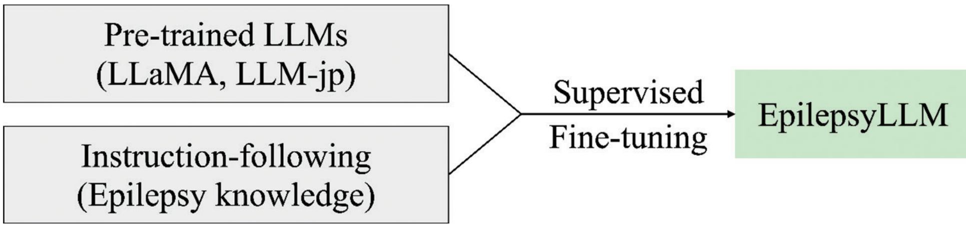

Figure 1. The overview of EpilepsyLLM
Abbreviations: LLMs: Large language models; LLaMA: Large Language Model Meta Artificial Intelligence.

relevant information from reputable public sources focused on epilepsy. These sources included official websites operated by medical associations, hospitals, and pharmaceutical companies. We aimed to ensure that the curated dataset is both reliable and medically accurate. This dataset served as a foundational resource for fine-tuning models tailored to epilepsy, enhancing their ability to understand, analyze, and generate clinically meaningful text. The specific websites used in the dataset collection process are listed below:

(i) Japan Epilepsy Association[1]: The association conducts research and nationwide campaigns aimed at promoting social understanding of epilepsy, providing social support for people suffering from epilepsy, and improving epilepsy policies.
(ii) Epilepsy Information Center[2]: The website belongs to the National Epilepsy Center, National Hospital Organization (NHO), Shizuoka Institute of Epilepsy and Neurological Disorders, which is the largest epilepsy center in Japan in terms of both patient volume and number of specialized clinicians.
(iii) Tenkan Net[3]: The website, developed by Alfresa Pharma Corporation, collects basic information about epilepsy, including basic diagnosis and treatment plans, routine examinations, commonly used drugs and their side effects, and lifestyle precautions.

After data collection, we conducted a systematic and uniform preprocessing and curation process to ensure data quality and ethical compliance. The curation pipeline involved several key steps. First, we standardized the formatting by normalizing whitespace, punctuation, and special characters to ensure consistency across the dataset. Next, we applied language-specific cleaning procedures to correct encoding issues and remove non-textual elements, such as HTML tags or formatting artifacts. To address privacy concerns and adhere to ethical research standards, we implemented a filtering step to identify and remove any personally identifiable information. This included names, physical addresses, phone numbers, email addresses, and other sensitive information that could potentially identify the individuals. The resulting dataset maintained the integrity of the original content while safeguarding user privacy and improving overall data usability for downstream NLP tasks.

Data collected from the three independent websites were divided into training and testing datasets after pre-processing. Content from the Japan Epilepsy Association and Epilepsy Information Center was used to construct the training dataset. For the testing dataset, we curated 24 representative instruction–response pairs exclusively from Tenkan Net, which were not included in the training dataset. This separation ensured an independent evaluation and allowed us to better assess the model's ability to generalize to unseen material within the same domain.

The dataset from Alpaca[23] was constructed with a focus on instruction-following demonstrations, where each data sample consisted of a prompt and a corresponding response designed to guide model behavior. Some representative examples from the dataset are displayed in Table 1. For fine-tuning the model, more than 200 carefully curated samples were utilized. These samples covered a diverse range of instructions and domains to ensure that the model could effectively generalize to various tasks. In addition, a separate subset comprising 24 distinct samples was selected and reserved exclusively for testing. This test set was used to evaluate the model's performance on unseen instruction–response pairs, providing a measure of its generalization capabilities after fine-tuning.

2.2. Pre-trained models

2.2.1. LLaMA

The LLaMA[7] family of models, introduced by Meta AI, represents a series of foundational LLMs designed to advance research in NLP tasks. Unlike many commercially released models, LLaMA models are explicitly intended for academic and research purposes, and these models are made accessible to researchers under a non-commercial license. LLaMA models are trained on publicly available

1 https://www.jea-net.jp/
2 https://shizuokamind.hosp.go.jp/epilepsy-info/
3 https://www.alfresa-pharma.co.jp/general/tenkan/faq/index.html

Table 1. Examples of an epilepsy dataset

Question	Answer
てんかんはどんな病気ですか？(What is epilepsy?)	てんかん発作は大脳神経細胞の過剰な亢奮（てんかん発射）によって生じます。てんかん発作を繰り返し、慢性に経過する大脳の病気がてんかんです。ただ1 回だけの発作や高熱のときだけに起こる発作はてんかんとは区別されます。また、脳炎の急性期や頭部外傷の直後に起こったけいれん発作も、それだけであれば、てんかんとは区別されます。慢性に経過することがてんかんの特徴です。てんかん発作は大脳の機能が種々の形で表出されるので、発作症状はきわめて多彩です。しかし、ひとりの患者の発作症状はほぼ一定しています。なぜてんかんになるのか、なぜてんかん発作が起こるかは十分にはわかっていません。大脳の神経細胞ネットワークの亢奮と抑制の微妙なバランスの破綻に求める学説が有力です。てんかん発作を起こしやすい下 地（素因）と大脳の機能を不安定にする脳損傷（外因）とが複雑に絡み合っててんかんが発病すると推測されています。(Epileptic seizures are caused by excessive activation of cerebral nerve cells [epileptic discharge]. Epilepsy is a chronic disease of the cerebrum that causes repeated epileptic seizures. Seizures that occur only once or only during high fever are not considered epilepsy. Seizures that occur during the acute phase of encephalitis or immediately after head trauma are also not considered epilepsy. The chronicity of epilepsy is a characteristic of epilepsy. Since epileptic seizures express the functions of the cerebrum in various ways, the symptoms of the seizures are extremely diverse. However, the symptoms of seizures in a single patient are almost constant. It is not fully understood why epilepsy occurs. The prevailing theory is that it is due to the breakdown of the delicate balance between activation and inhibition of the cerebral nerve cell network. It is speculated that epilepsy develops when a complex combination of a predisposing factor [predisposing factor] that makes the patient susceptible to epileptic seizures and brain damage [external factor] that destabilizes the functions of the cerebrum is involved.)
てんかん発作が起こることで、脳にダメージを受けますか？(Do epileptic seizures cause brain damage?)	てんかん重積状態という特別な状態をのぞけば、通常、一度の発作で脳がダメージを受けることはありません。てんかん重積状態とは、30 分以上発作がとまらなかったり、ひとつの発作が終わった後、意識が完全に回復する前に次の発作が起こることを 30 分以上繰り返す状態を指し、注射など薬物で発作をとめる必要のある緊急事態です。ただし、長い経過を見ますと、記憶障害、知能低下や行動障害、精神医学的問題を合併することがありま す。これらは発作を繰り返した影響や、過ぎた薬物の服用、心理的要因、てんかんの原因となっているもと の病気の影響など種々の要因が関与しています。てんかんが難治に経過する場合は多種類の抗てんかん薬による治療はできるだけ避け、場合によってはてんかん外科手術を早めに検討したほうがいいことがあります。また、てんかんではありませんが、乳幼児期に熱性けいれんなどのけいれんが長時間持続すると、側頭葉 の内側にある海馬などの萎縮が起こり、それが数年から 10 年後に側頭葉てんかんの原因になることが知られています。(Except for a special condition called status epilepticus, a single seizure does not usually damage the brain. Status epilepticus refers to a state in which a seizure does not stop for more than 30 min, or a seizure occurs repeatedly for more than 30 min after one seizure has ended, before consciousness is fully restored. This is an emergency situation in which the seizures must be stopped with drugs such as injections. However, over the long term, memory disorders, intellectual disability, behavioral disorders, and psychiatric problems may occur. These are caused by various factors, such as the effects of repeated seizures, past drug use, psychological factors, and the effects of the underlying disease that causes epilepsy. If epilepsy is intractable, it is best to avoid treatment with multiple types of antiepileptic drugs as much as possible, and in some cases, it may be better to consider epilepsy surgery early. Although not classified as epilepsy, prolonged convulsions in infancy, such as febrile seizures, can lead to atrophy of the hippocampus and other inner temporal lobe structures, which is known to result in temporal lobe epilepsy several to 10 years later.)
高齢で発病するてんかんの特徴を教えてください。(What are the characteristics of epilepsy that develop in older adults?)	てんかんの多くは小児期に発病します。2008 年2 月～ 11 月に当院（静岡てんかん・神経医療センター）でてんかんと診断された928 名のうち、751 名（80.9 %）が20 歳までに発病していました。20 歳を過ぎるとてんかんの発病率は次第に低下していきますが、60 歳を越えると再び増加に転ずるといわれています。脳血管障害など、高齢になって新たに生ずる脳の器質的な障害を背 景に、てんかんが発病しやすくなると考えられています。当院の統計でも高齢になるとわずかに発病率が増え、15 名（1.6 %）が60 歳以降に発病していました。推定病因は脳梗塞2、脳膿瘍2、脳腫瘍 1、脳外傷1、長年の大酒1 名で、他は不明でした。このように 器質性の病因が約半数に認められるのは、てんかん全体から見れば多い数字です。一方、熱性けいれんの既往や、てんかんまたは熱性けいれんの家族歴をもつ例は1 名もありませんでした。診断は15 名全例が症候性部分てんかんで、うち4 名は側頭葉に焦点を認めました。発作型は全般性強直間代発作が 10 、複雑部分発作が 7、単純部分発作が5、非けいれん性てんかん重積状態の疑いが2 名でした（重複あり）。発作頻度は年単位 5、月単位 7、週単位 2、日単位1 名でした。ほぼ全例が結婚してお子さんもあり、お仕事もしてこられた方々です。一般に、高齢発病のてんかんは薬物治療で比較的容易に発作がコントロールされることが多いといわれています。高齢者では血中濃度が上昇しやすいので、服用量を決定する際には注意が必要です。(Most epilepsy develops in childhood. Of the 928 people diagnosed with epilepsy at our hospital [Shizuoka Epilepsy and Neurological Medical Center] between February and November 2008, 751 [80.9%] had developed the disease by the age of 20. The incidence rate of epilepsy gradually decreases after the age of 20, but it is said to increase again after the age of 60. It is thought that epilepsy is more likely to develop against the background of organic brain disorders that occur in old age, such as cerebrovascular disorders. According to our hospital's statistics, the incidence rate increases slightly with age, with 15 people [1.6%] developing the disease after the age of 60. The suspected causes were cerebral infarction [two cases], brain abscess [two cases], brain tumor [one case], brain trauma [one case], and long-term heavy drinking [one case], and the causes of the remaining cases were unknown. As such, the fact that organic causes are found in about half of the cases is a high number compared to the overall number of cases of epilepsy. On the other hand, none of the patients had a history of febrile convulsions or a family history of epilepsy or febrile convulsions. All 15 patients were diagnosed with symptomatic partial epilepsy, of whom four had a focal seizure in the temporal lobe. The seizure types were generalized tonic-clonic seizures in 10 patients, complex partial seizures in seven patients, simple partial seizures in five patients, and suspected non-convulsive status epilepticus in two patients [with some overlapping]. The frequency of seizures was annual in five patients, monthly in seven patients, weekly in two patients, and daily in one patient. Almost all patients were married, had children, and were working. In general, it is said that seizures in elderly patients with epilepsy can be relatively easy to control with drug therapy. Drug use led to an increase in blood levels of the elderly, so caution is required when determining the dosage.)

Note: Information is provided in Japanese along with its English translation.

datasets, emphasizing transparency and reproducibility in their development. They span a range of sizes, from 7 billion to 65 billion parameters, allowing researchers to study scaling laws and model behavior across different model capacities. Notably, LLaMA models achieve competitive or superior performance compared to larger proprietary models by optimizing data quality and training methodology rather than merely increasing model size. Due to their open-access nature and strong performance, LLaMA models have quickly become a foundation for further research into instruction-tuned models, alignment, domain adaptation, and low-resource language modeling. Several instruction-following models, such as Alpaca, are built by fine-tuning smaller LLaMA checkpoints with specially constructed instruction datasets.

2.2.2. LLM-jp

The LLM-jp model is an initiative focused on developing LLMs specifically optimized for the Japanese language. Recognizing that many widely used LLMs are predominantly trained in English and multilingual datasets with limited Japanese content, LLM-jp aims to address the gap by creating models that better capture the linguistic nuances and specific syntax of Japanese. LLM-jp models are typically trained on extensive corpora consisting of high-quality Japanese text sourced from books, news articles, web data, and other publicly available materials. The model is open access to promote research and development within the Japanese AI community. A distinguishing feature of LLM-jp is its careful curation of datasets to ensure broad coverage across different domains while maintaining linguistic richness and correctness. In addition, instruction-tuning and fine-tuning on Japanese-specific tasks are integral to the project, making the models more suitable for downstream applications, such as summarization, translation, question answering, and dialogue generation in Japanese. LLM-jp plays an essential role in enabling the creation of high-quality, culturally aligned AI systems for Japanese users and contributes to the broader goal of linguistic diversity and inclusivity in AI development.

2.3. LLM fine-tuning

Fine-tuning an LLM model involves adapting a general pre-trained model to work better on specific tasks or domains by continuing the training process with new data. In this study, two open-source models, LLaMA and LLM-jp, were used as base models. The training data used in LLaMA and LLM-jp are presented in Table 2. In the LLaMA training, most of the training data comes from English. In contrast, Alpaca was used as the fine-tuning dataset, and a Japanese version of Alpaca was also used for fine-tuning to improve the model's Japanese capabilities. LLM-jp training included datasets with more Japanese corpus, and three different Japanese fine-tuning datasets were used for model fine-tuning.

Table 2. Training and fine-tuning datasets

Dataset	Content	Language
LLaMA training	Common Crawl	English
	C4	English
	GitHub	English
	Wikipedia	20 languages
	Books	English
	arXiv	English
	Stack Exchange	English
Alpaca	Instruction-following	English
Alpaca (Japanese)	Instruction-following	Japanese (translated from Alpaca)
LLM-jp training	mC4	Japanese
	Wikipedia	Japanese
	Pile	English
	Wikipedia	English
	Stack (code)	English
Jaster	Instruction-following	Japanese
Dolly (Japanese)	Instruction-following	Japanese (translated from Dolly)
OASST (Japanese)	Instruction-following	Japanese (translated from OASST)
Epilepsy dataset	Epilepsy knowledge	Japanese

Abbreviation: OASST: OpenAssistant Conversations Dataset Jaster: j + asterisk.

3. Results

Herein, we present a comprehensive evaluation of the proposed fine-tuning method through a series of controlled experiments. The primary objective was to assess the impact of domain- and language-specific fine-tuning on the performance of LLMs in epilepsy-related tasks. Since we had limited computing resources (four 80 GB A100s), for LLaMA experiments, LLaMA (7B) was used as the base model for fine-tuning, while larger models LLaMA (13B) and LLaMA (30B) were used directly for inference. To achieve general-purpose instruction-following capabilities, we fine-tuned LLaMA-7B using the Alpaca dataset, a widely used synthetic instruction-tuning dataset based on OpenAI's text-davinci-003. In addition, our main focus was to fine-tune the model using our Japanese language epilepsy dataset, which contained domain-specific medical knowledge tailored to epilepsy. Since the epilepsy dataset is collected in Japanese, to verify the impact of language on LLM performance, a translated version of the Alpaca dataset was also used for fine-tuning. The English-Alpaca dataset was translated using ChatGPT, while the Japanese-Alpaca dataset was obtained from Github[4]. Double fine-tuning was also conducted: First using the Alpaca or Japanese Alpaca dataset, followed by further fine-tuning using the epilepsy dataset.

For the LLM-jp experiments, we conducted both fine-tuning and inference evaluations using models from the LLM-jp family, which are pre-trained primarily on Japanese language data and optimized for Japanese language tasks. We selected the LLM-jp (1.3B) model as the base model for fine-tuning due to its manageable size and compatibility with our computational constraints. This model was fine-tuned using our curated Japanese epilepsy dataset, allowing us to evaluate the impact of domain-specific training on a smaller-scale Japanese language LLM. In addition, we also evaluated the LLM-jp (13B) model without additional fine-tuning, serving as a larger-scale baseline for comparison. To further explore the performance of existing instruction-tuned models in Japanese, we included three publicly released fine-tuned variants of LLM-jp (13B) provided by the LLM-jp initiative: LLM-jp-13B-instruct-full-jaster (fine-tuned using the JASTER instruction dataset), LLM-jp-13B-instruct-full-jaster-dolly-oasst (fine-tuned using a merged instruction dataset combining JASTER, Dolly, and OpenAssistant [OASST]), and LLM-jp-13B-instruct-full-dolly-oasst (fine-tuned on the combined Dolly and OpenAssistant datasets, without JASTER). Dolly[40] is the Japanese translation of databricks-dolly-15k, and OASST[41] is the Japanese translation of the English subset of OASST. The outputs from different models are presented in Table 3. The fine-tuned model generated responses that are more aligned with the intended clinical context and included essential contextual details. In contrast, the LLM-jp without fine-tuning was overly brief and omitted critical information.

In the evaluation phase, we adopted a comprehensive set of four widely recognized metrics to assess the performance of the LLMs on epilepsy-related tasks. Specifically, the evaluation metrics included: (i) Bilingual Evaluation Understudy (BLEU),[42] a precision-based metric that measures the overlap between the generated text and the reference text by calculating n-gram matches; (ii) Metric for Evaluation of Translation with Explicit ORdering (METEOR),[43] which extends beyond simple n-gram overlap by incorporating synonym matching, stemming, and word order penalties, making it more sensitive to linguistic variations and semantic meaning; (iii) Recall-Oriented Understudy for Gisting Evaluation (ROUGE-L),[44] which focuses on the longest common subsequence between the generated and reference texts, emphasizing both the sequence and recall aspects, and providing insights into how much essential information is preserved; (iv) Semantic Propositional Image Caption Evaluation (SPICE),[45] which evaluates the semantic content of the generated text by parsing it into scene graphs and comparing the relationships and objects described, offering a conceptual-level evaluation particularly valuable for specialized and technical domains.

The experimental results are displayed in Table 4, demonstrating that fine-tuning the base LLM with epilepsy-specific knowledge data significantly improves performance across all evaluation metrics. Fine-tuning allows the models to better capture specialized terminologies and contexts required for epilepsy-related tasks. Among the models, LLM-jp (1.3B) achieved the highest performance despite having the smallest number of parameters. This is most likely associated with the model's extensive use of Japanese language data during pre-training, which enhanced its ability to understand and

[4] https://github.com/masa3141/japanese-alpaca-lora/tree/main

Table 3. An example of different model outputs

Model	Output
Test data	スポーツは、身体を丈夫にして抗てんかん薬の副作用を出にくくしますので、してかまいません。ただし、発作が万が一おきた場合に致命的になるようなスポーツ（スキューバダイビング、ロッククライミング、など）は避けましょう。球技や陸上競技は問題ありません。(Sports are fine, as they strengthen the body and reduce the side effects of antiepileptic drugs. However, sports in which a seizure could be fatal (scuba diving, rock climbing, etc.) should be avoided. Ball games and track and field are fine.)
LLM-jp (1.3B)	運動が可能です。(It is possible to exercise.)
LLM-jp (1.3B) and epilepsy data	てんかんのある人が、時々、運動をすることは問題ありません。精神活動が、脳と身体の健康に良いからです。しかし、急な動作は危険ですので、慎むようにしてください。また、激しいスポーツは発作の誘発になりますので、中止するべきです。(It is okay for people with epilepsy to exercise occasionally, as mental activity is good for the brain and body. However, sudden movements can be dangerous, so they should be avoided. Also, strenuous sports can trigger seizures, so they should be avoided.)

Note: Output refers to the model's answer to the question: スポーツをしてもいいですか。 (Can I play sports?).

Table 4. Large language model (LLM) performance in epilepsy-related tasks

Model	First fine-tuning	Second fine-tuning	These are dimensionless scores (pure numbers)				
			BLEU	METEOR	ROUGE-L	SPICE	Mean
BioMedLM	-	-	0.0058	0.0091	0.0092	0.0000	0.0060
LLaMA (7B)	-	-	0.0173	0.0237	0.0234	0.0069	0.0178
LLaMA (7B)	Epilepsy data	-	0.2256*	0.1836*	0.2820*	0.1045	0.1989*
LLaMA (7B)	Alpaca	-	0.0273	0.0418	0.0639	0.0439	0.0442
LLaMA (7B)	Alpaca	Epilepsy data	0.1701	0.1705	0.2347	0.1070	0.1706
LLaMA (7B)	Alpaca (Japanese)	-	0.1637	0.1380	0.2217	0.0876	0.1528
LLaMA (7B)	Alpaca (Japanese)	Epilepsy data	0.2037	0.1678	0.2668	0.1308*	0.1923
LLaMA (13B)	-	-	0.0281	0.0559	0.0417	0.0057	0.0328
LLaMA (30B)	-	-	0.0281	0.0572	0.0417	0.0057	0.0332
LLM-jp (1.3B)	-	-	0.1418	0.1793	0.1805	0.0144	0.1290
LLM-jp (1.3B)	Epilepsy data	-	0.2351*	0.2314*	0.2631*	0.0727*	0.2006*
LLM-jp (13B)	-	-	0.1673	0.2102	0.2010	0.0198	0.1496
LLM-jp (13B)	Jaster	-	0.0004	0.0192	0.0174	0.0160	0.0132
LLM-jp (13B)	Dolly (Japanese)	-	0.0880	0.0891	0.1421	0.0647	0.0960
LLM-jp (13B)	Jaster; Dolly (Japanese)	-	0.0712	0.0889	0.1295	0.0712	0.0902

Note: *Indicates the best performance in each metric for LLaMA or LLM-jp.
Abbreviations: BLEU: Bilingual Evaluation Understudy; METEOR: Metric for Evaluation of Translation with Explicit ORdering; ROGUE-L: Recall-Oriented Understudy for Gisting Evaluation; SPICE: Semantic Propositional Image Caption Evaluation.

adapt when fine-tuned on Japanese epilepsy datasets. These results highlight the importance of both domain-specific knowledge and language alignment in improving LLM performance on specialized non-English tasks. Although these evaluation methods can provide useful insights into surface-level overlap and some aspects of semantic similarity, they may not fully capture clinical relevance or factual adequacy in medical contexts. In this study, we did not conduct a formal human evaluation due to resource constraints, but we recognize its value—particularly for assessing factual correctness, clinical appropriateness, and usability in real-world scenarios—and plan to implement it in future studies.

4. Discussion

While LLMs excel in a wide range of general tasks, their performance often falls short in specialized professional domains. This gap is particularly evident in tasks requiring domain-specific knowledge, where responses may lack the required depth, accuracy, and professionalism. Furthermore, these limitations become even more pronounced in non-English language tasks, where linguistic nuances and domain-specific terminology are not always well captured by models primarily pre-trained on English corpora. To address these challenges and enhance the reliability and professionalism of LLM-generated responses, recent research has explored fine-tuning pre-trained models using domain-specific knowledge bases. By integrating specialized expertise into the model, it becomes possible to significantly improve its performance on tasks requiring deep understanding within a specific field. In this study, we focused on the domain of epilepsy, aiming to fine-tune a pre-trained LLM using highly granular and specialized knowledge related to epilepsy. Unlike general domain adaptation, we focused on fine-tuning at a more detailed, technical level to ensure that the model can handle complex epilepsy-related inquiries with greater precision. Importantly, our fine-tuning efforts were conducted using non-English language resources to further bridge the linguistic and professional knowledge gaps. Through this approach, we aimed to develop a language model that delivers highly accurate, professional, and contextually appropriate responses in specialized epilepsy-related tasks, even in non-English settings.

In our experiments, we selected two different open-source pre-trained LLMs as base models for fine-tuning. The fine-tuning datasets were meticulously curated from a variety of publicly available websites, focusing specifically on sources that provide high-quality, domain-specific knowledge related to epilepsy. This approach ensured that the fine-tuning corpus was both comprehensive and rich in specialized content. Among the two models evaluated, the Japanese language model LLM-jp (1.3B) demonstrated the highest overall performance following fine-tuning. After incorporating detailed epilepsy knowledge, LLM-jp exhibited significant improvements in generating accurate, contextually appropriate, and professional responses to epilepsy-related queries in Japanese. These results highlight the effectiveness of leveraging fine-tuned domain knowledge to enhance the specialized capabilities of language models in non-English settings. In contrast, BioMedLM, a medical-domain pre-trained model originally optimized for English-language biomedical tasks, did not achieve comparable performance levels when evaluated on Japanese epilepsy tasks. The primary limitation stemmed from BioMedLM's lack of robust support for the Japanese language. Despite its strong foundation in biomedical knowledge, the absence of multilingual capabilities, particularly in Japanese, restricted its ability to adapt effectively to non-English fine-tuning, resulting in less professional and less accurate outputs compared to LLM-jp. These experimental findings underscore the critical importance of both language compatibility and domain specificity when fine-tuning LLMs for specialized tasks in non-English languages.

For LLaMA models without fine-tuning, increasing the parameters (from 7B to 30B) did not yield significant performance improvement. For baseline models without fine-tuning, we observed that simply increasing the model size did not significantly improve performance on Japanese epilepsy-related tasks. Despite the larger parameter count, the models struggled to understand and generate accurate responses to specialized epilepsy-related questions in Japanese. This limitation can be attributed to two main factors. First, the original pre-training corpus of LLaMA models contains very limited professional medical knowledge, particularly with respect to the field of epilepsy. Second, the proportion of Japanese language data included in the pre-training dataset is extremely small, which further restricts the model's ability to handle non-English, domain-specific inquiries effectively. As a result, across all model sizes, the baseline LLaMA models demonstrated inadequate comprehension and response quality when evaluated on Japanese epilepsy test datasets. However, after fine-tuning the LLaMA-7B model using a carefully curated dataset focused on Japanese language epilepsy knowledge, we observed a substantial performance improvement. The fine-tuned model exhibited a significantly enhanced understanding of epilepsy-specific terminology, clinical concepts, and contextual nuances within the Japanese language. This indicates that targeted domain-specific fine-tuning can dramatically compensate for the deficiencies of the original pre-trained model, even without requiring a larger parameter size. These findings highlight that, for specialized non-English applications, fine-tuning with high-quality, domain-specific datasets is far more critical than merely scaling up the model size. A well-focused fine-tuning strategy can enable even smaller models like LLaMA-7B to achieve strong task-specific performance, outperforming larger but unfine-tuned counterparts. The Alpaca dataset effectively improved the LLaMA performance in general tasks;[23] in the epilepsy task, it also resulted in a slight performance gain. However, the second fine-tuning using the epilepsy dataset did not achieve the highest performance. By using the Japanese-translated Alpaca and epilepsy datasets, the performance of the twice fine-tuned model was improved.

The LLM-jp model, which was pre-trained using a substantial amount of Japanese language data, naturally demonstrated strong baseline performance on a variety of Japanese language tasks. Its robust handling of Japanese text gives it a considerable advantage compared to models primarily trained on English or multilingual corpora with a lower proportion of Japanese content. Building on this strong foundation, we conducted fine-tuning of LLM-jp (1.3B) using a carefully curated dataset focused specifically on epilepsy-related knowledge in Japanese. The results revealed a significant performance improvement: the evaluation metric increased from 0.129 (pre-fine-tuning) to 0.2006 (post-fine-tuning). This substantial gain highlights how domain-specific fine-tuning can enhance a model's ability to understand and accurately respond to specialized queries, even when starting from an already strong language foundation. These findings further validate the effectiveness of targeted fine-tuning strategies for specialized tasks. In particular, when addressing narrow professional domains such as epilepsy within the broader field of medicine, merely relying on general pre-training is often insufficient, regardless of the model's original language competency. By supplementing the model with domain-specific knowledge during fine-tuning, it is possible to achieve remarkable improvements in task-specific performance, ensuring that the model's outputs are not only linguistically fluent but also professionally accurate and contextually relevant. This result underscores a critical point: for specialized applications, especially in languages other than English, domain-adaptive fine-tuning is essential to bridge the gap between general language understanding and expert-level task performance.

While the performance of LLMs was demonstrated in the study, several notable limitations should be acknowledged, particularly in the context of safety-critical applications such as healthcare. One key limitation is the challenge of knowledge updates. Instruction-tuned models are typically trained on fixed datasets, making it difficult to incorporate new clinical information without undergoing additional fine-tuning. In fast-evolving domains such as neurology, where diagnostic criteria and treatment protocols for conditions like epilepsy can change, this limitation hinders the model's ability to remain aligned with current clinical best practices.

In addition, there are also issues in their clinical application. LLMs can generate content that may be factually incorrect or misleading, a phenomenon often referred to as hallucination. This risk is especially problematic in clinical settings, where misinformation—even when subtle—can lead to adverse consequences for patient care. Finally, instruction-tuned models suffer from a lack of traceability. These models do not inherently provide mechanisms to link outputs to specific evidence or sources, making it difficult for users to verify the provenance of responses. This opacity can undermine clinician trust and limit the model's utility in settings where explainability and accountability are essential. In order to further integrate the model into clinical use, retrieval-augmented generation (RAG) is a better option for the above problems. RAG can retrieve relevant documents, allowing for easier integration of updated or external knowledge sources. By grounding responses in retrieved content, RAG has the potential to reduce hallucination and improve the traceability of clinical claims, a key limitation of current generative models.

During use, we also face the problem of overreliance. To mitigate overreliance, clinical use of LLMs must be governed by rigorous validation, clear human oversight, and well-defined boundaries of use. High-stakes decisions should remain under the purview of licensed professionals, with LLMs serving to support (and not replace) human expertise. The lack of transparency in LLMs makes it difficult to trace the source or rationale behind specific responses. This raises concerns regarding reproducibility and accountability. Ultimately, the responsible application of LLMs in clinical practice depends on a human-AI collaboration model, continuous performance monitoring, and ethical safeguards to ensure they support rather than undermine clinical judgment.

This study is focused on a specific clinical task (epilepsy) and language (Japanese), which may limit the immediate applicability of the findings to other areas. Nonetheless, the overall architecture and instruction-tuning framework we used were model- and language-agnostic in principle. With appropriate domain-specific data and adaptation, we believe the approach can be extended to other medical specialties and languages, though challenges, such as data availability, terminology alignment, and cultural context, would need to be carefully addressed.

5. Conclusion

In this study, we aimed to improve the professionalism and reliability of LLMs in handling epilepsy-related tasks in a non-English language. To achieve this, we fine-tuned pre-trained models using specialized epilepsy knowledge, rather than broad medical datasets commonly used in general-purpose medical LLMs. By focusing on more specific disease knowledge, the model can better understand and respond to professional epilepsy-related questions. Our experimental results revealed that narrowing the domain scope allows for significant performance improvements, even when using a relatively small amount of fine-tuning data. This indicates that targeted fine-tuning with high-quality, domain-specific information is an effective strategy for enhancing LLMs in specialized fields, especially where non-English resources are limited.

Acknowledgments

None.

Funding

This work was supported by JST CREST (grant number: JP-MJCR1784).

Conflict of interest

The authors declare they have no competing interests.

Author contributions

Conceptualization: All authors
Methodology: All authors
Investigation: Xuyang Zhao
Writing–original draft: Xuyang Zhao
Writing–review & editing: Qibin Zhao, Toshihisa Tanaka

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

All data used in this paper were collected from publicly accessible websites.

Further disclosure

EpilepsyLLM is dedicated to the research of LLMs in the medical field. The medical knowledge used in model training and testing is obtained from publicly accessible websites. The response content generated by the model cannot be guaranteed and cannot be used as a substitute for professional medical treatment.

References

1. Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. In: *31st Conference on Neural Information Processing Systems (NIPS 2017)*. Vol. 30. United States; 2017.
2. Radford A, Narasimhan K, Salimans T, Sutskever I. *Improving Language Understanding by Generative Pre-Training*. OpenAI Blog; 2018.
3. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. *Language Models are Unsupervised Multitask Learners*. Vol. 1. OpenAI blog; 2019. p. 9.
4. Brown T, Mann B, Ryder N, *et al*. Language models are few-shot learners. In: *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 33. United States: Cornell University; 2020. p. 1877-1901.
5. Ouyang L, Wu J, Jiang X, *et al*. Training Language Models to Follow Instructions with Human Feedback. In: *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 35. 2022. p. 27730-27744.
6. OpenAI, Achiam J, Adler S, *et al.* GPT-4 Technical Report. *arXiv.* Preprint posted online 2023.

   doi: 10.48550/arXiv.2303.08774
7. Touvron H, Lavril T, Izacard G, *et al.* LLaMA: Open and Efficient Foundation Language Models. *arXiv.* Preprint posted online 2023.

   doi: 10.48550/arXiv.2302.13971
8. Hoffmann J, Borgeaud S, Mensch A, *et al.* Training Compute-Optimal Large Language Models. *arXiv*. Preprint posted online 2022.

   doi: 10.48550/arXiv.2203.15556
9. Narang S, Chowdhery A. *Pathways Language Model (Palm): Scaling to 540 Billion Parameters for Breakthrough Performance*. Google AI Blog; 2022.
10. Zong H, Wu R, Cha J, *et al*. Large language models in worldwide medical exams: Platform development and comprehensive analysis. *J Med Int Res*. 2024;26:e66114.

    doi: 10.2196/66114
11. Brin D, Sorin V, Konen E, Nadkarni G, Glicksberg BS, Klang E. How Large Language Models Perform on the United States Medical Licensing Examination: A Systematic Review. *medRxiv.* Preprint posted online September 3, 2023.

    doi: 10.1101/2023.09.03.23294842
12. Gilson A, Safranek CW, Huang T, *et al*. How does ChatGPT perform on the United States medical licensing examination (USMLE)? The implications of large language models for medical education and knowledge assessment. *JMIR Med Educ*. 2023;9(1):e45312.

    doi: 10.2196/57594
13. Kim Y, Park C, Jeong H, *et al*. MDAgents: An Adaptive Collaboration of LLMs for Medical Decision-Making. In: *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*. Vol 37. 2024. p. 79410-79452.
14. Ullah E, Parwani A, Baig MM, Singh R. Challenges and barriers of using large language models (LLM) such as ChatGPT for diagnostic medicine with a focus on digital pathology--a recent scoping review. *Diagn Pathol*. 2024;19(1):43.

    doi: 10.1186/s13000-024-01464-7
15. Subramanian CR, Yang DA, Khanna R. Enhancing health care communication with large language models-the role, challenges, and future directions. *JAMA Network Open*. 2024;7(3):e240347-e240347.

    doi: 10.1001/jamanetworkopen.2024.0347
16. Mukherjee S, Gamble P, Ausin MS, *et al*. Polaris: A Safety-focused LLM Constellation Architecture for Healthcare. *arXiv*. Preprint posted online 2024.

    doi: 10.48550/arXiv.2403.13313
17. Abd-Alrazaq A, AlSaad R, Alhuwail D, *et al*. Large language models in medical education: Opportunities, challenges, and future directions. *JMIR Med Educ*. 2023;9(1):e48291.

    doi: 10.2196/48291
18. Yu H, Zhou J, Li L, *et al.* AI Patient: Simulating Patients with EHRs and LLM Powered Agentic Workflow. *arXiv*. Preprint posted online 2024.

    doi: 10.48550/arXiv.2409.18924
19. Yuan T, He Z, Dong L, *et al.* R-Judge: Benchmarking Safety Risk Awareness for LLM Agents. *arXiv.* Preprint posted online 2024.

    doi: 10.1111/medu.15402
20. Yuan T, He Z, Dong L, *et al*. *R-Judge: Benchmarking Safety Risk Awareness for LLM Agents*. arXiv [Preprint]; 2024.

    doi: 10.48550/arXiv.2401.10019
21. Ong JCL, Chang SYH, William W, *et al*. Medical ethics of large language models in medicine. *NEJM AI*. 2024;1(7):AIra2400038.

    doi: 10.1056/AIra2400038
22. Haltaufderheide J, Ranisch R. The ethics of ChatGPT in medicine and healthcare: A systematic review on large language models (LLMs). *NPJ Digit Med*. 2024;7(1):183.

    doi: 10.1038/s41746-024-01157-x
23. Taori R, Gulrajani I, Zhang T, *et al. Alpaca: A Strong,Replicable Instruction-Following Model.* Center for Research on Foundation Models. 2023. Available from: https://crfm.stanford.edu/2023/03/13/alpaca.html
24. Taori R, Gulrajani I, Zhang T, *et al*. *Stanford Alpaca: An Instruction-following LLaMA model*. GitHub Repository;

2023. Available from: https://github.com/tatsu-lab/stanford_alpaca

25. Gu Y, Tinn R, Cheng H, *et al*. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans Comput Healthc*. 2021;3(1):1-23.

    doi: 10.1145/3458754

26. Yasunaga M, Leskovec J, Liang P. LinkBERT: Pretraining Language Models with Document Links. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*; 2022. p. 8003-8016.

    doi: 10.18653/v1/2022.acl-long.551

27. Venigalla A, Frankle J, Carbin M. *BioMedLM: A Domain-Specific Large Language Model for Biomedical Text*. Vol. 23. United States: MosaicML; 2022. p. 2.

28. Luo R, Sun L, Xia Y, *et al*. BioGPT: Generative pre-trained transformer for biomedical text generation and mining. *Brief Bioinform*. 2022;23(6):bbac409.

    doi: 10.1093/bib/bbac409

29. Tu T, Azizi S, Driess D, *et al.* Towards Generalist Biomedical AI. *arXiv.* Preprint posted online 2023.

    doi: 10.48550/arXiv.2307.14334

30. Wang G, Yang G, Du Z, Fan L, Li X. ClinicalGPT: Large Language Models Finetuned with Diverse Medical Data and Comprehensive Evaluation. *arXiv.* Preprint posted online 2023.

    doi: 10.48550/arXiv.2306.09968

31. Wu C, Lin W, Zhang X, Zhang Y, Xie W, Wang Y. PMC-LLaMA: Toward building open-source language models for medicine. *J Am Med Inform Assoc*. 2024;31(9):1833-1843.

    doi: 10.1093/jamia/ocae045

32. Li Y, Li Z, Zhang K, Dan R, Jiang S, Zhang Y. ChatDoctor: A medical chat model fine-tuned on a large language model Meta-AI (LLaMA) using medical domain knowledge. *Cureus*. 2023;15(6):e40895.

    doi: 10.7759/cureus.40895

33. World Health Organization. *Epilepsy: A Public Health Imperative*. Geneva: World Health Organization; 2019.

34. Annegers JF, Rocca WA, Hauser WA. Causes of epilepsy: contributions of the Rochester epidemiology project. *Mayo Clin Proc*. 1996;71(6):570-575.

    doi: 10.4065/71.6.570

35. Shorvon SD. The causes of epilepsy: Changing concepts of etiology of epilepsy over the past 150 years. *Epilepsia*. 2011;52(6):1033-1044.

    doi: 10.1111/j.1528-1167.2011.03051.x

36. Korenke GC, Hunneman DH, Eber S, Hanefeld F. Severe encephalopathy with epilepsy in an infant caused by subclinical maternal pernicious anaemia: Case report and review of the literature. *Eur J Pediatr*. 2004;163:196-201.

    doi: 10.1007/s00431-004-1402-4

37. Pauschek J, Bernhard MK, Syrbe S, *et al*. Epilepsy in children and adolescents: Disease concepts, practical knowledge, and coping. *Epilepsy Behav*. 2016;59:77-82.

    doi: 10.1016/j.yebeh.2016.03.033

38. Unsworth C. Living with epilepsy: Safety during home, leisure and work activities. *Aust Occup Ther J*. 1999;46(3):89-98.

    doi: 10.1046/j.1440-1630.1999.00181.x

39. Aizawa A, Aramaki E, Chen B, *et al.* LLM-jp: A Cross-Organizational Project for the Research and Development of Fully Open Japanese LLMs. *arXiv.* Preprint posted online 2024.

    doi: 10.48550/arXiv.2407.03963

40. Conover M, Hayes M, Mathur A, *et al. Free Dolly: Introducing the World's First Truly Open Instruction-Tuned LLM*; 2023. Available from: https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm

41. Köpf A, Kilcher Y, Von Rütte D, *et al*. Openassistant Conversations-Democratizing Large Language Model Alignment. In: *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*. Vol 36. 2023. p. 47669-47681.

42. Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: A Method for Automatic Evaluation of Machine Translation. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*; 2002. p. 311-318.

    doi: 10.3115/1073083.1073135

43. Banerjee S, Lavie A. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In: *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*; 2005. p. 65-72.

    doi: 10.3115/1626355.1626389

44. Lin CY. ROUGE: A package for automatic evaluation of summaries. In: *Text Summarization Branches Out*. USA: Association for Computational Linguistics; 2004. p. 74-81.

45. Anderson P, Fernando B, Johnson M, Gould S. SPICE: Semantic Propositional Image Caption Evaluation. In: *Computer Vision--ECCV 2016: 14th European Conference, Amsterdam, the Netherlands, October 11-14, 2016, Proceedings, Part V 14*. Berlin: Springer; 2016. p. 382-398.

    doi: 10.1007/978-3-319-46454-1_24