A Japanese Language Model and Three New Evaluation Benchmarks for Pharmaceutical NLP

Anonymous ACL submission

Abstract

We present a Japanese domain-specific language model for the pharmaceutical field, developed through continual pretraining on 2 billion Japanese pharmaceutical tokens and 8 billion English biomedical tokens. To enable rigorous evaluation, we introduce three new benchmarks: **YakugakuQA**, based on national pharmacist licensing exams; **NayoseQA**, which tests cross-lingual synonym and terminology normalization; and **SogoCheck**, a novel task designed to assess consistency reasoning between paired statements. We evaluate our model against both open-source medical LLMs and commercial models, including GPT-4o. Results show that our domain-specific model outperforms existing open models and achieves competitive performance with commercial ones, particularly on terminology-heavy and knowledge-based tasks. Interestingly, even GPT-4o performs poorly on SogoCheck, suggesting that cross-sentence consistency reasoning remains an open challenge. Our benchmark suite offers a broader diagnostic lens for pharmaceutical NLP, covering factual recall, lexical variation, and logical consistency. This work demonstrates the feasibility of building practical, secure, and cost-effective language models for Japanese domain-specific applications, and provides reusable evaluation resources for future research in pharmaceutical and healthcare NLP. Our model, codes, and datasets will be released upon acceptance.

1 Introduction

Large Language Models (LLMs) have achieved remarkable performance across a wide range of natural language processing (NLP) tasks. However, their effectiveness remains limited in domain-specific settings such as manufacturing, finance, and medicine (Islam et al., 2023; Hager et al., 2024; Zhang et al., 2024), where deep contextual understanding and precise terminology handling are required. In these domains, general-purpose LLMs often fall short due to inadequate domain knowledge and difficulty handling complex or specialized queries. Moreover, while domain-specific fine-tuning can enhance surface-level performance, it has been shown that this does not necessarily lead to genuine knowledge acquisition (Zhou et al., 2023).

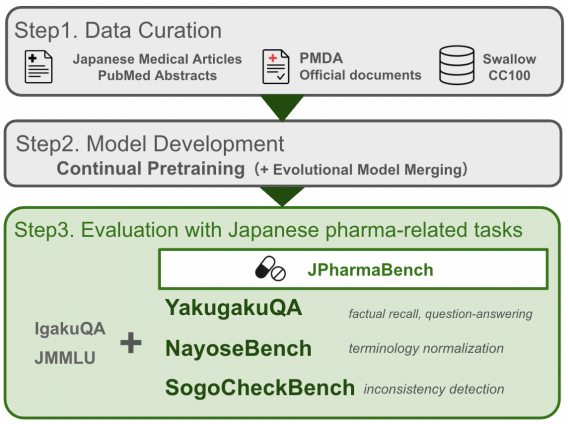


Figure 1: **JPharmatron and JPharmaBench.** The pipeline for data curation, continued pretraining, and evaluation of JPharmatron.

The pharmaceutical domain is no exception. In particular, the Japanese pharmaceutical industry faces significant administrative overhead in tasks such as document preparation, verification, and regulatory compliance—often governed by standards such as GMP (Chaloner-Larsson et al., 1999) and ICH guidelines[1]. Despite these challenges, little work has been done to develop LLMs tailored for pharmaceutical operations, especially in Japanese.

In this work, we present JPharmatron, a Japanese language LLM series specialized for pharmaceutical operations. To build JPharmatron, we perform continual pretraining of the Qwen2.5 (Yang et al., 2024) model using a cu-

[1] https://www.ich.org/page/ich-guidelines

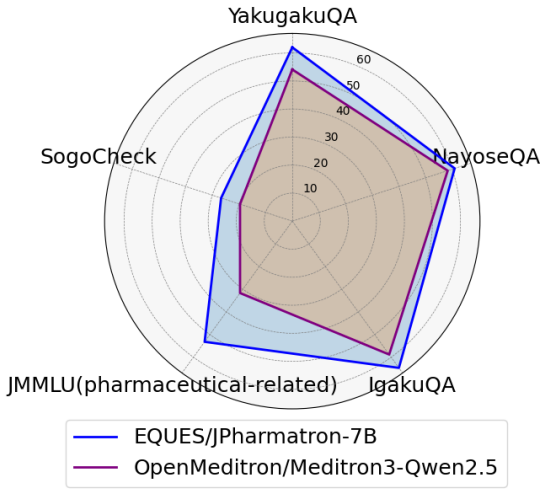


Figure 2: **Performance Comparison with Meditron.** JPharmatron consistently achieves higher scores than Meditron across JPharmaBench, IgakuQA, and JMMLU.

rated corpus consisting of Japanese pharmaceutical journals, web resources, and synthetic data (Appendix C). Unlike prior work focusing on drug discovery (Chaves et al., 2024; Tsuruta et al., 2024), our model targets real-world operational tasks, such as document standardization and terminology normalization.

To evaluate pharmaceutical reasoning and generation capabilities, we introduce three novel benchmarks:

(1) YakugakuQA (§3.2): a multiple-choice QA dataset based on the Japanese National Pharmacist Examination;

(2) NayoseQA (§3.3): a paraphrasing benchmark for standardizing drug names and active substances;

(3) SogoCheck (§3.4): a document consistency-check task reflecting real administrative workflows.

These benchmarks, collectively referred to as JPharmaBench, are designed to reflect practical scenarios encountered in pharmaceutical companies, particularly in regulatory and clerical operations. To the best of our knowledge, this is the first benchmark suite for evaluating LLMs in Japanese pharmaceutical applications.

We evaluate JPharmatron using in-context learning across JPharmaBench and two existing benchmarks additionally. Without task-specific fine-tuning, our model outperforms competitive LLMs including Meditron (§2.2), showing gains of 7.9% on YakugakuQA (Ours) and 5.9% on IgakuQA (Kasai et al., 2023). These results suggest that domain-adaptive continual pretraining can significantly enhance LLM performance in specialized pharmaceutical and medical settings.

Our contributions are threefold:

- We introduce the first LLMs and evaluation benchmarks specifically designed for Japanese pharmaceutical NLP.
- We develop tasks aligned with real-world workflows, ensuring practical relevance in pharmaceutical operations.
- We provide a complete methodology — from data collection to evaluation — that serves as a replicable and secure framework for domain-specific LLM development in regulated industries.

2 Related works

2.1 Domain-specific LLMs and benchmarks in healthcare

With the emergence of GPTs (Radford et al., 2018; Brown et al., 2020), domain-specific adaptations for healthcare have rapidly gained attention. Several English-centric LLMs have been developed to infuse medical knowledge into general-purpose models. For instance, Med-PaLM 2 (Singhal et al., 2023b), a specialized version of PaLM 2 (Anil et al., 2023), is fine-tuned on curated medical datasets and achieves performance comparable to medical professionals on exams.

Benchmarking has evolved in parallel. MultiMedQA (Singhal et al., 2023a) combines datasets to evaluate both factual knowledge and clinical reasoning. Other benchmarks, such as MedQA (Jin et al., 2020) and the medical subset of MMLU (Hendrycks et al., 2021), are commonly used to assess instruction-following and medical understanding.

In the Japanese context, GPT-style healthcare LLMs are still emerging. Recent projects (Sukeda et al., 2023, 2024a,b) have focused on adapting LLMs for Japanese medical question answering. The standard benchmarks are also being developed (Sukeda, 2024; Jiang et al., 2024), exemplified by IgakuQA (Kasai et al., 2023), based on the Japanese national medical licensing exam.

These developments in both English and Japanese highlight a global trend toward aligning LLMs with clinical expertise across languages and contexts. While significant progress has

been made in the medical field, efforts in the pharmaceutical domain remain limited, and the few existing models (Chen et al., 2024; Chaves et al., 2024) are not publicly available.

2.2 Meditron

Among existing domain-specific medical LLMs, Meditron (Chen et al., 2023) is particularly relevant to our work. Meditron is a family of open-source LLMs of 7B and 70B, built upon LLama2 (Touvron et al., 2023), and adapted with medical continual pretraining and supervised fine-tuning using curated English medical corpus. It demonstrates strong performance in MedQA (Jin et al., 2020), making it a prominent example of an open medical LLM. The work is further extended by Open Meditron Initiative[2].

In contrast, our work focuses on the Japanese language and the pharmaceutical domain, both of which remain underexplored. With strong performance on YakugakuQA, our model serves as a Japanese-pharmaceutical counterpart to Meditron. This parallel extends to benchmarks as well: Meditron is evaluated on MedQA (Jin et al., 2020), while our model is evaluated on YakugakuQA (ours) and IgakuQA (Kasai et al., 2023), which are all based on national licensing exams in their respective languages and domains.

3 Benchmark construction

Pharmaceutical domain has not received as much attention for LLM applications, resulting in a limited number of evaluation benchmarks, especially in Japanese. When the focus is solely on therapeutics data, a comprehensive benchmark for therapeutics machine learning called the *Therapeutic Data Commons* (Huang et al., 2022) can be applied to LLM evaluations (Chaves et al., 2024). However, the performance of LLMs in the broader pharmaceutical domain has only been evaluated on the North American Pharmacist Licensure Examination (NAPLEX) (Ehlert et al., 2024; Chen et al., 2024), with no evaluations conducted in Japanese. Although MMLU (Hendrycks et al., 2021) and JMMLU (Yin et al., 2024) cover related healthcare domains, neither includes pharmaceutics as a distinct category.

[2] https://huggingface.co/OpenMeditron

> *Which of the following is not an ideal property of a dilute solution? Choose one.*
> 1. *Vapor pressure lowering*
> 2. *Freezing point depression*
> 3. *Boiling point elevation*
> 4. *Surface tension reduction*
> 5. *Osmotic pressure*

Figure 3: **An example question from the Japanese National Pharmacist Licensing Examination.** The model is required to output "4" in this case. The question is originally in Japanese, but translated into English by ChatGPT for readability.

3.1 Overview of JPharmaBench

To evaluate language models in the Japanese pharmaceutical domain, we constructed three novel benchmarks, each reflecting a different type of reasoning or knowledge required in real-world pharmaceutical practice: factual recall, terminology normalization, and inconsistency detection (Table 1). All benchmarks are based on publicly available data and are structured as question-answering tasks, making them compatible with various LLMs.

3.2 YakugakuQA: National Licensing Exam

YakugakuQA is a question-answering dataset based on the Japanese national pharmacist licensing examinations (NPLE) administered by the Ministry of Health, Labour and Welfare. As illustrated in Figure 3, each question requires selecting one or two correct answers from five or six choices. As summarized in Table 2, YakugakuQA serves as a pharmaceutical counterpart to IgakuQA.

We have collected the exam data from the past 13 years, from 2012 to 2024. All questions, answers, and commentaries have been obtained from the website *yakugaku lab*[3] and manually processed. The category varies among pharmacy and eight other related areas: pharmacy, pharmacology, chemistry, pathology, hygiene, physics, practice, law, and biology.

Some questions in the NPLE require responses based on a provided image — for example, identifying a chemical reaction depicted in the image. However, such image-based questions are excluded from our experiments, as our study focuses on LLMs designed for text input. The number of

[3] https://yakugakulab.info/

Benchmark	Format	Main Skill	Source	#Examples	Language(s)
YakugakuQA	4-to-6-choice QA	Factual recall	Licensing exams	3,021	Japanese
NayoseQA	5-choice QA	Terminology normalization	KEGG DRUG Database	34,769	Japanese / English
SogoCheck	Sentence pair	Inconsistency detection	Japanese Pharmacopoeia	200	Japanese

Table 1: **An overview of JPharmaBench, the three pharmaceutical benchmarks for evaluation.** Each task is designed to assess different capabilities of LLMs in domain-specific settings.

	English	Japanese
Medicine	MedQA (Jin et al., 2020)	IgakuQA (Kasai et al., 2023)
Pharmacy	NAPLEX (not structured)	**YakugakuQA (Ours)**

Table 2: **National licensing exams.** These are typically used as benchmarks when evaluating domain-specific LLMs in medical-related fields.

questions by year and category used in our experiments is shown in Table 6.

3.3 NayoseQA: Synonym and Terminology Normalization in the Pharmaceutical Domain

NayoseQA is our original benchmark designed to evaluate LLMs' ability to handle lexical variation and term normalization in pharmaceutical texts written in Japanese. The task focuses on resolving different surface forms of the same underlying drug or chemical entity, including:

- Japanese name ↔ English name (e.g., 水 ↔ H2O)
- brand name ↔ generic name (e.g., Ganaton ↔ Itopride hydrochloride)
- chemical name ↔ common name (e.g., Prostaglandin E2 ↔ PGE2)

This type of normalization is commonly referred to as "nayose" in Japanese, a term used in information systems to describe the process of identifying and consolidating records that refer to the same real-world entity. In our context, it involves linguistic and domain-specific reasoning to recognize synonymous terms for pharmaceutical compounds. In real-world pharmaceutical documents and practice in Japan, such variations are common due to regulatory terminology, manufacturer-specific branding, and historical naming conventions. Accurately interpreting and normalizing these variations is essential for drug interaction checks, medical record standardization, and multilingual information retrieval.

Text A: *Storage method: sealed container. Temperature below 25° C. Humidity below 60%.*
Text B: *Storage method: sealed container. Temperature below 26° C. Humidity below 61%.*
Label: Change in temperature and humidity

Figure 4: **A simplest example from SogoCheck.** The numbers are inconsistent across two inputs. Originally in Japanese, but translated for readability.

3.4 SogoCheck: Inconsistency Detection in Paired Pharmaceutical Statements

SogoCheck is a novel benchmark we introduce to evaluate an LLM's ability to detect logical or factual inconsistencies (referred to as "sogo" in Japanese) between two pieces of text in the pharmaceutical domain. Unlike factual question-answering benchmarks, which assess whether a synthetic text contains any factual errors (Zhao et al., 2023), SogoCheck focuses on cross-text consistency. The task is inspired by a common practice in pharmaceutical quality assurance in Japan, where experts conduct consistency reviews to cross-validate information across documents such as package inserts, internal quality assurance logs, and regulatory submissions.

In this task, the model is presented with a pair of short Japanese texts, typically drawn from regulatory documents, drug descriptions, or quality assurance manuals. The model is asked to determine whether the two statements are consistent either explicit or implicit. Some examples are clear-cut (e.g., numerical mistakes, see Figure 4), while others require pharmacological reasoning or recognition of subtle semantic contradictions.

The final dataset includes 200 examples, synthesized with an LLM to balance clarity and realism. This benchmark is particularly valuable because inconsistency detection is crucial in practical workflows such as regulatory review, where conflicting information can lead to severe medical or legal consequences.

4 Model & Training

We developed a domain-specific language model, JPharmatron, through continual pretraining with three different data scales, based on Qwen2.5-7B (Yang et al., 2024), a multilingual open-source language model that also supports Japanese input, and evolutionary merging. This base model was chosen for its strong general performance, multilingual capacity, and availability under a commercially permissible license.

To inject domain-specific knowledge while preserving general language capabilities, we adopted continual pretraining rather than training from scratch. We prepared three variations of the training corpus:

2B tokens: Approximately 2B Japanese tokens sourced from pharmaceutical-related documents such as journal papers and drug package inserts;

10B tokens: The above 2B Japanese tokens combined with an additional 8B English tokens from PubMed Abstracts;

9B tokens: Based on the 10B-token corpus, further augmented with 1.2B tokens from the CC100 multilingual dataset. After removing duplicates, the number of tokens was finally 9B tokens (see Appendix C for details).

Training was conducted using standard autoregressive language modeling objectives with the original tokenizer of Qwen2.5. Table 3 provides an overview of the training configuration and data composition. In addition, model merging was performed to attach instruction-following ability to the model. Further details on data collection, cleaning, and preprocessing pipelines are defered to Appendix C.

We emphasize that our goal was not to outperform proprietary LLMs like GPT-4o, but to develop a practically deployable model as a first baseline that balances accuracy, efficiency, and privacy for real-world use in Japanese pharmaceutical contexts. This lightweight domain adaptation strategy enables enterprises to build specialized models without large-scale resources (§6.2).

Training Settings	
Method	Continual pretraining
Base model	Qwen2.5-7B
Japanese data	2B tokens (pharma-related)
English data	8B tokens (mainly PubMed Abstracts)
Tokenizer	Qwen2.5 tokenizer
Steps	67171
Batch size	16
Optimizer	hybridadam
Learning rate	1.0×10^{-5}
GPU	8 × NVIDIA H100
Framework	Pai-Megatron-Patch
GPU hours	444

Table 3: **Details of model training settings.**

5 Evaluation

5.1 Experimental Setups

We evaluated our domain-specific model against three types of baseline models: (1) a general-purpose Japanese LLM (Swallow series or equivalent), (2) a medical LLM (Meditron)[4], and (3) GPT-4o via the OpenAI API. Evaluation was conducted across three newly proposed benchmarks — YakugakuQA, NayoseQA, and SogoCheck — as well as two existing Japanese medical benchmarks: IgakuQA and a pharmaceutical subset of JMMLU. This setup enables direct comparison with prior work.

To ensure fairness, all models were prompted with consistent formatting (details provided in Appendix B). For multiple-choice questions, models were instructed to select one or more answer options as appropriate, where the accuracy was measured based on exact match.

5.2 Quantitative results

Table 4 shows the accuracy of each model on each benchmark. While GPT-4o achieved the highest accuracy overall, as expected from a frontier commercial LLM, our domain-specific model consistently outperformed both Meditron and the general-purpose Japanese model across all tasks. This highlights the effectiveness of domain-specific continual pretraining in Japanese, and establishes our model as the strongest open alternative for pharmaceutical NLP tasks in the Japanese language.

Breaking down by benchmark, on YakugakuQA, our model achieved an accuracy of 62.0%, outperforming Meditron3-Qwen2.5-7B by 7.9 points. This result suggests that factual pharmaceutical knowledge can be effectively captured through continual pretraining, even without training from scratch. In addition, it suggests that medical domain specialization alone may be

[4]We use Meditron3-Qwen2.5-7B from OpenMeditron for comparison, as the older version (Chen et al., 2023) lacks sufficient Japanese support and our model is also based on Qwen2.5-7B, ensuring a fair evaluation.

insufficient for handling pharmaceutical tasks effectively. The accuracy results by categories are listed in Table 5, along with additional larger models for references: Llama-3.1-Swallow-70B (Fujii et al., 2024), Qwen2.5-72B-Instruct (Yang et al., 2024), and o1-preview via OpenAI API.

In NayoseQA, which tests synonym normalization and cross-lingual terminology mapping, the performance gap between our domain-specific model and the general-purpose model (Llama3.1-Swallow) was surprisingly small. This suggests that the task primarily requires lexical and semantic matching capabilities rather than deep domain-specific pharmaceutical knowledge. While domain adaptation improved performance modestly, it appears that general LLMs with strong multilingual and synonym handling capabilities can already perform well on such terminology normalization tasks. This indicates that future pharmaceutical LLM development efforts may benefit more from enhancing complex reasoning and factual recall abilities rather than focusing solely on terminology alignment.

Finally, SogoCheck proved to be challenging for all models. While one of our models outperformed Meditron by 7.1 points, the absolute accuracy remained low. Notably, even GPT-4o achieved only 39.1% accuracy, suggesting that subtle consistency detection in specialized domains remains an open research challenge. Interestingly, many SogoCheck examples were intentionally designed to be solvable by simple textual comparison — identifying surface-level differences without requiring deep reasoning (see Figure 4). Despite this, LLMs often failed to detect such inconsistencies, indicating that current models still struggle with fine-grained semantic alignment even when superficial textual clues are available. This gap between human intuition and model behavior highlights a critical limitation in today's LLM architectures.

5.3 Error analysis

We analyze the 16.4% of incorrectly answered questions on YakugakuQA to identify common failure patterns and inform future improvements in domain-specific LLMs such as JPharmatron.

Positional Bias. Consistent with previous works (Marchisio et al., 2024; Trung et al., 2024), we observed a positional bias in GPT-4o's responses on YakugakuQA, where the model exhibited a tendency to favor the first answer choice. Specifically, the number of responses selecting option "1" exceeded the total number of questions (Figure 5a), and the error rate for option "1" was the lowest among all choices (Figure 5b).

Single vs. Multiple-Choice Question. GPT-4o exhibited a 4.4% higher error rate on multiple-choice questions compared to single-answer questions (Figure 5c).

Question category. Figure 5d shows that error rates for chemistry and physics are around 25%, while those for biology and pathology are below 10%. This indicates that GPT-4o performs better in biology and pathology, but struggles with calculation-heavy questions in chemistry and physics (Ahn et al., 2024; Li et al., 2024b). The higher performance in biology and pathology may be attributed to the prevalence of fact-based, single-answer questions in these domains. This pattern is commonly observed across various LLMs, as shown in Table 5, and also in JMMLU as shown in Table 8.

Complex questions. Based on the previous observation, we employed Qwen2.5-72B-Instruct (Yang et al., 2024) to annotate questions requiring complex reasoning or calculations, following the LLM-as-a-Judge framework (Li et al., 2024a). Although such questions accounted for fewer than 500 out of approximately 3000, they exhibited an error rate of 34.1% (Figure 5e). These results suggest that top-tier LLMs still struggle with calculation-intensive tasks within the pharmaceutical domain.

6 Discussion

6.1 Impact of our Benchmark Suite

Our benchmark suite is designed to evaluate a diverse range of language capabilities required for pharmaceutical NLP. While prior datasets such as IgakuQA and JMMLU primarily focus on factual recall, our benchmarks target additional competencies that better reflect the demands of real-world pharmaceutical decision-making.

Evaluation results confirm that this broader scope offers meaningful insights. YakugakuQA and NayoseQA showed consistent improvements across most models, suggesting that domain-specific pretraining effectively enhances factual recall and term-level understanding. In contrast, SogoCheck presented a more difficult challenge. Some models showed minor gains, while others

	Model	YakugakuQA	NayoseQA	SogoCheck	IgakuQA	JMMLU
(1)	TinySwallow-1.5B-Instruct	37.2	35.3	3.1	39.0	32.1
	sarashina2.2-3b-instruct	46.2	45.6	0.66	41.6	37.8
	Llama-3-Swallow-8B-Instruct-v0.1	42.6	29.8	-	41.5	20.6
	Llama-3.1-Swallow-8B-Instruct-v0.3	48.2	57.6	-	45.2	44.0
(2)	Meditron3-Qwen2.5-7B	54.1	58.3	19.6	58.8	31.7
(3)	GPT-4o	**83.6**	**86.0**	**39.1**	**86.6**	**79.1**
Ours	JPharmatron-7B /2B tokens	60.7	58.3	12.5	62.3	**55.0**
	JPharmatron-7B /10B tokens	54.8	**62.6**	22.0	60.1	48.7
	JPharmatron-7B /9B tokens	**62.0**	60.9	**26.7**	**64.7**	53.2

Table 4: **Performance of our LLMs in five pharmaceutical-related benchmarks**, compared to (1) a general-purpose Japanese LLM (Swallow series, or equivalent), (2) a medical LLM (Meditron), and (3) GPT-4o. Each value shows the accuracy (%). "-" denotes the lack of instruction-following capability to solve each task. The top two models for each task are highlighted in bold.

Model	Biology	Chemistry	Hygiene	Law	Pathology	Pharmacology	Pharmacy	Physics	Practice	Overall
TinySwallow-1.5B-Instruct	41.1	21.9	34.4	**46.5**	**44.3**	27.8	36.9	32.4	38.0	37.2
sarashina2.2-3b-instruct	46.3	36.7	45.8	**56.2**	**56.6**	37.8	41.5	29.2	48.6	46.2
Qwen2.5-7B-Instruct	**69.1**	18.2	52.9	54.3	**65.0**	46.6	47.4	49.4	55.7	53.9
Meditron3-Qwen2.5-7B	**69.1**	24.0	54.4	57.5	**63.8**	47.4	49.1	45.1	54.0	54.1
Llama-3-Swallow-8B-Instruct-v1	46.0	26.4	45.6	**56.1**	**47.3**	31.8	34.6	30.2	46.5	42.6
Llama-3.1-Swallow-8B-Instruct-v3	56.4	18.8	48.5	**57.5**	**56.9**	42.1	39.4	34.6	49.7	48.2
Llama-3.1-Swallow-70B-Instruct-v1	**81.7**	41.4	71.2	70.0	**82.1**	71.1	66.5	55.5	68.6	70.9
Qwen2.5-72B-Instruct	**89.8**	51.5	72.2	72.5	**84.4**	76.4	68.7	62.8	70.0	73.6
GPT-4o	**94.4**	76.1	80.9	83.4	**92.1**	88.7	81.8	72.6	78.6	83.6
o1-preview	**93.3**	88.3	88.1	83.3	**93.2**	90.8	85.0	89.1	84.5	87.9
JPharmatron-7B /2B tokens	**80.9**	28.4	55.9	66.6	**71.5**	55.7	55.1	55.2	58.6	60.7
JPharmatron-7B /10B tokens	**70.8**	19.3	53.6	57.3	**66.9**	46.2	48.8	51.7	55.3	54.8
JPharmatron-7B /9B tokens	**80.5**	45.7	57.9	63.8	**73.8**	58.4	54.9	51.6	61.3	62.0

Table 5: **Accuracy of YakugakuQA comparison by category.** Each value shows the accuracy (%). The top two categories for each model are highlighted in bold. Most models excel in biology and pathology.

failed to improve. As previously shown, the suprisingly low accuracy of GPT-4o indicates that current LLMs — even the state-of-the-art — struggle with subtle consistency checks in Japanese pharmaceutical contexts.

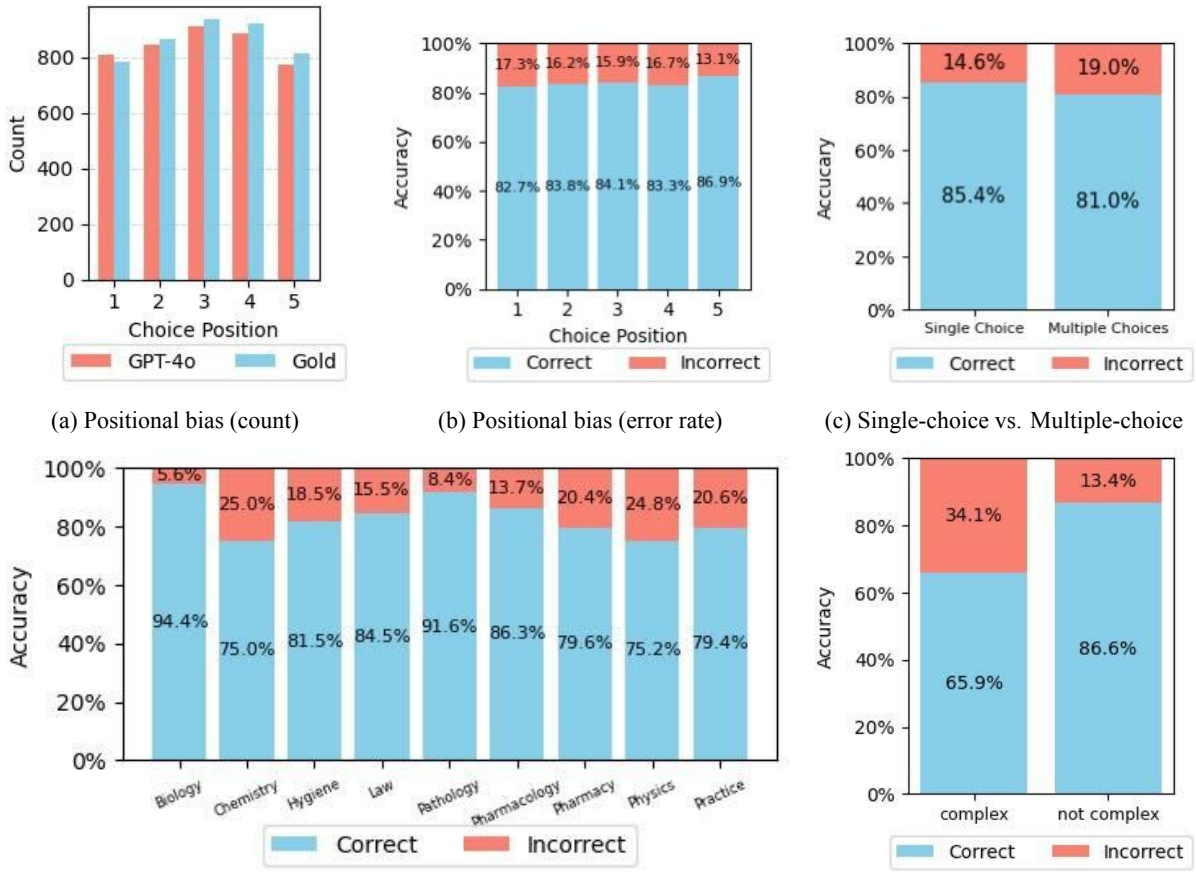
These findings highlight the diagnostic value of SogoCheck. Rather than being a standard QA task, it probes semantic understanding capabilities that go beyond surface-level knowledge. This suggests that inconsistency detection, especially in high-stakes domains like pharmacovigilance, requires capabilities not well-captured by general LLMs.

6.2 Deployable Domain-Specific Models: Challenges and Prospects

This study demonstrates the feasibility of building a high-performing, domain-specific LLM in Japanese without relying on commercial APIs. In pharmaceutical settings, where both data sensitivity and operational cost are critical concerns, locally trainable models such as ours present a practical and privacy-conscious alternative. Our open-source setup offers a replicable framework for enterprises and research groups seeking to train or fine-tune specialized models within secure environments. Moreover, our benchmark suite lays the groundwork for more practical evaluations of language models in healthcare and pharmaceutical contexts. In particular, tasks like SogoCheck capture practical detection abilities that are not assessed by conventional QA benchmarks, thereby suggesting promising directions for future model and dataset development.

Despite these advances, the deployment of domain-specific models faces a critical scalability-performance tradeoff. On one hand, 7B-parameter models such as JPharmatron are relatively feasible to deploy using a small cluster of GPUs. On the other hand, such models inevitably fall short of the performance levels achieved by larger models (e.g., 70B). Bridging this gap without compromising deployability remains an open challenge, and we believe our work represents a meaningful first step toward addressing this dilemma.

Our ultimate goal in this field is to achieve a strong and useful pharmaceutical LLM. To this



(a) Positional bias (count)

(b) Positional bias (error rate)

(c) Single-choice vs. Multiple-choice

(d) Category-wise accuracy

(e) Complex questions

Figure 5: **Error analysis on GPT-4o's responses in YakugakuQA.**

end, we need to further strengthen open models, as commercial models are often unavailable or restricted by regulations. Our experimental results, particularly those discussed in §5.3, suggest three directions for future work, listed in order of priority: (i) improving performance in core subjects to reach parity with commercial models, (ii) enhancing the overall capabilities of LLMs, and (iii) addressing weaknesses in lower-performing subjects. While the best open models already achieve acceptable performance, they still lag clearly behind commercial counterparts (Table 5). As a next step, it is essential to evaluate how much performance can be improved in targeted subject areas, depending on the intended application of the model, by simply incorporating a substantial amount of relevant training data. For the lower-performing subjects, including the improvement in chemistry and physics, both domain knowledge and reasoning ability must be significantly strengthened. However, considering development costs, we argue that addressing these weaknesses may not be a high priority in practice, as they can often be circumvented by limiting the task scope from application sides.

7 Conclusion

We presented **JPharmatron**, a Japanese domain-specific LLM for the pharmaceutical field, trained via continual pretraining on a bilingual pharmaceutical corpus. Alongside the model, we introduced **JPharmaBench**, the first benchmark suite covering diverse pharmaceutical language tasks. Our model outperforms existing open medical LLMs across diverse pharmaceutical tasks, highlighting that general medical specialization alone is insufficient for pharmaceutical applications. Notably, the benchmark includes tasks such as SogoCheck, which reflect real-world document validation workflows unique to the pharmaceutical domain. Beyond releasing a domain-specific model and benchmark, our work demonstrates the feasibility of building cost-effective, specialized LLMs deployable in secure, resource-constrained environments, which is critical for real-world use in privacy-sensitive domains like pharmaceuticals.

8 Limitations

Lack of Complete Instruction-Following Ability in LLMs

Some smaller models tend to deviate from instructions, often generating output that includes extraneous text beyond the expected format. A common error is the inclusion of additional phrases or explanations following a colon or line break. To ensure a fair comparison in our experiments, we post-processed model outputs by extracting only the selected choice and discarding any extra text.

Limitations of YakugakuQA

Firstly, questions with images should be addressed. In particular, the chemistry category lacks sufficient coverage due to the high proportion of image-based questions. While the rise of multimodal models, especially vision-language models, is an important development, this study focuses exclusively on text-only large language models. Therefore, image-based questions were excluded from our evaluation. In the future, this limitation should be revisited when assessing multimodal models.

Moreover, YakugakuQA is a simple five-choice question-answering task, which may not be sufficient for practical implementation, although it could serve as a minimum requirement.

Last but not least, the prompting strategy can also be improved. In our work, we used a simple setup as an initial step in this field. It should be noted that in-context learning of LLMs has the potential to boost performance, as demonstrated by Medprompt (Nori et al., 2023) in medical question-answering for example. This point remains controversial (Nori et al., 2024) and was not addressed in this study.

Limitations of NayoseQA

Although we introduce a novel benchmark NayoseQA, its current format is limited to multiple-choice QA. While this format enables controlled evaluation, it may not fully reflect the practical needs of real-world entity normalization systems, where open-ended or instruction-following formats are more appropriate. To address this, we have separately released an instruction-style (SQuAD (Rajpurkar et al., 2016)-type) variant of NayoseQA, which is not included in the main results but may serve as a valuable resource for future work on more realistic applications.

Limitations of SogoCheck

SogoCheck is currently limited in scale, with only a small number of consistency pairs included in the benchmark. This restricts the statistical robustness of evaluation and may limit its confidence across different model types and domains. In addition, generating realistic inconsistencies is inherently challenging. While we employed LLM-based generation methods to create contradictory statement pairs, it remains difficult to simulate subtle, human-like inconsistencies that naturally occur in real-world pharmaceutical texts. Many automatically generated inconsistencies tend to be either too trivial or too artificial, reducing their diagnostic value. Developing more authentic and diverse inconsistency examples remains an open challenge for future work.

References

Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, and Ari S Morcos. 2023. Semdedup: Data-efficient learning at web-scale through semantic deduplication. *arXiv preprint arXiv:2303.09540*.

Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. 2024. Large language models for mathematical reasoning: Progresses and challenges. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 225–237, St. Julian's, Malta. Association for Computational Linguistics.

Takuya Akiba, Makoto Shing, Yujin Tang, Qi Sun, and David Ha. 2025. Evolutionary optimization of model merging recipes. *Nature Machine Intelligence*, pages 1–10.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, et al. 2025. Smollm2: When smol goes big–data-centric training of a small language model. *arXiv preprint arXiv:2502.02737*.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy,

Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. PaLM 2 Technical Report. *Preprint*, arXiv:2305.10403.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Gillian Chaloner-Larsson, Roger Anderson, Anik Egan, Manoel Antonio Da Fonseca Costa Filho, Jorge F Gomez Herrera, World Health Organization. Vaccine Supply, and Quality Unit. 1999. A WHO guide to good manufacturing practice (GMP) requirements / written by Gillian Chaloner-Larsson, Roger Anderson, Anik Egan; in collaboration with Manoel Antonio da Fonseca Costa Filho, Jorge F. Gomez Herrera.

Juan Manuel Zambrano Chaves, Eric Wang, Tao Tu, Eeshit Dhaval Vaishnav, Byron Lee, S. Sara Mahdavi, Christopher Semturs, David Fleet, Vivek Natarajan, and Shekoofeh Azizi. 2024. Tx-LLM: A Large Language Model for Therapeutics. *Preprint*, arXiv:2406.06316.

Linqing Chen, Weilei Wang, Zilong Bai, Peng Xu, Yan Fang, Jie Fang, Wentao Wu, Lizhi Zhou, Ruiji Zhang, Yubin Xia, et al. 2024. PharmaGPT: Domain-Specific Large Language Models for Bio-Pharmaceutical and Chemistry. *arXiv preprint arXiv:2406.18045*.

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*.

Alexa Ehlert, Benjamin Ehlert, Binxin Cao, and Kathryn Morbitzer. 2024. Large Language Models and the North American Pharmacist Licensure Examination (NAPLEX) Practice Questions. *American Journal of Pharmaceutical Education*, 88(11):101294.

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.

Charles Goddard, Shamane Siriwardhana, Malikeh Ehghaghi, Luke Meyers, Vladimir Karpukhin, Brian Benedict, Mark McQuade, and Jacob Solawetz. 2024. Arcee's MergeKit: A toolkit for merging large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 477–485, Miami, Florida, US. Association for Computational Linguistics.

P. Hager, F. Jungmann, R. Holland, et al. 2024. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature Medicine*, 30(11):2613–2622.

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

Kexin Huang, Tianfan Fu, Wenhao Gao, Yue Zhao, Yusuf Roohani, Jure Leskovec, Connor W Coley, Cao Xiao, Jimeng Sun, and Marinka Zitnik. 2022. Artificial intelligence foundation for therapeutic science. *Nature Chemical Biology*.

Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. 2023. Financebench: A new benchmark for financial question answering. *arXiv preprint arXiv:2311.11944*.

Junfeng Jiang, Jiahao Huang, and Akiko Aizawa. 2024. JMedBench: A Benchmark for Evaluating Japanese Biomedical Large Language Models. *Preprint*, arXiv:2409.13317.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What Disease does this Patient Have? A Large-scale Open Domain Question Answering Dataset from Medical Exams. *arXiv preprint arXiv:2009.13081*.

Jungo Kasai, Yuhei Kasai, Keisuke Sakaguchi, Yutaro Yamada, and Dragomir Radev. 2023. Evaluating GPT-4 and ChatGPT on Japanese medical licensing examinations. *Preprint*, arXiv:2303.18027.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024a. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*.

Qintong Li, Leyang Cui, Xueliang Zhao, Lingpeng Kong, and Wei Bi. 2024b. GSM-plus: A comprehensive benchmark for evaluating the robustness of LLMs as mathematical problem solvers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2961–2984, Bangkok, Thailand. Association for Computational Linguistics.

Kelly Marchisio, Saurabh Dash, Hongyu Chen, Dennis Aumiller, Ahmet Üstün, Sara Hooker, and Sebastian Ruder. 2024. How does quantization affect multilingual LLMs? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15928–15947, Miami, Florida, USA. Association for Computational Linguistics.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. 2023. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*.

Harsha Nori, Naoto Usuyama, Nicholas King, Scott Mayer McKinney, Xavier Fernandes, Sheng Zhang, and Eric Horvitz. 2024. From Medprompt to o1: Exploration of Run-Time Strategies for Medical Challenge Problems and Beyond. *arXiv preprint arXiv:2411.03590*.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023a. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. 2023b. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari Morcos. 2022. Beyond neural scaling laws: beating power law scaling via data pruning. *Advances in Neural Information Processing Systems*, 35:19523–19536.

Issey Sukeda. 2024. Development and bilingual evaluation of Japanese medical large language model within reasonably low computational resources. *arXiv preprint arXiv:2409.11783*.

Issey Sukeda, Risa Kishikawa, and Satoshi Kodera. 2024a. 70B-parameter large language models in Japanese medical question-answering. *arXiv preprint arXiv:2406.14882*.

Issey Sukeda, Masahiro Suzuki, Hiroki Sakaji, and Satoshi Kodera. 2023. JMedLoRA: medical domain adaptation on Japanese large language models using instruction-tuning. *arXiv preprint arXiv:2310.10083*.

Issey Sukeda, Masahiro Suzuki, Hiroki Sakaji, and Satoshi Kodera. 2024b. Development and analysis of medical instruction-tuning for Japanese large language models. *Artificial Intelligence in Health*, 1(2):107–116.

Kushal Tirumala, Daniel Simig, Armen Aghajanyan, and Ari Morcos. 2023. D4: Improving llm pretraining via document de-duplication and diversification. *Advances in Neural Information Processing Systems*, 36:53983–53995.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Luong Trung, Xinbo Zhang, Zhanming Jie, Peng Sun, Xiaoran Jin, and Hang Li. 2024. ReFT: Reasoning with reinforced fine-tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7601–7614, Bangkok, Thailand. Association for Computational Linguistics.

Hirofumi Tsuruta, Hiroyuki Yamazaki, Ryota Maeda, Ryotaro Tamura, and Akihiro Imura. 2024. A SARS-cov-2 interaction dataset and VHH sequence corpus for antibody language models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. 2023. Ties-merging: Resolving interference when merging models. *Advances in Neural Information Processing Systems*, 36:7093–7115.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Ziqi Yin, Hao Wang, Kaito Horio, Daisuke Kawahara, and Satoshi Sekine. 2024. Should We Respect LLMs? A Cross-Lingual Study on the Influence of Prompt Politeness on LLM Performance. *arXiv preprint arXiv:2402.14531*.

Jingqing Zhang, Kai Sun, Akshay Jagadeesh, Parastoo Falakaflaki, Elena Kayayan, Guanyu Tao, Mahta Haghighat Ghahfarokhi, Deepa Gupta, Ashok Gupta, Vibhor Gupta, et al. 2024. The potential and pitfalls of using a large language model such as chatgpt, gpt-4, or llama as a clinical assistant. *Journal of the American Medical Informatics Association*, 31(9):1884–1891.

Yiran Zhao, Jinghan Zhang, I Chern, Siyang Gao, Pengfei Liu, Junxian He, et al. 2023. Felm: Benchmarking factuality evaluation of large language models. *Advances in Neural Information Processing Systems*, 36:44502–44523.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2023. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021.

A Ethical considerations

While JPharmatron is designed to complete pharmaceutical tasks resembling the real tasks in pharmacy companies, it is not yet confirmed to accomplish the real tasks within professional acceptable quality. It raises several ethical considerations that must be addressed to ensure responsible development and deployment.

Importantly, the model may still generate factually incorrect or misleading content. We recommend to further finetune our model with the company's real data and conduct additional use-case alignment and testing before deploying it in real-world practice. We further emphasize that the model is not intended for clinical use. Instead, it is suitable for document processing tasks, where potential risks can be mitigated through human review and validation of the generated content.

The training data may contain biases related to demographics, geographic representation, or commercial interests. Additionally, if any data were to originate from patents, proprietary databases, or unpublished sources, there would be a risk of inadvertently disclosing protected content or facilitating unauthorized reuse. Although all training data used in this study were sourced from publicly available datasets, we acknowledge that this issue was not directly addressed in the current work.

B Supplementary information on our benchmarks

B.1 YakugakuQA

The number of YakugakuQA is listed in Table 6. Among the available questions online, only those with texts were extraceted.

Prompt Below are the three-shot examples included in the prompt throughout our experiments. All of them are originally in Japanese, but translated into English by ChatGPT-4o mini for this article.

> Question: Which of the following insomnia medications inhibits the orexin receptor? Please select exactly one from the options 1, 2, 3, 4, or 5.
> 1: Brotizolam
> 2: Flunitrazepam
> 3: Eszopiclone
> 4: Ramelteon
> 5: Lemborexant
> Answer: 5
> Question: Which two mechanisms of action describe the effects of sacubitril/valsartan? Please select exactly two from the options 1, 2, 3, 4, or 5.
> 1: Inhibits neprilysin, thereby preventing the breakdown of endogenous natriuretic peptides, resulting in vasodilation and diuretic effects.
> 2: Inhibits angiotensin II receptors, suppressing aldosterone secretion from the adrenal cortex, thereby causing vasodilation.
> 3: Acts on ANP receptors in the blood vessels and kidneys, activating guanylate cyclase, resulting in vasodilation and diuretic effects.
> 4: Blocks aldosterone receptors in the collecting ducts, leading to diuretic effects.
> 5: Inhibits angiotensin-converting enzyme, thereby preventing the formation of angiotensin II, resulting in vasodilation.
> Answer: 1,2
> Question: Which of the following migraine prophylactic drugs inhibits calcitonin gene-related peptide (CGRP)? Please select exactly one from the options 1, 2, 3, 4, or 5.
> 1: Basiliximab
> 2: Trastuzumab
> 3: Benralizumab
> 4: Galcanezumab
> 5: Tocilizumab
> Answer: 4

	Biology	Chemistry	Hygiene	Law	Pathology	Pharmacology	Pharmacy	Physics	Practice	Total
2012	17	4	30	29	37	38	36	17	65	273
2013	16	3	32	28	36	34	33	11	63	256
2014	15	4	28	29	35	37	28	13	63	252
2015	8	3	26	27	35	35	31	9	60	234
2016	10	3	30	27	37	40	29	12	50	238
2017	11	2	28	26	37	36	27	10	54	231
2018	11	4	31	27	36	35	25	10	53	232
2019	9	1	28	28	32	33	26	12	46	215
2020	12	4	25	26	33	33	17	12	42	204
2021	6	2	30	27	35	30	19	10	55	214
2022	9	3	25	27	33	33	24	15	48	217
2023	10	3	23	25	27	33	22	15	47	205
2024	11	11	33	23	28	36	31	18	59	250

Table 6: **The number of questions used in our experiments by year and category.** The questions that include images have been excluded from the original NPLE.

Category	The number of questions
clinical_knowledge	150
college_biology	143
college_chemistry	99
college_medicine	150
college_physics	100
high_school_biology	148
high_school_chemistry	149
high_school_physics	150
high_school_statistics	150
medical_genetics	99
nutrition	149
professional_medicine	150
virology	150
Total	1787

Table 7: **The number of questions by categories included in pharmaceutical-related JMMLU.**

B.2 Pharmaceutical-related subset of JMMLU

The number of questions included in each category of JMMLU which was used in our evaluation experiments is listed in Table 7. The category-wise accuracy is shown in Table 8. Consistent with the results in YakugakuQA (Table 5), the overall trend that biology tends to score higher than chemistry and physics is observed.

C Model & Training

C.1 Data accumulation

The continual pretraining corpus used for JPharmatron is composed of five categories of text, collected from publicly available sources. Each data type was selected to contribute domain-relevant knowledge or general linguistic fluency. An overview is provided below:

Journal Articles Academic papers and review articles related to pharmacology, pharmacy practice, and clinical medicine. These texts provide rich domain-specific vocabulary and formal written structures.

PubMed Abstract Subset A curated selection of English abstracts from the PubMed database, focusing on drug-related publications. This source contributes approximately 8 billion tokens and provides a biomedical foundation to complement the Japanese data.

Package Inserts approved by PMDA Texts published by Japan's Pharmaceuticals and Medical Devices Agency (PMDA), such as drug approval summaries, review reports, and safety alerts. These documents contribute approximately 87 million tokens and reflect regulatory terminology.

Official Documents from Governmental Institutes Documents from government-affiliated organizations including the Pharmaceuticals and Medical Devices Act.

General-Domain Corpus A part of FineWeb[5] and Swallow Dataset[6].

C.2 Data Filtering

We constructed a high-quality, domain-specific corpus for the pharmaceutical domain by leveraging a multi-stage filtering pipeline built upon large

[5] https://huggingface.co/datasets/HuggingFaceFW/fineweb

[6] https://huggingface.co/datasets/tokyotech-llm/swallow-magpie-ultra-v0.1

Model	clinical_ knowledge	college_ biology	college_ chemistry	college_ medicine	college_ physics	high_school_ biology	high_school_ chemistry	high_school_ physics	high_school_ statistics	medical_ genetics	nutrition	professional_ medicine	virology	Over -all
TinySwallow-1.5B-Instruct	41.3	28.0	29.3	36.0	28.0	40.5	26.8	25.3	28.7	31.3	34.2	30.7	34.0	32.1
sarashina2.2-3b-instruct	39.3	45.5	29.3	42.0	35.0	52.7	26.2	27.3	34.0	40.4	47.7	44.7	24.7	37.8
Qwen2.5-7B-Instruct	52.7	46.9	30.3	41.3	37.0	50.7	36.2	28.7	32.7	48.5	57.7	49.3	41.3	42.9
Meditron3-Qwen2.5-7B	48.7	27.3	19.2	26.7	33.0	37.8	23.5	28.7	34.7	28.3	44.3	33.3	22.0	31.7
Llama-3-Swallow-8B-Instruct-v0.1	30.7	12.6	17.2	25.3	11.0	26.4	20.1	21.3	27.3	11.1	16.1	30.0	11.3	20.6
Llama-3.1-Swallow-8B-Instruct-v0.3	52.0	45.5	35.4	47.3	37.0	55.4	35.6	30.0	36.7	55.6	53.7	44.7	42.0	44.0
GPT-4o	82.7	93.0	60.6	81.3	69.0	85.1	76.5	70.0	82.0	88.9	82.6	94.7	56.7	79.1
Ours (best)	58.7	64.3	44.4	48.7	50.0	65.5	48.3	46.0	64.7	59.6	62.4	58.7	40.7	55.0

Table 8: **Accuracy comparison on JMMLU across different subject categories and different LLMs.**

language models (LLMs) and trained classifiers. Following SmolLM2 (Allal et al., 2025), the overall procedure consists of three steps:

1. We first sampled a subset of documents from the Common Crawl dataset (CC100). A high-performing LLM (Qwen2.5-72B) was prompted to assign each page a pharmaceutical relevance score ranging from 0 (irrelevant) to 5 (highly relevant).

2. Using 54,056 LLM-labeled samples, we trained a classifier to predict the pharmaceutical relevance score of input documents. Pages scoring 1 or higher were retained.

3. The retained documents were further evaluated using the same LLM to assign an educational quality score (0-5). A second classifier, trained on 5,478 LLM-labeled samples, was used to filter out documents with an educational quality score 3 or lower. This ensured that the resulting data not only pertains to pharmaceutical content but is also of pedagogical value.

All training data for both classifiers were generated using high-confidence outputs from the Qwen2.5-72B model. Both classifiers were trained following the configuration of the finemath-classifier[7] framework.

As a result of this filtering pipeline, we collected 904,651 high-quality, pharmaceutical-related documents (totalling 1.2 billion tokens) from the deduplicated Common Crawl (llm-jp-corpus-v3[8]).

C.3 Data cleansing

In this study, we employed the D4 algorithm (Tirumala et al., 2023) to perform data deduplication, aiming to reduce redundant information. D4 is primarily composed of SemDeDup (Semantic deduplication) (Abbas et al., 2023) and SSL Prototype (Self-Supervised Learning Prototypes) (Sorscher et al., 2022). The former incorporates k-means clustering to eliminate texts with cosine similarity larger than $1 - \epsilon$. We set $\epsilon = 3 \times 10^{-8}$ for the discarding threshold in SemDeDup and $R = 0.95$ for the discarding proportion in SSL Prototype, respectively. In summarization, the total number of tokens were reduced from 10B to 9B.

C.4 Base model selection

Discussing industrial applications often lead to the cost perspectives. Different from research purpose development, the operational cost in inference phase also should be taken into account, otherwise no institution can afford to utilize the trained model. Training a model from scratch to learn Japanese was deemed prohibitively costly. Therefore, in selecting the base model, we prioritized the use of a pretrained model that had already been trained on Japanese data, and we also sought a model with a commercially viable license that would facilitate its adoption within the pharmaceutical industry. We restricted the model size to around 7B for better usability considering the training cost and inference cost. Based on these criteria, we chose Qwen2.5-7B (Yang et al., 2024) as the base model.

C.5 Enhancing Instruction Following via Model Merging

Our domain-specific model trained through continued pretraining exhibited poor instruction-following capabilities. As a result, these models struggle to answer multiple-choice questions correctly, rendering them ineffective for standard benchmark evaluations which rely heavily on such tasks.

Instead of applying supervised fine-tuning (SFT), which can be resource-intensive and require carefully aligned datasets, we adopt a lightweight approach by leveraging model merging. Specifically, we aim to endow a domain-adapted model with strong instruction-following and reasoning ca-

[7] https://huggingface.co/HuggingFaceTB/finemath-classifier

[8] https://gitlab.llm-jp.nii.ac.jp/datasets/llm-jp-corpus-v3

Merge method	YakugakuQA (%)
TIES (weight 8:2)	57.2
TIES (weight 7:3)	59.0
TIES (weight 6:4)	60.4
DARE TIES by EvoLLM	**60.7**

Table 9: **Accuracy comparison on YakugakuQA across different merging methods.** Qwen2.5-7B-Instruct was used as the base model and JPharmatron-7B (Ours) was used as the auxiliary model.

pabilities by merging it with a general-purpose instruction-tuned model.

To this end, we designate Qwen2.5-7B-Instruct as the base model, given its demonstrated strength in instruction adherence and task generalization. The domain-specific model, pretrained on 2B tokens of pharmaceutical texts, serves as the knowledge-rich counterpart in the merge.

We employ the TIES merging strategy (Yadav et al., 2023) provided by *mergekit* (Goddard et al., 2024), and assign a weight to balance the retention of domain knowledge while preserving the core reasoning and output structure of the instruction-tuned base model. Table 9 shows the superiority of EvoLLM (Akiba et al., 2025) coupled with DARE TIES merging.