

J-ClinicalBench: A Benchmark for Evaluating Large Language Models on Practical Clinical Tasks in Japanese

Seiji Shimizu¹, Tomohiro Nishiyama¹, Shohei Hisada¹, Yamato Himi¹
Shoko Wakamiya¹, Yuki Yanagisawa², Masami Tsuchiya²
Satoko Hori², Eiji Aramaki¹

¹Nara Institute of Science and Technology, Nara, Japan

²Keio University, Tokyo, Japan

{shimizu.seiji.so8, aramaki}@is.naist.jp

Abstract

Recent advances in large language models (LLMs) have accelerated the NLP applications in the medical and clinical domains. However, evaluations remain limited for non-English languages, such as Japanese, where clinical corpora are particularly scarce. To address this gap, we present **J-ClinicalBench**, a publicly available benchmark designed to reflect realistic Japanese clinical tasks. We first created 227 expert-authored clinical documents and newly constructed five datasets for core clinical tasks. Building on these datasets, **J-ClinicalBench** comprises nine clinical tasks spanning clinical language reasoning, generation, and understanding. We establish baseline performance on **J-ClinicalBench** by evaluating state-of-the-art proprietary and Japanese open-source LLMs, providing the first assessment of their utility in practical clinical scenarios. By releasing this benchmark, we aim to foster the development and evaluation of clinically applicable LLMs in Japanese healthcare, bridging the current gap between clinical NLP research and clinical practice. **J-ClinicalBench** and our evaluation are available at <https://github.com/seiji-shimizu/J-ClinicalBench-release>

Keywords: clinical NLP, corpus construction, evaluation benchmark

1. Introduction

In recent years, there has been rapid progress in adapting large language models (LLMs) for applications in the medical and clinical domain (Labrak et al., 2024; Han et al., 2023; Chen et al., 2023; Toma et al., 2023). Alongside this progress, a growing body of research has focused on benchmarking and systematically evaluating LLMs across diverse clinical tasks (Chen et al., 2025; Liu et al., 2024; Dorfner et al., 2024; Dada et al., 2025). Supported by the availability of medical and clinical resources in the English language (Adams et al., 2025; Johnson et al., 2016, 2023; Sun et al., 2013), these efforts have enabled comprehensive assessments of LLM performance in practical clinical settings.

However, conducting such evaluations is more challenging for non-English languages, such as Japanese, where clinical corpus scarcity is especially severe. Because of strict privacy regulations, publicly available clinical documents are extremely scarce for the Japanese language (Nakamura et al., 2025b; Yada et al., 2024). Existing Japanese medical benchmarks are largely based on translations of English datasets and focus primarily on simplified tasks, such as board examination question answering (Jiang et al., 2025; Kasai et al., 2023). Yet, translation fails to capture the diversity and complexity of clinical documentation, which can vary significantly across hospitals and individual practitioners (Laparra et al., 2020). Moreover, exam-

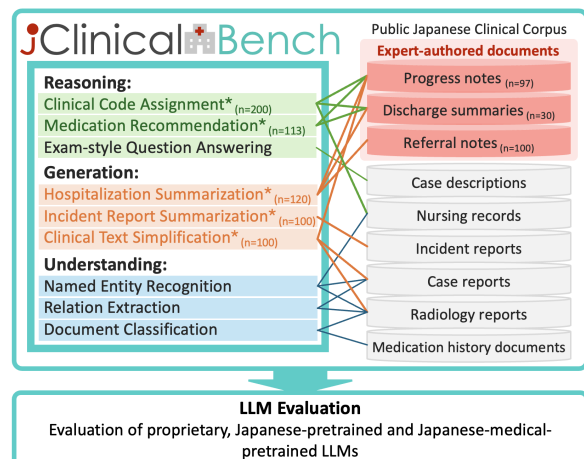


Figure 1: Overview of **J-ClinicalBench**. Datasets marked with * are newly created. We first created 227 expert-authored clinical documents in Japanese and constructed five new datasets. Building on these, **J-ClinicalBench** comprises nine tasks spanning clinical reasoning, generation, and understanding, enabling a comprehensive evaluation of LLMs in realistic clinical settings.

style QA tasks are insufficient for assessing the practical utility and reliability of LLMs in the clinical domain (Griot et al., 2025).

To bridge this gap, we introduce **J-ClinicalBench**, a publicly available benchmark designed to reflect realistic Japanese clinical

tasks, comprising nine tasks spanning clinical language reasoning, generation, and understanding (Fig. 1). To mitigate the scarcity of Japanese clinical corpora, we first **created 227 expert-authored clinical documents**, including discharge summaries and progress notes carefully designed to simulate in-hospital documentation. We then **constructed five new datasets for clinical tasks** such as diagnostic clinical code assignment and medication recommendation. Building upon these datasets, **J-ClinicalBench** enables a comprehensive evaluation of LLMs on tasks closely aligned with clinical practice.

By releasing this benchmark and conducting an initial evaluation of state-of-the-art proprietary and open-source Japanese LLMs, we aim to advance the development of clinically applicable LLMs in Japanese healthcare, thereby bridging the current gap between clinical NLP research and practical clinical utility. Our main contributions are as follows:

- Releasing J-ClinicalBench, a benchmark of nine clinically relevant NLP tasks to foster evaluation in practical clinical scenarios
- Creating a publicly available 227 expert-authored clinical documents in Japanese and newly constructing five clinical datasets to address the lack of realistic non-English clinical corpora
- Evaluation of proprietary, Japanese-pretrained, and Japanese-medical-pretrained LLMs on these tasks, providing a performance baseline in clinical tasks

2. J-ClinicalBench

In this work, we introduce **J-ClinicalBench**, a publicly available benchmark for clinical NLP in Japanese. All tasks are grounded in expert-authored Japanese clinical documents rather than translations, thereby reflecting the linguistic and contextual complexity of real-world clinical documentation. Table 1 provides an overview of the included tasks. The benchmark covers nine tasks in clinical language understanding, reasoning, and generation, following the categorization in Liu et al. (2024). In addition to existing datasets, we introduce five newly constructed datasets (marked with * in Table 1), which are described in Sect. 3.2.

2.1. Reasoning Tasks

Reasoning tasks include Clinical Code Assignment (CCA), Medication Recommendation (MR), and Exam-style Question Answering (ExamQA). These tasks evaluate whether LLMs can integrate medical knowledge with information explicitly stated in

the document to support clinically informed decisions (Nguyen et al., 2024). Exam-style QA is adapted from the IGAKU QA benchmark (Kasai et al., 2023), while CCA and MR are based on newly constructed datasets.

Clinical Code Assignment: Assigning diagnostic codes is essential for billing, epidemiological studies, and facilitating secondary use of clinical data (Falter et al., 2024). Supporting this process with LLMs has the potential to reduce clinician workload and improve coding consistency. This task requires assigning all diagnostic codes that correspond to the patient’s symptoms, diseases, and conditions, given descriptions in clinical notes. We employ MedDRA as the coding system, a widely adopted international standard that provides a controlled vocabulary for medical terminology (Brown et al., 1999).

Medication Recommendation: Inspired by Liu et al. (2024), this task evaluates whether LLMs can recommend suitable medications based on a patient’s clinical description. Provided with descriptions in clinical notes, LLMs are required to recommend all possible medications suitable for the patient’s condition, taking into account contextual information. This task simulates a decision-support scenario, where accurate recommendations can aid clinicians.

Exam-style Question Answering: This task is derived from the Japanese national medical board examination. Specifically, we use “clinical questions” (those beginning with “The patient is an XX-year-old...”), drawn from the IGAKU QA benchmark (Kasai et al., 2023). The objective is to evaluate whether LLMs can combine information from the text with relevant medical knowledge to select the correct option in a multiple-choice setting.

2.2. Generation Tasks

Generation tasks include Hospitalization Summarization (HS), Incident Report Summarization (IRS), and Clinical Text Simplification (CTS). These tasks focus on the ability of LLMs to rewrite clinical documents clearly and accurately. All three are based on newly created datasets.

Hospitalization Summarization: Writing clinical notes constitutes a major portion of healthcare professionals’ workload, contributing to stress and burnout (Gesner et al., 2019). This task aims to assess LLMs’ ability to generate concise narratives of a patient’s hospital course. Specifically, the task requires models to (i) summarize progress notes into discharge summaries and (ii) further condense discharge summaries into referral notes. Such automatic support could reduce documentation burden and enhance the quality of clinical communication across institutions.

Task	Description	Document Type(s)	Metrics	
Reasoning	Clinical Code Assignment (CCA)*	Assign diagnostic code for diseases, symptoms, and conditions reported in clinical documents	Progress notes, Discharge summaries, Nursing Records	F1
	Medication Recommendation (MR)*	Recommend appropriate medications based on the patient’s condition as described in progress notes and discharge summaries.	Progress notes, Discharge summaries	F1
	Exam-style Question Answering (ExamQA)	Multiple-choice question answering based on medical cases derived from IgakuQA	Case descriptions (exam questions)	Accuracy
Generation	Hospitalization Summarization (HS)*	Summarize notes during hospitalization in discharge summaries and further summarize them for patient referral	Progress notes, Discharge summaries, Referral Notes	ROUGE-L
	Incident Report Summarization (IRS) *	Generate short topic summaries based on clinical incident reports	Incident reports	ROUGE-L
	Clinical Text Simplification (CTS)*	Simplify radiology reports, progress notes and clinical research texts for patients.	Radiology reports, Progress notes, Case reports	ROUGE-L
Understanding	Named Entity Recognition (NER)	Extract clinical entities (disease and medication entities) from clinical texts	Radiology reports, Case reports, Nursing records	F1
	Relation Extraction (IE)	Identify relationships between medical entities (temporal and between-entity relations)	Radiology reports, Case reports	F1
	Document Classification (DC)	Classify clinical texts into predefined categories Comprised of ADE and TNM classification	Medication history documents, Radiology reports	Accuracy, F1

Table 1: Overview of J-ClinicalBench Tasks. We curated nine clinical tasks spanning clinical reasoning, generation, and understanding. Datasets marked with * are newly created.

Incident Report Summarization: Incident reports are central to patient safety, capturing errors, near-misses, and adverse events. NLP can potentially support the extraction of key information from these reports, facilitating large-scale analysis and timely detection of safety risks (Denecke and Paula, 2024). This task evaluates the ability of LLMs to extract essential information from incident reports (e.g., event type, contributing factors, outcomes) and generate concise summaries.

Clinical Text Simplification: Clinical documents often contain technical jargon and complex expressions that hinder comprehension by patients and non-specialist readers. This task requires LLMs to rewrite clinical notes in plain Japanese, while preserving medical accuracy and essential information. Successful simplification has the potential to improve communication between patients and healthcare providers and support accessibility of medical information for diverse populations (Xiao et al., 2025).

2.3. Understanding Tasks

Understanding tasks include Named Entity Recognition (NER), Relation Extraction (RE), and Document Classification (DC). These tasks target information extraction and classification from clinical documents.

Named Entity Recognition: NER is a foundational step for downstream tasks such as knowledge base construction, cohort identification, and automated summarization. This task involves iden-

tifying mentions of clinically relevant concepts such as diseases and medications in free-text narratives. We adapt MedTxt corpus (Yada et al., 2024) to cover multiple document types, including radiology reports, nursing records, and case reports.

Relation Extraction: Relation Extraction aims to identify associations between clinical entities, such as body part–disease links or temporal relations among events. Accurate extraction of such relations is critical for constructing structured patient histories and supporting clinical decision-making. Based on MedTxt corpus (Yada et al., 2024), we incorporate relation types for entity–entity relation and temporal relation.

Document Classification: This task requires assigning pre-defined categories to full clinical documents. We adapt two subtasks: (i) classifying whether a document reports an adverse drug event (ADE), and (ii) predicting tumor staging following the TNM system (e.g., T2a, N0, M0) derived from Nakamura et al. (2025b). ADE classification is important for pharmacovigilance and drug safety monitoring (McMaster et al., 2023), while TNM staging classification reflects clinically meaningful categorization used for prognosis and treatment planning (Nakamura et al., 2025b).

3. Corpus Construction

To establish the foundation of **J-ClinicalBench**, we construct a publicly available Japanese clinical corpus. Motivated by the scarcity of Japanese clin-

Discharge summary (Hospital course)	Progress note
<p>高血圧性心不全の診断で利尿薬、降圧薬調整を開始した。また、二次性高血圧の精査を開始した。第2病日は尿量流出良好であり、酸素化は改善がみられ、食事を開始した。内服も可能となったためARB,MRA,SGLT2を開始した。降圧は不十分であり、第3病日にニフェジピンを導入した。全身状態が安定してきたため第4病日にβ遮断薬を導入した。第5病日に酸素需要はなくなった。二次性高血圧について精査を行った血液検査では明らかな異常はみられず代謝異常は積極的に疑わない。腎動脈の狭窄は否定できず退院後造影CTを検討している。β遮断薬投与したが安定したため第7病日に増量した。塩分摂取は過多であったことより、栄養指導を家族へ行い食事改善を促した。内服に切り替えたが全身状態増悪なく安定していたため、第10病日に退院とした。</p> <p>(A patient was admitted with hypertensive heart failure, and treatment was initiated with diuretics and antihypertensives, alongside an evaluation for secondary hypertension. On hospital day 2, urine output improved, oxygenation stabilized, and oral intake began, allowing the introduction of ARB, MRA, and SGLT2 inhibitors. Because blood pressure control was still inadequate on day 3, nifedipine was added; on day 4, a β-blocker was started. By day 5, oxygen was no longer required. Tests for secondary hypertension showed no significant abnormalities; renal artery stenosis was not ruled out, and contrast CT was scheduled post-discharge. As the patient remained stable, the β-blocker dose was increased on day 7. Dietary counseling was given due to high salt intake. The patient remained stable and was discharged on hospital day 10.)</p>	<p>S: 少し楽になりました。 O: BP 140/78mmHg HR 90bpm BW 85kg SpO2 95%(6Lマスク) 尿量 1000ml/6h 心エコー EF 70%(eyeball), 有意な弁膜症なし, IVC 22mm RV(-) A/P: #高血圧性心不全 #HT ラシックス1Aで尿量流出良好。胸水貯留は著明であり、2A2xで行っていく。血圧高値であり、外来でニトロを導入し現在8mL/hで安定している。明日、二次性高血圧の精査を含めたL/Dを提出予定。内服は明日朝よりアジルバ20mgから開始。酸素投与量が減り食事開始すればMRA, SGLT2を開始する予定。退院までにβ blockerを導入する。</p> <p>(S: I am feeling slightly better. O: BP 140/78 mmHg, HR 90 bpm, BW 85 kg, SpO2 95% (6 L mask), urine output 1000 ml/6 h Echocardiography: EF 70% (eyeball), no significant valvular disease, IVC 22 mm, RV(-) A/P: #Hypertensive heart failure #HTN The patient is on Lasix 1 A with good urine output; due to significant pleural effusion, the dose will be increased to 2 A twice daily. Blood pressure is still high, so nitroglycerin infusion (8 mL/h) was added as an outpatient measure. Tomorrow, labs and diagnostics will be submitted, including a workup for secondary hypertension. Oral therapy will start with Azilva 20 mg in the morning. Once oxygen demand lessens and oral intake becomes possible, MRA and SGLT2 inhibitors will be begun. A β-blocker will be introduced before discharge.)</p>
<p>MedDRA code: Preferred Term</p> <p>10020823: 高血圧性心疾患 (Hypertensive heart disease) 10039834: 二次性高血圧 (Secondary hypertension) 10033317: 酸素飽和度異常 (Oxygen saturation abnormal) 10038378: 腎動脈狭窄症 (Renal artery stenosis)</p>	<p>Medications (drug names)</p> <p>フロセミド (Furosemide) ニトログリセリン (Nitroglycerin) アジルサルタン (Azilsartan)</p>

Table 2: Examples of the hospital course section from a discharge summary and a corresponding progress note for a patient with hypertensive heart failure. English translations are provided in brackets. The progress note is derived from the discharge summary by incorporating additional clinical details, such as laboratory values, and by adapting the realistic clinical documentation style, including the use of abbreviations. At the bottom, we provide example annotations for Clinical Code Assignment (CCA) using the discharge summary and Medication Recommendation (MR) using the progress notes.

ical corpora, we first create 227 **expert-authored clinical documents** including progress notes, discharge summaries, and referral notes (Sect. 3.1). Building on these documents, we newly **created datasets** that reflect practical clinical tasks (Sect. 3.2).

3.1. Expert-authored Clinical Documents

We recruited licensed physicians to author synthetic clinical documents, namely, discharge summaries, progress notes and referral notes, aiming to emulate authentic in-hospital documentation while mitigating privacy concerns associated with real patient data. These documents were designed to capture the linguistic and stylistic characteristics of Japanese clinical notes.

Discharge Summaries: We first reconstructed 30 discharge summaries based on clinical cases from the Japanese national board examination. An example snippet from a discharge summary (hospital course) for the case of hypertensive heart failure is provided in the Table 2. Since the original exam questions do not include full hospitalization details, physicians supplemented missing information (e.g., discharge follow-up) and reformatted the content into plausible discharge summaries.

Progress Notes: For each discharge summary, four physicians created multiple progress notes (3 to 4 on average and 97 in total) to reflect longitudinal care (example in Table 2). Authors were instructed to (i) supplement omitted details such as laboratory values and clinical course, and (ii) ensure that, when combined, the progress notes comprehensively covered the content of the corresponding discharge summary. These notes were expected to resemble realistic clinical documentation by incorporating abbreviations and omissions, rather than direct copy-paste from the discharge summary.

Referral Notes: To simulate inter-specialty communication, ten physicians authored 100 referral notes based on discharge summaries, each tailored to a specific medical specialty (e.g., surgery, emergency department). These notes also reflect realistic clinical documentation, including the use of abbreviations commonly found in clinical notes.

3.2. Created Datasets

Below, we outline five newly created datasets included in **J-ClinicalBench**. In addition to the expert-authored clinical documents introduced in Sect. 3.1, we incorporated other document types

such as radiology reports, nursing records, case reports, and incident reports (Yada et al., 2024; Nakamura et al., 2025b) derived from existing work to further diversify the datasets.

Clinical Code Assignment: We annotated progress notes, discharge summaries, and nursing records with medical codes corresponding to symptoms, diseases, and clinical conditions mentioned in the documents. Specifically, we assigned Japanese Preferred Terms (PT) from the MedDRA ontology. An example is shown in Table 2. The resulting dataset is composed of 200 documents in total with 403 unique MedDRA codes annotated. The annotation process was conducted in two stages: an initial round by an experienced medical annotator, followed by validation from professors of medicine. The first-stage annotations achieved an accuracy of 0.80, as measured during the validation stage, and the corrected labels were incorporated into the final dataset.

Medication Recommendation: We annotated progress notes and discharge summaries with possible prescription medications (an example in Table 2). A licensed pharmacist performed the annotation task, assigning generic drug names that could reasonably be prescribed at the time of discharge or during the corresponding visit. The annotation was guided by clinical practice guidelines and professional experience. In total, 113 annotated documents were obtained with 310 unique drug names.

Hospitalization Summarization: We constructed two types of hospitalization summarization datasets: HS-PN and HS-DS. HS-PN aims to generate the hospitalization course and discharge status sections of a discharge summary from four to five SOAP-format progress notes as input. HS-DS, on the other hand, targets the generation of a referral note conditioned on a discharge summary and physician-specific attributes, including clinical experience and medical specialty. The HS-PN dataset consists of 20 discharge summaries, while HS-DS contains 100 reference referral notes.

Incident Report Summarization: Incident reports document the details of accidents or near-miss events that may pose potential risks in the future. Because the narrative descriptions in the reports are often lengthy, a concise summary is usually prepared manually. To this end, we constructed datasets in which the input consists of full incident descriptions, and the output is a short, phrase-level summary. Summaries are created for 100 incident reports by medical experts. Below is an example.

Incident Report: オキシコドン10mg (2錠)を9:00と21:00に定期投与予定だったが、日勤で9:00分を投与していないことに10:10に気づいた。処方カレンダーは確認済みだったが、検温後に判明。

(Oxycodone 10 mg, 2 tablets, was scheduled for regular administration at 9:00 and 21:00. During the day shift, it was noticed at 10:10 that the 9:00 dose had not been administered.)

Summary: オキシコドンの投与忘れ
(Missed administration of oxycodone.)

Clinical Text Simplification: We created 100 simplified documents, including progress notes, radiology reports, and case reports, rewritten in clearer, layperson-oriented language. Two nurses with more than 10 years of experience performed the annotation, following instructions that emphasized simplification rather than summarization. In total, 40 progress notes, 45 case reports, and 15 radiology reports were simplified. An example of a part of a simplified progress note is provided below.

Progress note: TKA術後6POD。関節の熱感・腫脹なし。ROMは問題なし。

(TKA POD6. No warmth or swelling of the joint. ROM is normal.)

Simplification: 人工膝関節全置換術（傷んだ膝の関節を人工関節に置き換える手術）術後6日目。関節は熱くなく、腫れもなし。関節の動く範囲に問題なし。

(Total knee arthroplasty (surgery to replace a damaged knee joint with an artificial joint). Postoperative day 6. The joint is neither warm nor swollen. The range of joint motion is normal.)

4. Baseline Performance

4.1. Settings

We establish baseline performance by evaluating both proprietary models, namely GPT-4o (Achiam et al., 2023) and GPT-5 (OpenAI, 2025) and a set of open-source Japanese LLMs. The open-source models include:

Llama-3¹: a general-purpose model from Meta, not specifically specialized for Japanese, used as a baseline for comparison (AI@Meta, 2024).

Swallow²: based on Llama-3, Swallow is continually pre-trained using large Japanese corpora to

¹<https://huggingface.co/meta-llama/Meta-Llama-3-70B-Instruct>

²<https://huggingface.co/tokyotech-llm/Llama-3-Swallow-70B-Instruct-v0.1>

Example Prompt (original Japanese)	Example Prompt (Translated in English)
<p>### 指示: あなたのタスクは、提供された電子カルテ文書の内容に基づいて、処方される可能性のある薬剤をすべてリストアップすることです。...</p> <p>### 入力: S 薬が効いてきたみたいで痛みは弱くなりました。 O BP 144/82mmHg SpO 99%(RA) A/P #右緊張性気胸 ...</p> <p>選択肢: ['ミダゾラム', 'レバミピド', 'ニコランジル'...]</p> <p>### 出力: ['セレコキシブ', 'ケトプロフェン', 'レバミピド'...]</p>	<p>### Instruction: Your task is, based on the contents of the provided electronic medical record document, to list all medications that might be prescribed...</p> <p>### Input: S It seems the medicine has begun working; the pain has lessened. O BP 144/82 mmHg SpO 99% (RA) A/P #Right tension pneumothorax ...</p> <p>Choices: ['Midazolam', 'Rebamipide', 'Nikorandil'...]</p> <p>### Output: ['Celecoxib', 'Ketoprofen', 'Rebamipide'...]</p>

Table 3: Example prompt for Clinical Code Assignment task.

Model	Reasoning			Generation				Understanding					
	ExamQA	CCA	MR	HS-PN	HS-DS	IRS	CTS	NER-D	NER-Med	RE-Ent	RE-Time	DC-TMN	DC-ADE
GPT-4o	92.4	44.1	28.6	33.7	35.3	38.7	50.8	41.2	51.2	76.8	69.7	80.0	86.4
GPT-5	94.5	59.1	33.7	29.1	25.6	34.1	45.4	33.2	64.9	78.4	78.0	62.1	86.2
Llama-3	54.5	17.7	8.2	27.6	34.7	29.4	54.3	27.9	17.3	75.2	32.7	68.1	91.3
Swallow	64.2	11.8	7.7	31.6	34.9	36.5	57.9	32.4	27.4	69.4	52.7	67.7	91.8
MedSwallow	80.3	15.6	11.2	31.1	35.9	37.6	55.0	38.5	33.6	69.2	59.5	71.8	92.0
llm-jp	61.3	4.4	8.8	32.4	34.6	38.0	51.8	15.5	15.8	45.7	36.6	61.1	87.2
SIP-jmed-llm	76.3	8.4	8.9	25.1	34.0	41.5	51.3	17.6	16.7	60.8	46.4	49.6	79.2

Table 4: Performance comparison across tasks in J-ClinicalBench.

enhance Japanese-language understanding and generation (Fujii et al., 2024).

MedSwallow³: extends Swallow by further pre-training on medical domain corpora, in order to better cover clinical and biomedical terminology.

LLM-jp⁴: trained from scratch in Japanese (rather than adapted from a multilingual or English-centric base), giving strong native language capacity (Nakamura et al., 2025a).

SIP-jmed-llm⁵: developed under the SIP-jmed-llm initiative, this model builds upon LLM-jp by incorporating medical corpora in its continual pre-training and instruction tuning.

The Llama-3, Swallow, and MedSwallow models are 70B-parameter models, while LLM-jp and SIP-jmed-llm are 8×13B (MoE) models. All of them were evaluated with 4-bit quantization for inference. For every task, prompts were designed as a combination of a task-specific instruction and a one-shot example. In CCA and MR, the model was additionally provided with a set of candidate codes or drugs and instructed to select all appropriate options (see Table 3 for an illustration).

4.2. Results

Table 4 summarizes model performance across all tasks in **J-ClinicalBench**. Overall, LLMs demonstrate strong performance in ExamQA, RE, and

DC. In contrast, tasks requiring complex reasoning, such as CCA, MR, and NER remain challenging for all models, consistent with findings from prior work (Chen et al., 2025; Liu et al., 2024; Boyle et al., 2023). Proprietary models (GPT-4o, GPT-5) achieve the strongest overall results, reflecting their high capability across diverse clinical tasks. Performance in the generation tasks shows less variation, with proprietary and open-source models achieving comparable results. Notably, GPT-5 exhibits lower performance on tasks such as HS-DS and HS-PN. We provide a detailed analysis of the potential reasons for this observation in Sect. 5.2.

Among the open-source models, Swallow, which incorporates continual pre-training on Japanese corpora, occasionally surpasses its base model, Llama-3. The improvements are more consistent for MedSwallow, which outperforms Llama-3 in 9 out of 13 settings. A similar trend is observed for SIP-jmed-LLM, which outperforms LLM-jp in 8 out of 13 settings. These results indicate that pre-training on medical corpora yields measurable gains, particularly for tasks that rely heavily on domain-specific knowledge.

5. Discussion

Based on our evaluation results, we discuss the challenges of current LLMs for clinical applications (Sect. 5.1), highlighting challenges such as limited reasoning and task-specific knowledge gaps. We also conduct a human evaluation of hospitalization summarization to assess model-generated summaries (Sect. 5.2).

³<https://huggingface.co/pfnet/Llama3-Preferred-MedSwallow-70B>

⁴<https://huggingface.co/llm-jp/llm-jp-3-8x13b-instruct3>

⁵<https://huggingface.co/SIP-med-LLM/SIP-jmed-llm-2-8x13b-OP-instruct>

Japanese	English Translation
<p>S: おはようございます。熱も下がりました。尿も少しずつ出ています。 O: 体温37.0°C。脈拍80/分。整。血圧126/60mmHg。呼吸数16/分。SpO2 99% (room air)。 白血球9,800、CRP 7.2mg/dL。 A/P: 昨日より抗菌薬治療を開始した。解熱されており、炎症反応も改善傾向を認める。培養結果を確認していく。 尿閉の原因となる薬剤を中止しており、症状も改善を認める。これらの薬剤は中止継続とする。</p> <p>Annotations Gold: ["尿閉", "白血球数増加", "C-反応性蛋白増加", "炎症"] GPT-5: ["尿閉", "白血球数増加", "C-反応性蛋白増加", "発熱", "炎症"] MedSwallow: ["発熱", "炎症"]</p>	<p>S: Good morning. My fever has gone down. Urine output is gradually improving. O: Temp 37.0°C, pulse 80/min regular, BP 126/60 mmHg, RR 16/min, SpO2 99% (room air). WBC 9,800, CRP 7.2 mg/dL. A/P: Antibiotic therapy was initiated yesterday. The fever has subsided, and inflammatory markers show improvement. Will continue to monitor culture results. The medication suspected of causing urinary retention was discontinued, and symptoms have improved. These medications will remain discontinued.</p> <p>Annotations Gold: ["Urinary retention", "Increased white blood cell count", "Increased C-reactive protein", "Inflammation"] GPT-5: ["Urinary retention", "Increased white blood cell count", "Increased C-reactive protein", "Fever", "Inflammation"] MedSwallow: ["Fever", "Inflammation"]</p>

Table 5: Examples of human annotation (Gold), GPT-5 answer and MedSwallow answer in Clinical Code Assignment tasks

5.1. Challenges in Clinical Applications

Consistent with prior work, LLMs achieve strong performance on exam-style QA and classification tasks with predefined answer options, while showing limited ability in open-ended tasks, including CCA, MR, and NER.

Table 5 presents an illustrative CCA example. GPT-5 produced an almost complete set of MedDRA PTs but incorrectly included “fever,” despite explicit instructions to exclude negated symptoms. In comparison, MedSwallow extracted only a subset of entities (“fever” and “inflammation”) and missed clinically important mentions such as laboratory findings. Overall, proprietary models tend to over-predict moderately, while open-source models show more pronounced degradation in tasks that require careful reasoning and integration of clinical knowledge.

A second example demonstrates challenges in NER of medication values:

腹部<m-val>20Gy</m-val>骨盤<m-val>15Gy</m-val>のRadiation施行後当院に転院した。

After receiving radiation therapy of 20 Gy to the abdomen and 15 Gy to the pelvis, the patient was transferred to our hospital.

Here, “Gy” represents the absorbed dose of ionizing radiation. Whether to treat radiation dosage as a distinct entity depends on the annotation scheme. All evaluated models failed to recognize this mention as medication volume, highlighting need for task-specific knowledge in NER tasks.

Taken together, these results indicate that instruction tuning on non-clinical-domain data alone is insufficient for tasks requiring fine-grained clinical reasoning and task-specific knowledge. To improve performance, particularly for open-ended tasks and NER, LLMs require dedicated instruction tuning on clinical corpora (Corbeil et al., 2025) or task-specific fine-tuning (Shimizu et al., 2025).

Model	Consistency	Saliency	Fluency
GPT-4o	3.21	3.79	3.84
GPT-5	3.84	3.79	3.63
llm-jp	2.63	2.89	3.84
SIP-jmed-llm	2.21	2.21	3.37

Table 6: Human evaluation of the HS-PN summarization task. Scores are on a 1-5 Likert scale averaged over summarizations.

5.2. Qualitative Evaluation of Hospitalization Summarization

While automated metrics like ROUGE-L provide a quantitative overview of performance, they may not fully capture the clinical utility of generated text. To this end, we manually compare the expert-authored reference summaries and LLM outputs, conducting a detailed qualitative analysis of the HS-PN summarization task for four models (GPT-4o, GPT-5, llm-jp, and SIP-jmed-LLM). Specifically we evaluated medical consistency (Tang et al., 2023; Croxford et al., 2025), saliency (Liu et al., 2023; Tang et al., 2023) and fluency (Ive et al., 2020; Hiebel et al., 2023) of the generated discharge summaries compared to the expert-authored ones. The assessment was conducted by a clinical expert (nurse) on a set of 20 discharge summaries.

Table 6 summarizes the Likert scores for each evaluation category. In the manual evaluation, GPT-5 achieved the highest mean scores for medical consistency and saliency, despite its relatively low ROUGE-L score shown in Sect. 4.2. The human expert favored outputs that were clinically faithful and correctly normalized abbreviations, even when their surface form deviated from the reference. Table 7 illustrates such an example. The GPT-5’s summary, while stylistically different, was clinically comprehensive and thus received a high consistency score. In contrast, other models omitted critical information (e.g., the post-discharge plan), leading to lower expert scores, yet sometimes achieved higher ROUGE-L scores due to superficial n-gram

Japanese	English Translation
<p>GPT-5 Output ## 入院経過 入院時、黄疸所見と高アンモニア血症 (NH3 180) を伴い、肝硬変が疑われる背景での肝性脳症と判断。誘因として便秘を認め、ラクツロース開始とともにタンパク制限下にアミノレバン投与を行った。入院翌日、排便が得られ意識は悪化なく、NH3 152と軽度改善。経過良好で、食事摂取可能となったためアミノレバンは終了し、分岐鎖アミノ酸製剤をリーバクト内服へ切替。以後も1日1回の排便を維持し、意識は清明、NH3 62まで改善。全身状態安定を確認し、ラクツロース・リーバクト内服継続とし、肝硬変の原因精査 (NAFLD疑い) は外来継続の方針で退院とした。 ## 退院時状況 意識清明・全身状態安定</p> <p>Human Evaluation Consistency: 5 (非常に良い) Saliency: 4 (良い) Fluency: 4 (良い) ROUGE-L: 0.24 (GPT-4o: 0.29, llm-jp: 0.26)</p>	<p>GPT-5 Output ## Hospital Course The patient was admitted with jaundice and hyperammonemia (NH3 180) and diagnosed with hepatic encephalopathy secondary to suspected liver cirrhosis. Constipation was the likely trigger, so lactulose and Aminoleban were started with protein restriction. The following day, the patient had a bowel movement, mental status remained stable, and NH3 improved to 152. As the patient's condition improved and oral intake was possible, Aminoleban was discontinued and switched to oral Livact. Consciousness remained clear with daily bowel movements, and NH3 normalized to 62. The patient was discharged in stable condition on lactulose and Livact, with plans for outpatient investigation into the cause of cirrhosis (suspected NAFLD). ## Status at Discharge Alert; general condition stable.</p> <p>Human Evaluation Consistency: 5 ("Excellent") Saliency: 4 ("Good") Fluency: 4 ("Good") ROUGE-L for this case: 0.24 (GPT-4o: 0.29, llm-jp: 0.26)</p>

Table 7: Case example where GPT-5 receives a high human score but a relatively low ROUGE-L.

overlap.

These results highlight the limitations of n-gram-based metrics such as ROUGE-L, which primarily capture surface-level overlap and can underestimate medically faithful but stylistically divergent summaries, while failing to penalize omissions of critical clinical information. Future work could explore using LLMs themselves as evaluators to better assess summary quality (Shakil et al., 2024) by enabling more reliable assessment of clinical content.

6. Related Work

The rapid progress of LLMs has spurred their adaptation to medical and clinical applications (Woo et al., 2025; Van Veen et al., 2024). Recent studies have demonstrated that domain-adaptive pretraining and instruction tuning on biomedical corpora enhance model performance on medical multiple-choice question (Labrak et al., 2024; Han et al., 2023; Chen et al., 2023; Toma et al., 2023). Also, a growing body of work has focused on systematically benchmarking LLMs across diverse clinical tasks. Recent efforts include large-scale benchmarks (Chen et al., 2025; Liu et al., 2024; Dorfner et al., 2024; Dada et al., 2025), which assess reasoning, generation, and understanding on medical and clinical documents. These benchmarks are supported by rich English-language clinical resources (Johnson et al., 2016, 2023), and evaluation datasets (Adams et al., 2025; Sun et al., 2013).

By contrast, benchmarks in non-English languages remain scarce due to the limited availability of clinical corpora. Although Japanese corpora such as radiology and case reports exist (Nakamura et al., 2025b; Yada et al., 2024), there remains a lack of publicly accessible, authentic clinical notes, such as discharge summaries and progress notes, which are essential for evaluating LLMs in practical scenarios. Due to the lack of corpora, existing efforts for evaluation in Japanese largely rely on translations of English datasets, such as JMed-

Bench (Jiang et al., 2025) and board examination QA tasks (Kasai et al., 2023). However, these translated benchmarks do not adequately reflect the heterogeneity of clinical documentation, which varies across hospitals, specialties, and practitioners (Laparra et al., 2020). Moreover, evaluation focused on exam-style QA tasks overlooks practical aspects of clinical language reasoning (Griot et al., 2025). This gap highlights the need for Japanese clinical corpora and benchmarks derived from expert-authored realistic documents to enable the evaluation of large language models beyond narrowly defined exam-like scenarios and English translations.

7. Conclusion

In this work, we introduced **J-ClinicalBench**, a publicly available benchmark designed to evaluate LLMs on realistic Japanese clinical tasks. By constructing a high-quality corpus of expert-authored clinical documents and datasets spanning five diverse tasks, we address the scarcity of non-English clinical corpora. Covering clinical language reasoning, generation, and understanding tasks, **J-ClinicalBench** provide a resource for assessing LLM utility in Japanese healthcare. Our evaluation demonstrates that proprietary models such as GPT-4o and GPT-5 achieve the strongest overall performance, while open-source models benefit from continual pre-training on Japanese medical corpora. Nevertheless, reasoning-intensive tasks such as clinical code assignment and medication recommendation remain challenging, highlighting persistent gaps in current LLM capabilities and clinical practice. We hope that **J-ClinicalBench** will serve as a foundation for future research, enabling systematic evaluation of both proprietary and open-source models, and fostering the development of domain-specialized LLMs that can better support clinical practice.

8. Limitations

While **J-ClinicalBench** provides a comprehensive benchmark for evaluating LLMs on Japanese clinical tasks, several limitations remain.

First, the performance of models can be sensitive to prompt design, and our study employed relatively simple instruction formats. Although prompt engineering was explored to ensure fairness and reproducibility across models, further work is needed to systematically investigate the impact of prompt formulation, few-shot examples, and task-specific instructions on model behavior. Second, for most summarization tasks, evaluation relied primarily on ROUGE-L as an automated metric. As discussed in Sect. 5.2, ROUGE-based metrics may not adequately capture clinical faithfulness or the preservation of medically salient content. Future work could incorporate LLM-based or human-in-the-loop evaluation frameworks to better assess factual accuracy and clinical utility. Third, the dataset size for certain tasks, e.g., HS-PN, remains limited due to the cost of corpus creation. We plan to expand these datasets in subsequent releases by collaborating with clinical institutions under appropriate privacy safeguards. Finally, our evaluation primarily focused on text-based clinical reasoning and understanding tasks. Extending the benchmark to multimodal clinical data (e.g., imaging reports, laboratory results, or structured EHR data) would provide a more comprehensive assessment of LLM capabilities in real-world clinical scenarios.

9. Acknowledgements

This work was supported by Cross-ministerial Strategic Innovation Promotion Program (SIP) on "Integrated Health Care System" Grant Number JPJ012425 and JST, CREST Grant Number JP-MJCR22N1, Japan.

10. Bibliographical References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Lisa Adams, Felix Busch, Tianyu Han, Jean-Baptiste Excoffier, Matthieu Ortala, Alexander Löser, Hugo JWL Aerts, Jakob Nikolas Kather, Daniel Truhn, and Keno Bresssem. 2025. Longhealth: A question answering benchmark with long clinical documents. *Journal of Healthcare Informatics Research*, pages 1–17.

AI@Meta. 2024. [Llama 3 model card](#).

Joseph S Boyle, Antanas Kascenas, Pat Lok, Maria Liakata, and Alison Q O’Neil. 2023. Automated clinical coding using off-the-shelf large language models. *arXiv preprint arXiv:2310.06552*.

Elliot G Brown, Louise Wood, and Sue Wood. 1999. The medical dictionary for regulatory activities (meddra). *Drug safety*, 20(2):109–117.

Qingyu Chen, Yan Hu, Xueqing Peng, Qianqian Xie, Qiao Jin, Aidan Gilson, Maxwell B Singer, Xuguang Ai, Po-Ting Lai, Zhizheng Wang, et al. 2025. Benchmarking large language models for biomedical natural language processing applications and recommendations. *Nature communications*, 16(1):3280.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. 2023. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

Jean-Philippe Corbeil, Amin Dada, Jean-Michel Attendu, Asma Ben Abacha, Alessandro Sordoni, Lucas Caccia, Francois Beaulieu, Thomas Lin, Jens Kleesiek, and Paul Vozila. 2025. [A modular approach for clinical SLMs driven by synthetic data with pre-instruction tuning, model merging, and clinical-tasks alignment](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19352–19374, Vienna, Austria. Association for Computational Linguistics.

Emma Croxford, Yanjun Gao, Nicholas Pellegrino, Karen Wong, Graham Wills, Elliot First, Frank Liao, Cherodeep Goswami, Brian Patterson, and Majid Afshar. 2025. Current and future state of evaluation of large language models for medical summarization tasks. *npj Health Systems*, 2(1):6.

Amin Dada, Osman Alperen Koraş, Marie Bauer, Jean-Philippe Corbeil, Amanda Butler Contreras, Constantin Marc Seibold, Kaleb E Smith, Julian Friedrich, and Jens Kleesiek. 2025. [Does biomedical training lead to better medical performance?](#) In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 46–59, Vienna, Austria and virtual meeting. Association for Computational Linguistics.

Kerstin Denecke and Helmut Paula. 2024. Analysis of critical incident reports using natural language processing. In *dHealth 2024*, pages 1–6. IOS Press.

- Felix J Dorfner, Amin Dada, Felix Busch, Marcus R Makowski, Tianyu Han, Daniel Truhn, Jens Kleesiek, Madhumita Sushil, Jacqueline Lammer, Lisa C Adams, et al. 2024. Biomedical large languages models seem not to be superior to generalist models on unseen medical data. *arXiv preprint arXiv:2408.13833*.
- Maarten Falter, Dries Godderis, Martijn Scherrenberg, Sevda Ece Kizilkilic, Linqi Xu, Marc Mertens, Jan Jansen, Pascal Legroux, Hanne Kindermans, Peter Sinnaeve, Frank Neven, and Paul Dendale. 2024. [Using natural language processing for automated classification of disease and to identify misclassified icd codes in cardiac disease](#). *European Heart Journal - Digital Health*, 5(3):229–234.
- Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. 2024. Continual pre-training for cross-lingual llm adaptation: Enhancing japanese language capabilities. In *Proceedings of the First Conference on Language Modeling*, COLM, page (to appear), University of Pennsylvania, USA.
- Emily Gesner, Priscilla Gazarian, and Patricia Dykes. 2019. The burden and burnout in documenting patient care: an integrative literature review. *MEDINFO 2019: Health and wellbeing e-networks for all*, pages 1194–1198.
- Maxime Griot, Jean Vanderdonckt, Demet Yuksel, and Coralie Hemptinne. 2025. [Pattern recognition or medical knowledge? the problem with multiple-choice questions in medicine](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5321–5341, Vienna, Austria. Association for Computational Linguistics.
- Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressen. 2023. MedAlpaca—an open-source collection of medical conversational AI models and training data. *arXiv preprint arXiv:2304.08247*.
- Nicolas Hiebel, Olivier Ferret, Karen Fort, and Aurélie Névéol. 2023. [Can synthetic text help clinical named entity recognition? a study of electronic health records in French](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2320–2338, Dubrovnik, Croatia. Association for Computational Linguistics.
- Julia Ive, Natalia Viani, Joyce Kam, Lucia Yin, So-main Verma, Stephen Puntis, Rudolf N Cardinal, Angus Roberts, Robert Stewart, and Sumithra Velupillai. 2020. Generation and evaluation of artificial mental health records for natural language processing. *npj Digital Medicine*, 3(1):69.
- Junfeng Jiang, Jiahao Huang, and Akiko Aizawa. 2025. [JMedBench: A Benchmark for Evaluating Japanese Biomedical Large Language Models](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5918–5935, Abu Dhabi, UAE. Association for Computational Linguistics.
- Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. 2023. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1.
- Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3(1):1–9.
- Jungo Kasai, Yuhei Kasai, Keisuke Sakaguchi, Yutaro Yamada, and Dragomir Radev. 2023. Evaluating GPT-4 and ChatGPT on Japanese medical licensing examinations. *arXiv preprint arXiv:2303.18027*.
- Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickaël Rouvier, and Richard Dufour. 2024. BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5848–5864.
- Egoitz Laparra, Steven Bethard, and Timothy A Miller. 2020. Rethinking domain adaptation for machine learning over clinical language. *JAMIA open*, 3(2):146–150.
- Fenglin Liu, Zheng Li, Hongjian Zhou, Qingyu Yin, Jingfeng Yang, Xianfeng Tang, Chen Luo, Ming Zeng, Haoming Jiang, Yifan Gao, Priyanka Nigam, Sreyashi Nag, Bing Yin, Yining Hua, Xuan Zhou, Omid Rohanian, Anshul Thakur, Lei Clifton, and David A. Clifton. 2024. [Large language models are poor clinical decision-makers: A comprehensive benchmark](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13696–13710, Miami, Florida, USA. Association for Computational Linguistics.
- Yixin Liu, Alex Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq

- Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023. [Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4140–4170, Toronto, Canada. Association for Computational Linguistics.
- Christopher McMaster, Julia Chan, David F.L. Liew, Elizabeth Su, Albert G. Frauman, Wendy W. Chapman, and Douglas E.V. Pires. 2023. [Developing a deep learning natural language processing algorithm for automated reporting of adverse drug reactions](#). *Journal of Biomedical Informatics*, 137:104265.
- Taishi Nakamura, Takuya Akiba, Kazuki Fujii, Yusuke Oda, Rio Yokota, and Jun Suzuki. 2025a. [Drop-upcycling: Training sparse mixture of experts with partial re-initialization](#). In *Proceedings of the Thirteenth International Conference on Learning Representations*.
- Yuta Nakamura, Koji Fujimoto, Jonas Kluckert, Michael Krauthammer, Jun Kanzawa, Akira Katayama, Tomohiro Kikuchi, Ryo Kurokawa, Wataru Gono, Peitao Han, et al. 2025b. NTCIR-18 RadNLP 2024 Overview: Dataset and Solutions for Automated Lung Cancer Staging. In *Proceedings of the NTCIR-18 Conference*.
- Tuan Dung Nguyen, Thanh Trung Huynh, Minh Hieu Phan, Quoc Viet Hung Nguyen, and Phi Le Nguyen. 2024. [CARER - ClinicAI reasoning-enhanced representation for temporal health risk prediction](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10392–10407, Miami, Florida, USA. Association for Computational Linguistics.
- OpenAI. 2025. [Introducing gpt-5](#). OpenAI.
- Hassan Shakil, Atqiya Munawara Mahi, Phuoc Nguyen, Zeydy Ortiz, Jugal Kalita, and Mamoun T Mardini. 2024. [Evaluating text summaries generated by large language models using openai's gpt](#). In *2024 International Conference on Machine Learning and Applications (ICMLA)*, pages 951–956. IEEE.
- Seiji Shimizu, Tomohiro Nishiyama, Hiroyuki Nagai, Shoko Wakamiya, and Eiji Aramaki. 2025. [Toward cross-hospital deployment of natural language processing systems: Model development and validation of fine-tuned large language models for disease name recognition in japanese](#). *JMIR Med Inform*, 13:e76773.
- Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. 2013. [Evaluating temporal relations in clinical text: 2012 i2b2 challenge](#). *Journal of the American Medical Informatics Association*, 20(5):806–813.
- Liyan Tang, Zhaoyi Sun, Betina Idray, Jordan G Nestor, Ali Soroush, Pierre A Elias, Ziyang Xu, Ying Ding, Greg Durrett, Justin F Rousseau, et al. 2023. [Evaluating large language models on medical evidence summarization](#). *npj Digital Medicine*, 6(1):158.
- Augustin Toma, Patrick R Lawler, Jimmy Ba, Rahul G Krishnan, Barry B Rubin, and Bo Wang. 2023. [Clinical camel: An open expert-level medical language model with dialogue-based knowledge encoding](#). *arXiv preprint arXiv:2305.12031*.
- Dave Van Veen, Cara Van Uden, Louis Blanke-meier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, et al. 2024. [Adapted large language models can outperform medical experts in clinical text summarization](#). *Nature medicine*, 30(4):1134–1142.
- Elizabeth Geena Woo, Michael C Burkhart, Emily Alsentzer, and Brett K Beaulieu-Jones. 2025. [Synthetic data distillation enables the extraction of clinical information at scale](#). *npj Digital Medicine*, 8(1):267.
- Chenghao Xiao, Kun Zhao, Xiao Wang, Siwei Wu, Sixing Yan, Tomas Goldsack, Sophia Ananiadou, Noura Al Moubayed, Liang Zhan, William Cheung, and Chenghua Lin. 2025. [Overview of the biolaysumm 2025 shared task on lay summarization of biomedical research articles and radiology reports](#). In *The 24th Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, Vienna, Austria. Association for Computational Linguistics.
- Shuntaro Yada, Yuta Nakamura, Shoko Wakamiya, and Eiji Aramaki. 2024. [Cross-lingual natural language processing on limited annotated case/radiology reports in English and Japanese: Insights from the Real-MedNLP workshop](#). *Methods of Information in Medicine*, 63(05/06):145–163.

11. Language Resource References

- Yuta Nakamura and Koji Fujimoto and Jonas Kluckert and Michael Krauthammer and Jun Kanzawa and Akira Katayama and Tomohiro Kikuchi and Ryo Kurokawa and Wataru Gono and Peitao Han and others. 2025. *NTCIR-18 RadNLP 2024*

Overview: Dataset and Solutions for Automated Lung Cancer Staging. NTCIR.

Shuntaro Yada and Yuta Nakamura and Shoko Wakamiya and Eiji Aramaki. 2024. *Cross-lingual natural language processing on limited annotated case/radiology reports in English and Japanese: Insights from the Real-MedNLP workshop.* Georg Thieme Verlag KG.