MedNormJ: A Benchmark Dataset for Medical Concept Normalization in Japanese Clinical Documents

Yuki Tashiro, Seiji Shimizu, Tomohiro Nishiyama, Shoko Wakamiya, Eiji Aramaki
Nara Institute of Science and Technology, Japan
tashiro.yuki.ty3@naist.ac.jp
{shimizu.seiji.so8, nishiyama.tomohiro.ns5, wakamiya, aramaki}@is.naist.jp

Abstract

Medical concept normalization in clinical text is a fundamental technology for the secondary use of clinical data. However, constructing annotated resources for this task is challenging because annotation is both expertise-intensive and methodologically complex. As a result, a standard evaluation dataset for Japanese has yet to be established. In this study, we introduce a Japanese dataset for medical concept normalization, **MedNormJ** , which will be publicly available. The dataset consists of 397 pairs of medical expressions and their corresponding normalized disease names, manually curated from 96 medical documents, including case reports and radiology reports. Furthermore, we conduct comparative experiments using existing normalization approaches to benchmark their performance on this dataset in terms of both accuracy and computational efficiency. Through these experiments, we clarify the present performance level and identify remaining challenges specific to Japanese medical concept normalization.

Keywords: Medical Concept Normalization, Entity Linking, Clinical Text Processing, Benchmark Dataset, Large Language Model

1. Introduction

In recent years, there has been a growing global demand to accelerate drug development and clinical research through the use of real-world data derived from clinical text (Usuyama et al., 2025). Clinical text is a high-density information source that records patients' conditions in detail. However, such text is largely unstructured and exhibits substantial lexical variation, including synonyms, abbreviations, and typographical errors, which hinders large-scale automated analysis. To address this challenge, medical concept normalization has emerged as a foundational technology in clinical natural language processing.

This task can be formulated as mapping disease mentions within medical text (hereafter, mentions) to their corresponding concepts in a standardized terminology set (hereafter, an ontology). The mapped concepts are referred to as normalized forms. As illustrated in Figure 1, when utilizing the ICD-10-based Standard Disease Code Master, a Japanese disease terminology system derived from the International Classification of Diseases (ICD-10)[1], normalization entails selecting the most clinically appropriate concept from a search space of approximately 20,000 candidates for a given mention. For example, mentions such as "well differentiated adenocarcinoma" or "advanced sigmoid colon cancer" would be mapped to the normalized form "sigmoid colon cancer." This process of linking mentions to unique identifiers within ontologies like SNOMED CT (Chang and Mostafa, 2021) or ICD-10 is often referred to as medical coding.

Although the task of normalization is conceptually clear, constructing normalization datasets in the medical domain is challenging for two main reasons. First, annotation is costly: accurately handling complex disease names requires the involvement of experts with medical knowledge, and securing annotators and managing annotation effort pose stronger constraints than in general-domain settings (Wei et al., 2019; Fries et al., 2021). Second, there is uncertainty in deciding normalization granularity. Because medical concepts form deep hierarchical structures in ontologies (Figure 1), it can be difficult to determine which level of the hierarchy a mention should be linked to (Minh et al., 2025). For instance, for a mention such as "highly advanced sigmoid colon cancer," one must decide whether to preserve site information by selecting "sigmoid colon cancer" or to aggregate it into broader concepts such as "colon cancer" or "tumor," depending on downstream analysis goals and context. In other words, medical concept normalization faces two major challenges: a resource challenge, referring to the difficulty of securing expert annotators and a definition challenge, referring to ambiguity in determining the correct normalization, even among experts. These challenges have prevented the establishment of a standard evaluation dataset for Japanese medical text, making it difficult to evaluate normalization methods in terms of both accuracy and computational efficiency.

Given this situation, the development of a carefully designed benchmark dataset becomes essen-

[1] https://www2.medis.or.jp/stdcd/byomei/index.html

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 324–335

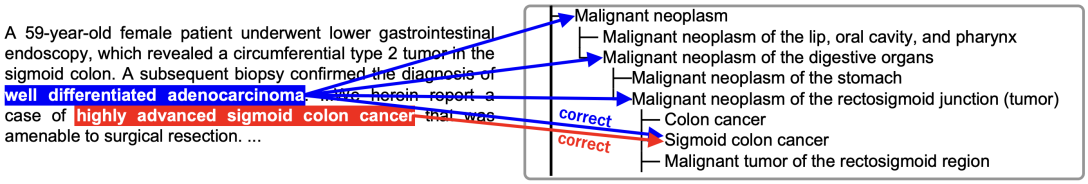

Figure 1: Overview of the medical concept normalization task. In this study, each disease mention in clinical text is mapped to a corresponding ontology concept. Mentions highlighted in red can be normalized literally (e.g., to “sigmoid colon cancer”), whereas those highlighted in blue are ambiguous in isolation and require contextual information for correct normalization. Thus, the task involves context-aware disambiguation based on surrounding text.

Guidelines	INPUT: target term (bold) and its context	OUTPUT: Normalized form and explanation
Principle: Context-dependent specificity When the disease concept to be annotated can be associated with a specific anatomical site (e.g., an organ) or clinical characteristics based on the context, it must be normalized to the **most specific disease name possible**. If the same concept is described at different levels of granularity within a document, the most specific disease name that can be inferred from the context should be selected as the normalized form.	胃底部に隆起性腫瘍を認めた. (An **elevated tumor** was observed in the gastric fundus.)	胃腫瘍 (**Gastric tumor**) Note: As the specific anatomical site is identifiable from the context, the general term “tumor” (腫瘍) must be avoided in favor of the more granular term “gastric tumor”.
Rule 1: Inclusion of lexical modifiers Modifiers that constitute an established medical observation or a standardized clinical concept must be retained in the normalized form.	盲腸に隆起性病変を認めた. (An **elevated lesion** was observed in the cecum.)	隆起性病変 (**Elevated lesion**) Note: The modifier “elevated” is an integral component of the established clinical observation; therefore, it should not be reduced to the more generic “lesion”.
Rule 2: Explicit causative factors When the causative factor is explicitly mentioned, the normalization should include the corresponding clinical category representing that cause.	乳アレルギーがあるため食事制限を行った. (Dietary restriction was implemented due to a **milk allergy**.)	食物アレルギー (**Food allergy**) Note: The specific cause “milk” indicates the category as food-related; therefore, the broader term “allergy” should be avoided.
Rule 3: Pathological diagnosis When a definitive clinical diagnosis is described in pathological findings, the diagnosis should be adopted as the normalized form.	病理所見は低分化型腺癌であった… 上行結腸癌の1例を経験した. (Pathology revealed a **poorly differentiated adenocarcinoma** … a case of ascending colon cancer.)	上行結腸癌 (**Ascending colon cancer**) Note: The final clinical diagnosis takes precedence over the histological description “poorly differentiated adenocarcinoma”.
Rule 4: Abbreviation expansion All medical abbreviations must be expanded to their full, stardardized technical expressions.	既往に**GERD**がある. (The patient has a history of **GERD**.)	胃食道逆流症 (**Gastroesophageal reflux disease**) Note: Abbreviations must be expanded to their full standardized form.
Exception 1: Abstruction of multiple findings When multiple specific findings are listed and a higher-level concept encompasses them, the higher-level concept should be adopted.	胸部X線で小結節，すりガラス影，腫瘤影を認めた. (Chest X-ray revealed **small nodules, ground-glass opacities, and mass shadows**.)	胸部異常陰影 (**Abnormal chest shadow**) Note: Multiple findings are grouped under a single encompassing concept; therefore, individual findings like “small nodule” are not selected.
Exception 2: Unification of synonymous findings Synonymous or closely related terms referring to the same clinical entity (e.g., mass-like pulmonary lesions) should be normalized to a single representative term.	肺に腫瘍／腫瘤性病変を認めた. (A **tumor / mass-like lesion** was observed in the lung.)	肺腫瘤 (**Pulmonary mass**) Note: Synonymous expressions should be unified into a representative term; the general term “tumor” is avoided here for consistency.
Exception 3: Malignancy preference If the context suggests malignancy, the normalized form should prioritize “cancer” over neutral terms like “tumor” to reflect the diagnostic intent.	画像上腫瘤性病変を認め，悪性が強く疑われた. (Imaging revealed a **mass-like lesion**, strongly suggestive of malignancy.)	癌 (**Cancer**) Note: Given the strong clinical suspicion of malignancy, the specific term “cancer” is preferred over the neutral “tumor”.

Table 1: Abstract guidelines for medical concept normalization and corresponding annotation examples.

tial. In this study, we use the concept normalization to collectively refer to the process of linking mentions to standardized concepts. We construct a Japanese medical concept normalization evaluation dataset, **MedNormJ**. We first performed pilot annotations by experts and analyzed cases of disagreement between annotators. Based on the analysis, we created annotation guidelines with explicit rules to unify normalization granularity. Specifically, the principle is to refine to more specific concepts whenever site, characteristics, or stage can be identified from context (Table 1). We then annotated the full dataset according to the guidelines and evaluated existing normalization methods in terms of performance and computational efficiency. The constructed dataset is publicly released for research purposes. [2]

[2] https://github.com/sociocom/mednormj

In this study, we make three contributions: (1) constructing **MedNormJ**, a Japanese benchmark dataset for medical concept normalization; (2) designing annotation guidelines to address normalization ambiguity; and (3) benchmarking existing normalization methods with respect to accuracy and computational efficiency.

2. Related Work

This section reviews prior research on medical concept normalization from both the dataset and methodological perspectives.

2.1. Medical Concept Normalization Datasets

A variety of benchmark datasets have been developed for medical concept normalization in English. The NCBI Disease Corpus (Doğan et al., 2014) is one of the most widely used datasets for disease normalization, while BC5CDR (Li et al., 2016) and MedMentions (Mohan and Li, 2019) extend coverage to chemical entities and a broad range of UMLS concepts, respectively. These resources have supported the development of representative normalization approaches, including DNorm (Leaman et al., 2013) and SapBERT (Liu et al., 2021a).

In contrast, the Japanese medical domain still lacks a widely shared benchmark dataset for medical concept normalization. Although several studies have constructed proprietary datasets (Morita et al., 2013; Aramaki et al., 2014), these resources either do not provide normalization annotation guideline or restrict normalization to ICD-10 codes. Because ICD-10 is primarily intended for statistical and administrative use, it may not capture the diverse disease expressions observed in clinical text at clinically sufficient granularity.

Annotation is also an essential component of dataset construction. A critical aspect of the annotation process is the assessment of both the effectiveness of the annotation scheme itself and the extent to which annotators correctly understand and follow the guidelines (Fernandes et al., 2025). The success of annotation largely depends on the clarity, consistency, and comprehensiveness of the documentation, as well as on the annotators' training and familiarity with the scheme (Artstein and Poesio, 2008). Well-developed guidelines, which include explicit definitions and illustrative examples, are indispensable for achieving reliable and accurate annotations (Stubbs and Pustejovsky, 2012).

The validation of an annotation scheme typically involves a combination of pilot annotations, iterative refinement of the guidelines, and qualitative analysis of problematic cases (Fernandes et al., 2025). In the annotation process, it is common practice to collect judgments from multiple annotators for each data instance, a methodology that has been widely recognized as effective for improving annotation quality (Snow et al., 2008). To quantitatively assess annotation quality, inter-annotator agreement (IAA) is the most commonly used metric, as it measures the consistency of annotations across annotators (Artstein and Poesio, 2008). Representative measures for computing IAA include Cohen's κ coefficient (Cohen, 1960), Krippendorff's α coefficient (Krippendorff, 2018), and simple agreement rates, commonly referred to as accuracy. While high IAA scores suggest that the guidelines are clear and effective, low agreement may arise from a variety of factors (Bayerl and Paul, 2024), often revealing underlying ambiguities or conceptual difficulties in the annotation task.

2.2. Medical Concept Normalization Methods

Existing approaches to medical concept normalization can be broadly categorized into dictionary-based, machine learning-based, bi-encoder-based, and large language model (LLM)-based methods.

Dictionary-based methods provide a simple baseline by directly matching an input mention against entries in a controlled vocabulary. String-based matching includes exact match and edit distance measures such as Levenshtein distance (Levenshtein et al., 1966). Although these techniques offer fast inference and low implementation cost, they are sensitive to out-of-vocabulary mentions and struggle to resolve context-dependent ambiguity.

To mitigate these limitations, machine learning–based approaches have been proposed. In particular, learning-to-rank methods such as DNorm (Leaman et al., 2013) model the correspondence between mentions and normalized forms using engineered features, and have been shown to outperform purely string-based baselines. For Japanese, extensions of DNorm have also been explored (Ujiie et al., 2020), where language-specific preprocessing yields additional gains.

More recently, bi-encoder approaches based on pretrained language models (e.g., BERT) have become a mainstream paradigm. These methods embed mentions and candidate normalized forms into a shared vector space and perform normalization by nearest-neighbor retrieval under an embedding similarity metric. Models such as SapBERT (Liu et al., 2021a,b) , which leverage synonym supervision during representation learning, achieve strong performance on biomedical normalization tasks.

LLM-based approaches have recently gained increasing attention. LLMs can incorporate broader contextual information and have been used to rerank candidate lists produced by bi-encoder re-

trieval, often improving accuracy (Abdulnazar et al., 2024). Nevertheless, LLM-based reranking typically incurs substantially higher latency and computational cost, which may limit its applicability in operational settings.

3. MedNormJ

In this section, we present the construction process of **MedNormJ**, a Japanese medical concept normalization dataset. Section 3.1 details the source materials used as the basis of **MedNormJ**, while Section 3.2 reports the pilot annotation and analyzes IAA. Section 3.3 describes the development of annotation guidelines based on the pilot analysis, and Section 3.4 presents the main annotation process along with dataset statistics and IAA.

3.1. Materials

As the basis of **MedNormJ**, we adopt MedTxt, an existing Japanese named entity recognition dataset. MedTxt consists of two subsets: MedTxt-CR (Yada et al., 2022), which comprises 50 case reports, and MedTxt-RR (Nakamura et al., 2022), which comprises 46 radiology reports. Their statistics are shown in Table 3. Case reports are academic articles describing rare clinical cases, whereas radiology reports are written by radiologists who interpret medical images such as CT and MRI and summarize imaging findings and associated diseases. In this study, we newly link each disease mention in these datasets to a standardized disease name in the ICD10-based Standard Disease Code Master, which serves as the ontology for normalization.

3.2. Pilot Annotation

To assess the reliability of medical concept normalization and identify sources of annotation ambiguity, we conducted a pilot annotation and computed IAA. For the pilot annotation, we randomly sampled a total of 100 mentions (50 from each dataset) and asked two Japanese healthcare professionals to annotate them independently.

Table 4 shows that the overall accuracy was 0.570, and disagreements were observed in 43 instances. Analysis of the disagreements revealed systematic mismatches mainly due to differences in concept granularity.

Table 2 presents an error analysis of the pilot annotation and shows the top three error categories that were consistently inconsistently annotated among annotators. It reports representative examples of inter-annotator disagreements and their corresponding counts. In total, 24 of the 43 observed errors were attributable to these three categories. For example, for the mention “abnormal lung field shadow,” surface variation occurred where different synonyms such as “abnormal chest shadow” and “abnormal lung shadow” were assigned; additionally, granularity differences were evident depending on whether context supplementation was applied, such as “mass” vs. “lung mass” (Table 2). These results indicate that medical concept normalization is ambiguous and challenging.

3.3. Guideline Development

Based on the pilot analysis, we developed annotation guidelines to ensure consistent normalization quality and to systematize decision criteria. These rules are shown in Table 1. Our primary rule is to refine to lower-level concepts as much as possible when site, characteristics, or stage can be identified from context.

However, the pilot disagreements included patterns that were difficult to resolve with the refinement principle alone. First, radiology reports often describe multiple findings in parallel; normalizing each finding individually can lead to greater variability across annotators. Indeed, in the pilot annotation, the agreement for MedTxt-RR was 0.360, lower than 0.780 for MedTxt-CR (Table 4). Therefore, we defined an exception rule: if multiple findings can be interpreted as the same condition and there exists a reasonable higher-level concept that subsumes them, we normalize to that higher-level concept (Exception 1).

Second, we observed systematic variation arising from synonymous or closely related expressions that refer to the same clinical entity. For example, expressions such as “tumor” and “mass-like lesion” were sometimes normalized differently despite referring to essentially the same concept in context. To reduce variability caused by lexical differences, we introduced a unification rule (Exception 2): synonymous or closely related findings should be normalized to a single representative term.

Third, we observed ambiguity not only in hierarchical granularity but also in diagnostic interpretation. In particular, for expressions where malignancy is not explicitly specified (e.g., “tumor” or “mass”) versus expressions that clearly indicate malignancy (e.g., “cancer” or “adenocarcinoma”), annotators differed in whether to interpret the mention as a neutral lesion concept or as a malignant disease concept. This distinction is not purely hierarchical, but reflects differences in clinical interpretation of malignancy. Therefore, we introduced a term-specific rule (Exception 3): if malignancy is explicitly indicated, normalize to “cancer” or “malignant neoplasm”; otherwise, normalize to “tumor.”

3.4. Main Annotation

We constructed the final evaluation dataset using all available documents, based on the developed

Annotator H1	Annotator H2	Freq.
腫瘤 (Tumor)	肺腫瘤 (Pulmonary mass)	9
肺野異常陰影 (Abnormal lung field shadow)	胸部異常陰影 (Abnormal chest shadow)	8
肺野異常陰影 (Abnormal lung field shadow)	肺部異常陰影 (Abnormal pulmonary shadow)	7

Table 2: Representative examples of systematic disagreements between two annotators (H1 and H2) in the pilot phase. Each row shows normalized forms independently assigned by Annotators H1 and H2 to the same mention in context, along with their frequency. These disagreements mainly reflect differences in normalization granularity and synonym selection.

	CR	RR	MedNormJ
Reports	50	46	96
Mentions	270	127	397
Unique mentions	234	62	295
Unique normalized forms	187	24	206
Mentions/normalized ratio	1.4	5.3	1.9

Table 3: Statistics of MedNormJ (H1&H2 agreement subset). CR and RR correspond to MedTxt-CR and MedTxt-RR, respectively.

Phase	Metric	MedTxt-CR	MedTxt-RR	Overall
Pilot	Accuracy	0.780	0.360	0.570
	Cohen's κ	0.776	0.346	0.565
Main	Accuracy	0.761	0.784	**0.768**
	Cohen's κ	0.759	0.722	**0.762**

Table 4: IAA in the pilot and main phases. Accuracy and Cohen's κ are shown for MedTxt-CR, MedTxt-RR, and overall. Agreement improved in the main phase after guideline refinement.

guidelines. Dataset statistics are shown in Table 3. The dataset covers a total of 96 medical documents, containing 397 normalized mentions in total. The number of unique normalized forms is 206.

We recomputed IAA on the constructed dataset using simple accuracy and Cohen's κ coefficient. As a result, the overall agreement for **MedNormJ** improved from 0.570 to 0.768. Note that the pilot annotation (n=100) and the main annotation (n=515) involve different sample sizes, so the comparison is not strictly controlled. According to commonly used interpretation criteria (Landis and Koch, 1977), this level of agreement corresponds to substantial agreement, indicating that the constructed dataset exhibits a reasonable level of annotation reliability.

On the other hand, systematic disagreements for certain terms were also observed (Table 5). For example, for expressions containing "muscle tone," annotator H1 tended to assign "muscle tone disorder," whereas annotator H2 tended to assign "spasticity." For "atelectasis," disagreements arose regarding whether to normalize to a lower-level concept with site information (e.g., "lower-lobe atelectasis") based on context, or to use the surface form "atelectasis." These results suggest the need for more detailed guidelines for clinically ambiguous expressions, or flexible evaluation schemes such as allowing multiple labels.

4. Experiments

This section summarizes our experimental setup and results on **MedNormJ**. Since the compared methods are described in detail in the subsection below, we focus on the overall evaluation design, including the evaluation metrics and how the results are organized.

4.1. Evaluation Setup

We evaluate medical concept normalization on **MedNormJ** from two perspectives: performance and computational efficiency. Performance is measured by Accuracy@1, and efficiency by average inference time per sample.

Accuracy@1 is defined as the proportion of instances for which the top-ranked normalized form exactly matches the gold label. This metric is widely used in prior work on medical concept normalization and entity linking (Kate, 2021; Luo et al., 2020; Liu et al., 2021a). Accuracy@1 is particularly appropriate for this task because normalization is formulated as selecting a single standardized concept from a predefined candidate set, and exact match evaluation directly reflects whether the model successfully identifies the correct concept.

To evaluate computational efficiency, we measure the average inference time per sample and analyze the trade-off between normalization accuracy and processing cost. All experiments were conducted on a single NVIDIA A100-PCIE-40GB GPU with an AMD EPYC 7702P 64-core CPU. GPT-5.2 (model version: `gpt-5.2-2025-12-11`) was accessed via the OpenAI API under default decoding settings (temperature = 0.0), and the reported inference time includes API latency.

Mentions	Normalized form by H1	Normalized form by H2	Freq.
... Muscle tone (increased)...	Muscle tone disorder	Spasticity	4
Atelectasis	Lower lobe atelectasis	Atelectasis	4

Table 5: Representative examples of systematic disagreement between two annotators (H1 and H2), including contextual information for each mention.

Model	Acc. (H1) (n=511)	Acc. (H2) (n=511)	Acc. (H1 & H2) (n=397)	Inference time (H1 & H2) (sec/sample)
Exact-match	0.399	0.411	0.476	2.10×10^{-5}
Levenshtein	0.406	0.418	0.479	4.13×10^{-4}
DNorm-J	0.451	0.466	0.547	1.00×10^{-3}
Bi-encoder	0.464	0.470	0.549	5.15×10^{-3}
Bi-encoder + GPT-5.2	0.521	0.536	0.602	7.77
Bi-encoder FT	0.534	0.540	0.625	6.00×10^{-3}
Bi-encoder FT + GPT-5.2	**0.572**	**0.560**	**0.653**	8.49
GPT-5.2 (zero)	0.311	0.301	0.358	3.83
GPT-5.2 (zero w/ all candidates)	0.466	0.452	0.565	5.59

Table 6: Comparison of normalization performance (Accuracy@1) and inference time. Acc. (H1) and Acc. (H2) denote accuracy when using annotator H1's and annotator H2's labels as gold, respectively. Acc. (H1 & H2) denotes accuracy when using only instances where both annotators agree as gold. Inference time denotes the average processing time per sample.

4.2. Compared Methods

To assess the difficulty of medical concept normalization on **MedNormJ**, we evaluated a set of representative normalization methods covering diverse modeling paradigms. Our objective is not to propose a new normalization model, but to characterize the current performance landscape of existing approaches for Japanese clinical text, considering both accuracy and inference efficiency.

We selected methods ranging from simple dictionary-based baselines to more advanced Bi-encoder-based and LLM-based approaches. Below, we briefly describe each method used in our experiments.

Dictionary-based As dictionary-based approaches, we employed two methods: an exact-match method and a more flexible edit-distance-based method. In the exact-match method, the input mention is matched against mentions registered in the medical terminology dictionary JMED-DICT mini (Nagai et al., 2025), which contains pairs of mentions and their corresponding normalized forms. If an exactly matching mention is found from this dictionary, the corresponding normalized form is returned as the candidate; otherwise, the input mention itself is returned.

The edit-distance-based method addresses typographical errors and orthographic variations by measuring string similarity. Specifically, we compute the Levenshtein distance (Levenshtein et al., 1966) between the input mention and each mentions in the dictionary. The Levenshtein distance is defined as the minimum number of character insertions, deletions, and substitutions required to transform one string into another. In this method, the dictionary mentions with the minimum distance to the input mention is selected, and its corresponding normalized form is returned as the candidate.

DNorm-J DNorm-J is a Japanese adaptation of DNorm (Leaman et al., 2013) that formulates medical concept normalization as a learning-to-rank problem using TF-IDF-based feature representations. In this study, we evaluated DNorm-J using an existing implementation configured for Japanese clinical text, without additional task-specific tuning.

Bi-encoder The Bi-encoder approach independently encodes mentions and normalized forms into a shared vector space and performs normalization based on embedding similarity. We evaluated two variants: a Bi-encoder without task-specific fine-tuning, and a fine-tuned version trained on the **MedNormJ** training data (Bi-encoder FT).

Bi-encoder + LLM Recent studies have reported improved normalization accuracy by reranking candidate lists generated by a Bi-encoder using LLMs with access to contextual information. Following this line of work, we included a Bi-encoder + LLM configuration in which GPT-5.2 reranks the top candidates. While this approach can improve accuracy, it incurs substantially longer inference time and higher computational cost, which we analyze from an operational perspective.

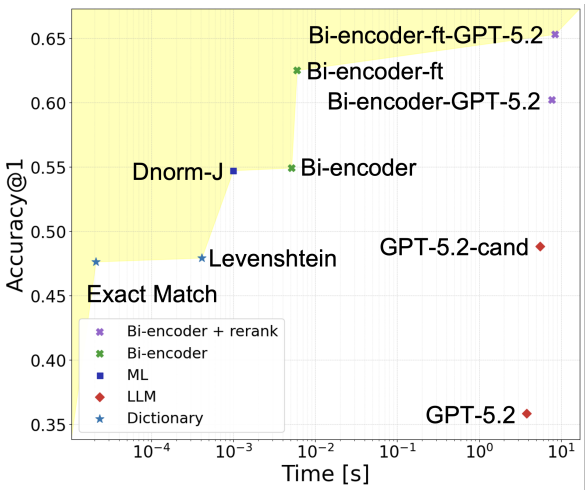

Figure 2: Inference time per sample [s] and Accuracy@1 for each method.

LLM We evaluate two zero-shot prompting strategies for GPT-5.2. The first prompt includes task and output-format instructions, together with the input mention and its context, and requires the model to directly generate the normalized form. The second prompt further provides a candidate list derived from JMED-DICT mini, enabling us to assess the effect of candidate constraints on performance.

4.3. Results

The results are shown in Table 6. While we report performance across all three settings, our analysis primarily focuses on the agreement subset (Acc. (H1 H2)), as it represents the most reliable ground truth derived from annotator agreement. Among the methods we evaluated, the hybrid approach, combining a Bi-encoder fine-tuned on JMED-DICT mini and GPT-5.2-based reranking, achieved the highest Accuracy@1 of 0.653. In contrast, GPT-5.2 in the zero-shot setting yielded the lowest performance, with an Accuracy@1 of 0.358.

Figure 2 plots Accuracy@1 against the average inference time for each method. The horizontal axis shows inference time on a logarithmic scale, and the vertical axis shows Accuracy@1. Methods in the upper-left region of the plot are both faster and more accurate. The yellow frontier line traces, for each latency regime, the maximum Accuracy@1 achieved. Methods on this frontier (exact-match, Levenshtein distance, the fine-tuned Bi-encoder, and the Bi-encoder with GPT-5.2-based reranking) represent the most efficient choices at their respective runtime scales.

In terms of latency, dictionary-based methods were the fastest, requiring only 2.10×10^{-5} seconds per sample. Bi-encoder-based approaches also maintained practical inference times while substantially improving accuracy.

By contrast, LLM-based reranking required over 1 second per sample, indicating much higher latency than dictionary-based methods. Within the reranking framework, using a reasoning-oriented model such as GPT-5.2 also incurred substantially higher computational cost than the other reranking models we evaluated, underscoring the trade-off between accuracy and efficiency.

5. Discussion

5.1. Principal Findings

Future direction: By analyzing expert annotations, we confirmed that medical concept normalization in Japanese medical text inevitably involves variability arising from differences in contextual interpretation and concept granularity. As shown in Table 5, systematic disagreements persist for specific expressions, highlighting the difficulty of defining a single unique correct answer in medical concept normalization. Therefore, future evaluation designs may benefit from more detailed guidelines and flexible frameworks, such as allowing multiple acceptable labels.

Language gap: In the baseline experiments, the highest score was 0.653. In this study, we used SapBERT, which was proposed in previous work (Liu et al., 2021b). The result obtained in our Japanese setting was higher than the previously reported Japanese result for the same model (0.240) on the Japanese subset of the XL-BEL benchmark reported in prior work (Liu et al., 2021b). However, it was still 0.129 lower than the reported English performance (0.782). These results show that performance in Japanese has improved compared to earlier reports, but a clear gap remains between Japanese and English. One main reason for this gap is the difference in the size of the training data. During the pretraining of SapBERT, about one million training instances were available for English, while only about 220,000 were available for Japanese. In addition, our experiments used only about 10,000 Japanese instances. This difference likely limits vocabulary coverage and the variety of expressions in Japanese. Another possible reason is the difference in the evaluation datasets and annotation policies. In Japanese, we observed variation in annotations due to differences in contextual interpretation and concept granularity. As a result, evaluation methods that assume only one correct answer may underestimate the true performance of the model. Overall, these findings suggest that improving performance in Japanese requires not only increasing the amount of training data, for example through synthetic data generation, but also improving the evaluation design and developing training strategies that consider hierarchical relationships

Error Type	Count	Ratio (%)
Prediction of semantically related terms	69	50.0
Hierarchical error (overly abstract prediction)	39	28.3
Mention bias	14	10.1
Hierarchical error (overly specific prediction)	9	6.5
Mentions appearing among candidate normalizations	4	2.9
Others	3	2.1

Table 7: Error types

between medical concepts.

Practical option: Evaluating not only accuracy but also inference time is essential when deploying medical concept normalization systems in real clinical and research settings. Our results with inference time provide practical guidance for selecting appropriate methods in such environments. No single method works best in every situation. The appropriate method depends on the purpose of the system, the required level of accuracy, the acceptable response time, and the available computational resources. When GPUs or other resources needed for deep learning models such as bi-encoders are limited, dictionary-based methods are effective. These approaches are especially suitable for batch processing, where large volumes of data must be handled efficiently. When sufficient computational resources are available and both high accuracy and fast inference per case are required, such as in real-time clinical decision support, a Bi-encoder and LLM are currently the most reliable option.

Our dataset is currently limited to two document types, namely case reports and radiology reports, and thus its coverage is restricted. Because writing styles and the availability of contextual information vary substantially across document types, an important future direction is to expand the dataset to include additional clinical document types, such as progress notes and discharge summaries, thereby connecting normalization research to more practical clinical applications.

Generalizability of exception rules: Although the proposed exception rules improved consistency, their generalizability to larger clinical corpora remains unclear, and additional rules or more flexible frameworks may be required.

5.2. Error Analysis

The most frequent error type involves the prediction of terms that are semantically related to, but not identical with, the gold standard normalized forms, accounting for 69 cases (50.0%). The second most common category consists of hierarchical errors, in which the model predicts a concept that is either more abstract or more specific than the correct one. In particular, overly abstract predictions were observed in 39 cases (28.3%), indicating difficulty in selecting the appropriate level of specificity within the disease name hierarchy.

In addition, 14 errors (10.1%) were attributed to mentions bias, where the model was overly influenced by string-level similarities in the input mention and consequently predicted a contextually inappropriate normalized form with a different clinical meaning. Such errors illustrate the model's tendency to rely on lexical cues at the expense of contextual interpretation.

Overly specific hierarchical errors were observed in 9 cases (6.5%), while errors in which the mention itself appeared among the candidate normalized forms accounted for 4 cases (2.9%). These latter cases suggest that the model failed to sufficiently leverage contextual information, defaulting instead to the mentions even when a more appropriate normalized form existed. The remaining errors (3 cases, 2.1%) were categorized as others.

Overall, this analysis indicates that, despite strong overall performance, the primary sources of error remain challenges in context-dependent decision making and hierarchical concept selection, highlighting directions for future improvements in disease concept normalization models.

6. Conclusion

We constructed a high-quality Japanese medical concept normalization dataset, **MedNormJ**, and reported a detailed annotation process and baseline results. In particular, by developing annotation guidelines that emphasize consistent normalization granularity and contextual interpretation, we achieved high IAA. Our results indicate that medical concept normalization on the dataset presents challenges, particularly due to ambiguity in concept granularity and contextual interpretation, and that there exists a trade-off between performance and computational efficiency across methods. We expect the dataset to contribute to further advances in Japanese medical NLP, including method development, and we plan to expand medical concept normalization datasets in future work.

Acknowledgments

This work was supported by the Cross-ministerial Strategic Innovation Promotion Program (SIP) on "Integrated Health Care System" (Grant Number JPJ012425), JST CREST (Grant Number JPMJCR22N1), and the Japan Society for the Promotion of Science (JSPS) Research Start-up Grant (JP25K24412), Japan.

Bibliographical References

Akhila Abdulnazar, Roland Roller, Stefan Schulz, and Markus Kreuzthaler. 2024. Large language models for clinical text cleansing enhance medical concept normalization. *IEEE Access*.

Eiji Aramaki, Mizuki Morita, Yoshinobu Kano, and Tomoko Ohkuma. 2014. Overview of the ntcir-11 mednlp-2 task. In *NTCIR*.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational linguistics*, 34(4):555–596.

Petra Saskia Bayerl and Karsten Ingmar Paul. 2024. What determines inter-coder agreement in manual annotations? a meta-analytic investigation. *Computational Linguistics*.

Eunsuk Chang and Javed Mostafa. 2021. The use of snomed ct, 2013-2020: a literature review. *Journal of the American Medical Informatics Association*, 28(9):2017–2026.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.

Ana Luísa Fernandes, Purificação Silvano, António Leal, Nuno Guimarães, Rita Ribeiro da Silva, Luís Filipe Cunha, and Alípio Jorge. 2025. The incremental process of building an annotation scheme for clinical narratives in portuguese: the contribution of human variation analysis. In *Research Challenges in Information Science-19th International Conference, RCIS 2025, Seville, Spain, May 20-23, 2025, Proceedings, Part II*.

Jason A Fries, Ethan Steinberg, Saelig Khattar, Scott L Fleming, Jose Posada, Alison Callahan, and Nigam H Shah. 2021. Ontology-driven weak supervision for clinical entity classification in electronic health records. *Nature communications*, 12(1):2017.

Rohit J Kate. 2021. Clinical term normalization using learned edit patterns and subconcept matching: system development and evaluation. *JMIR Medical Informatics*, 9(1):e23104.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Robert Leaman, Rezarta Islamaj Doğan, and Zhiyong Lu. 2013. Dnorm: disease name normalization with pairwise learning to rank. *Bioinformatics*, 29(22):2909–2917.

Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021a. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4228–4238.

Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2021b. Learning domain-specialised representations for cross-lingual biomedical entity linking. In *Proceedings of ACL-IJCNLP 2021*.

Yen-Fu Luo, Sam Henry, Yanshan Wang, Feichen Shen, Ozlem Uzuner, and Anna Rumshisky. 2020. The 2019 n2c2/umass lowell shared task on clinical concept normalization. *Journal of the American Medical Informatics Association*, 27(10):1529–e1.

Dao Sy Duy Minh, Nguyen Lam Phu Quy, Pham Phu Hoa, Tran Chi Nguyen, Huynh Trung Kiet, and Truong Bao Tran. 2025. Dragon: Dual-encoder retrieval with guided ontology reasoning for medical normalization. In *Proceedings of the 23rd Annual Workshop of the Australasian Language Technology Association*, pages 230–239.

Sunil Mohan and Donghui Li. 2019. Medmentions: A large biomedical corpus annotated with umls concepts. *arXiv preprint arXiv:1902.09476*.

Mizuki Morita, Yoshinobu Kano, Tomoko Ohkuma, Mai Miyabe, and Eiji Aramaki. 2013. Overview of the NTCIR-10 MedNLP Task. In *NTCIR*.

Hiroyuki Nagai, Tomohiro Nishiyama, Yuka Otsuki, Takako Fujimaki, Kyoko Kawabata, Noriko Kudo, Yuka Yamazaki, Haruya Shiraishi, Tomoyuki Kajiwara, Hiroyuki Shindo, Yoshimasa Kawazoe, Takeru Imai, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki. 2025. Jmed-dict: Construction of a large-scale medical terminology dictionary. In *Proceedings of the 31st Annual Meeting of the Association for Natural Language Processing (NLP2025)*, pages 3509–3514.

Yuta Nakamura, Shohei Hanaoka, Yukihiro Nomura, Naoto Hayashi, Osamu Abe, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki. 2022. Clinical Comparable Corpus Describing the Same Subjects with Different Expressions. *Stud Health Technol Inform*, 290:253–257.

Rion Snow, Brendan O'connor, Dan Jurafsky, and Andrew Y Ng. 2008. Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 254–263.

Amber Stubbs and James Pustejovsky. 2012. Natural language annotation for machine learning.

Shogo Ujiie, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki. 2020. Developing japanese disease normalization system. In *Proceedings of the 40th Joint Conference on Medical Informatics (The 21st Annual Meeting of the Japan Association for Medical Informatics)*.

Naoto Usuyama, Cliff Wong, Sheng Zhang, Tristan Naumann, and Hoifung Poon. 2025. Biomedical natural language processing in the era of large language models. *Annual Review of Biomedical Data Science*, 8.

Qiang Wei, Yukun Chen, Mandana Salimi, Joshua C Denny, Qiaozhu Mei, Thomas A Lasko, Qingxia Chen, Stephen Wu, Amy Franklin, Trevor Cohen, et al. 2019. Cost-aware active learning for named entity recognition in clinical text. *Journal of the American Medical Informatics Association*, 26(11):1314–1322.

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

A. Prompts: Bi-encoder FT + GPT-5.2

To ensure the reproducibility of our experiments, we provide the full prompts used for the GPT-5.2-based reranking step of the Bi-encoder FT + GPT-5.2 method below.

System Prompt (English Translation)

You are a skilled medical terminology expert who can deeply understand medical papers and clinical notes and accurately interpret subtle nuances in terminology. Rather than relying on simple string matching, always prioritize the medical meaning in context and perform the task with the highest possible accuracy.

System Prompt (Original Japanese)

あなたは、医学論文やカルテのテキストを深く理解し、用語の微妙なニュアンスを読み解く「熟練した医学用語専門家」です．単なる文字列のマッチングではなく、文脈上の医学的な意味を最優先し、常に最高の精度でタスクを実行してください．

Reranker Prompt (English Translation)

Role and Objective
As an expert medical terminology specialist, your task is to reorder (rerank) a given “candidate list” in order of the highest medical appropriateness and contextual relevance, based on three inputs: “mention form,” “original text,” and “candidate list.”

Instructions
1. Use only the terms provided in the “candidate list” and change only the order of the items.
2. Do not add any terms not present in the list, and do not delete or modify any terms within the list.
3. The final output must strictly follow the instructions in the “Output Format” section below.

Reasoning Steps
Follow the reasoning process below to determine the optimal order.
1. Understand the medical meaning of the mention form: First, accurately grasp what the <term> (mention form) within the <input> refers to.
2. Strict analysis of the context: Next, carefully read the <context> (original text) and identify the situation in which the <term> is used (e.g., patient’s chief complaint, physical findings, test results, or a confirmed diagnosis). In particular, never infer a disease name that is not directly stated in the original text.
3. Evaluate the candidate list: Finally, compare each term in the <candidates> list against the context analyzed above, evaluating their suitability one by one. If synonyms or related terms are present, carefully judge whether their level of specificity or generality matches the context.
4. Determine final ranking: Based on the evaluation above, reorder the list from most to least relevant.

Output Format
- Output only the reranked list as a Python list string (e.g., `['term1', 'term2', ...]`), without any additional explanations, introductions, concluding remarks, reasoning processes, or excuses.
- The list must contain exactly 10 elements.

Examples
Below is a concrete example demonstrating how to execute this task.

<example n="1" type="Good Case: Handling Symptoms">
<input>
<term>swelling of fingers</term>
<context>Since around July 1989, lower leg edema, swelling of fingers, skin sclerosis, and diffuse pigmentation appeared, and the first visit was in January 1990.</context>
<candidates>['lymphedema', 'swelling of fingers', 'rheumatoid arthritis', 'cerebral edema', 'swelling', 'parotid gland swelling', 'angioedema', 'edema', 'cellulitis', 'allergic edema']</candidates>
</input>
<reasoning>
The original text merely lists multiple symptoms in parallel. There is insufficient information to conclude a specific disease (e.g., lymphedema) causing the “swelling of fingers.” Therefore, “swelling of fingers,” which directly indicates the symptom itself, is the most appropriate. Interpretations that leap beyond the context are incorrect.
</reasoning>
<output>
['swelling of fingers', 'swelling', 'edema', 'lymphedema', 'angioedema', 'parotid gland swelling', 'allergic edema', 'cerebral edema', 'cellulitis', 'rheumatoid arthritis']
</output>
</example>

Final Prompt
Strictly follow the instructions and reasoning process above, and perform the task for the input information below.

<input>
<term>{{input}}</term>
<context>{{context}}</context>
<candidates>{{terms}}</candidates>
</input>

Reranker Prompt (Original Japanese)

Role and Objective
あなたは熟練した医学用語専門家として、与えられた「出現形」、「原文」、「候補リスト」の3つの入力情報に基づき、医学的に最も適切かつ文脈に即した重要度順に「候補リスト」を並べ替える（再ランキングする）任務を担います．

Instructions
1. 与えられた「候補リスト」内の用語のみを使用し、リストの順序だけを変更してください．
2. リストにない用語を追加したり、リスト内の用語を削除・変更したりしてはいけません．
3. 最終的な出力は、後述する「Output Format」セクションの指示に厳密に従ってください．

Reasoning Steps
モデルは以下の思考プロセスに従って、最適な順序を決定してください．
1. 出現形の医学的意味の把握: まず、<input>内の<term>（出現形）が何を指しているかを正確に理解します．
2. 文脈の厳密な分析: 次に<context>（原文）を精読し、<term>がどのような状況（例：患者の主訴、身体所見、検査結果、確定診断名）で使われているかを特定します．特に、原文に直接的な記載がない病名を推測することは絶対に避けてください．
3. 候補リストの評価: 最後に<candidates>（候補リスト）内の各用語を、上記で分析した文脈と照合し、一つずつ適合度を評価します．同義語や関連語が含まれる場合は、その具体性や一般性のレベルが文脈に合っているかを慎重に判断します．
4. 最終順位決定: 上記の評価に基づき、最も関連性が高いと判断した順にリストを並べ替えます．

Output Format
- 追加の説明、前置き、後書き、思考プロセス、言い訳などを一切含めず、再ランキングしたPythonのリスト形式の文字列 `['term1', 'term2', ...]` のみを出力してください．
- リストの要素は必ず10個にしてください．

Examples
以下は、このタスクをどのように実行すべきかを示す具体例です．

<example n="1" type="Good Case: Handling Symptoms">
<input>
<term>手指腫脹</term>
<context>平成1年7月頃より下腿浮腫,手指腫脹,皮膚硬化,びまん性色素沈着出現し,平成2年1月初診．</context>
<candidates>['リンパ浮腫', '手指腫脹', '関節リウマチ', '脳浮腫', '腫脹', '耳下腺腫脹', '血管性浮腫', '浮腫', '蜂窩織炎', 'アレルギー性浮腫']</candidates>
</input>
<reasoning>
原文は複数の症状を並列に記述しているに過ぎない．「手指腫脹」の原因となる特定の疾患（例：リンパ浮腫）を断定するだけの情報はないため、症状そのものを指す「手指腫脹」が最も適切である．文脈から飛躍した解釈は誤りである．
</reasoning>

<output>
['手指腫脹', '腫脹', '浮腫', 'リンパ浮腫', '血管性浮腫', '耳下腺腫脹', 'アレルギー性浮腫', '脳浮腫', '蜂窩織炎', '関節リウマチ']
</output>
</example>

Final Prompt
上記の指示と思考プロセスに厳密に従い、以下の入力情報に対するタスクを実行してください.

<input>
<term>{{input}}</term>
<context>{{context}}</context>
<candidates>{{terms}}</candidates>
</input>