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

Shinnosuke Ono[1,2,*] **Issey Sukeda**[1,2,*] **Takuro Fujii**[**] **Kosei Buma**[1,3] **Shunsuke Sasaki**[1,2]

[1]EQUES Inc. [2]The University of Tokyo [3]University of Tsukuba

Correspondence: ono@ms.k.u-tokyo.ac.jp

Abstract

We present JPHARMATRON, a Japanese domain-specific large language model (LLM) for the pharmaceutical field, developed through continual pre-training on two billion Japanese pharmaceutical tokens and eight billion English biomedical tokens. For rigorous evaluation, we introduce JPHARMABENCH, a benchmark suite consisting of three new benchmarks: YakugakuQA, based on national pharmacist licensing exams; NayoseQA, which tests cross-lingual synonym and terminology normalization; and SogoCheck, a novel task involving cross-document consistency checking. We evaluate our model against open-source medical LLMs and commercial models, including GPT-4o. Experimental results show that JPHARMATRON outperforms existing open models and achieves competitive performance with commercial ones. Interestingly, even GPT-4o performs poorly on SogoCheck, suggesting that cross-sentence consistency reasoning remains an open challenge. JPHARMATRON enables secure and local model deployment for pharmaceutical tasks, where privacy and legal constraints limit the use of closed models. Besides, JPHARMABENCH offers a reproducible framework for evaluating Japanese pharmaceutical natural language processing. Together, they demonstrate the feasibility of practical and cost-efficient language models for Japanese healthcare and pharmaceutical sectors. Our model, codes, and datasets are available on HuggingFace[1].

1 Introduction

Large language models (LLMs) have achieved remarkable performance across a wide range of general-purpose 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).

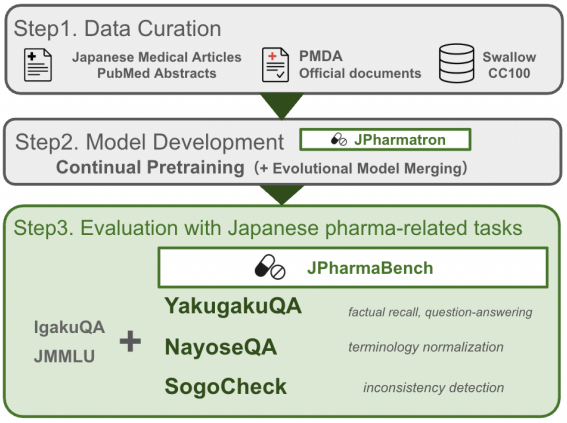


Figure 1: **JPHARMATRON and JPHARMABENCH.** The pipeline for data curation, continual pre-training, and evaluation of JPHARMATRON.

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 Good Manufacturing Practices (Chaloner-Larsson et al., 1999) and the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use guidelines[2]. Despite these challenges, little work has been done to develop LLMs tailored for pharmaceutical opera-

[*]Equal contributions. [**]Independent researcher.

[1]https://huggingface.co/collections/EQUES/jpharmatron and /jpharmabench.

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

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 1316–1332

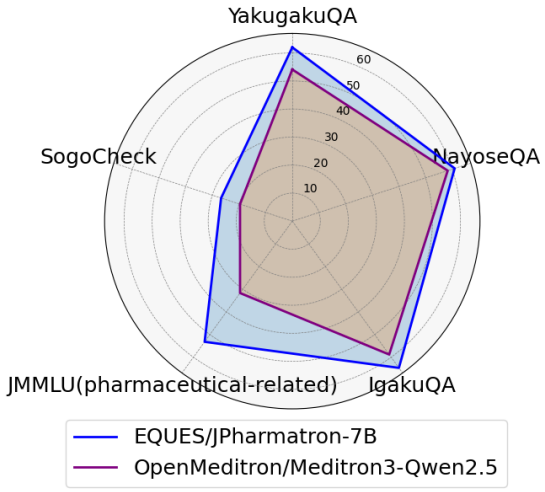


Figure 2: **Performance comparison with Meditron.** JPHARMATRON consistently achieves higher scores than Meditron (Chen et al., 2023) across JPHARMABENCH, IgakuQA (Kasai et al., 2023), and JMMLU (Yin et al., 2024).

tions, especially in Japanese.

In this work, we provide a complete methodology from data collection to evaluation, as depicted in Figure 1. First, we present JPHARMATRON, a Japanese LLM specialized for pharmaceutical operations. To build JPHARMATRON, we perform continual pre-training (CPT) of the Qwen2.5 (Yang et al., 2024) model using a curated 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 capabilities, we introduce three novel benchmarks: (1) YakugakuQA (§3.1): a multiple-choice question-answering (QA) benchmark based on the Japanese National License Examination for Pharmacists; (2) NayoseQA (§3.2): a paraphrasing benchmark for standardizing drug names and active substances; (3) SogoCheck (§3.3): a cross-document consistency check task that reflects 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 evaluated JPHARMATRON using in-context learning with JPHARMABENCH and two existing benchmarks. Without task-specific fine-tuning, JPHARMATRON outperformed similar-sized LLMs including Meditron (Chen et al., 2023), a medical domain-specific model, by 7.9% on YakugakuQA and by 23.3% on the pharmaceutical subset of JMMLU (Yin et al., 2024), as highlighted in Figure 2. These results demonstrate the limited transfer ability of medical domain-specific LLMs, as well as underscoring the need for pharmaceutical-specific models.

In summary, our contributions are threefold:

- We introduce the first LLM and evaluation benchmarks specifically designed for Japanese pharmaceutical NLP.
- Our benchmark design is inspired by concrete industrial problems, ensuring alignment with real-world regulatory and operational workflows.
- Together, the model and benchmarks provide a practical and reproducible foundation for developing domain-specific LLMs in regulated industries such as pharmacy.

2 Related Work

Models and benchmarks in healthcare. There has been growing interest in the development of domain-specific LLMs and evaluation benchmarks in the medical and clinical fields. This trend spans multiple languages with significant efforts in English (Singhal et al., 2023a,b; Chen et al., 2023), Chinese (Li et al., 2023; Chen et al., 2025), and also Japanese (Sukeda et al., 2023, 2024a,b) to align LLM capabilities with healthcare expertise.

Despite advances in medicine, the pharmaceutical domain remains relatively underexplored in LLM research. Only a handful of models such as PharmaGPT (Chen et al., 2024) and Tx-LLM (Chaves et al., 2024) have been developed with a pharmaceutical focus, and none of them are publicly available. Evaluation benchmarks for pharmaceutical LLMs are also notably limited, especially in non-English contexts. Existing evaluations have largely relied on the North American Pharmacist Licensure Examination (Ehlert et al., 2024; Chen et al., 2024), with no publicly available pharmaceutical QA datasets in Japanese. Existing benchmarks including medical topics (Hendrycks et al., 2021; Yin et al., 2024; Kasai et al., 2023) do not feature pharmaceutics as a distinct category.

Even with the rapid progress of LLMs by 2025, relatively basic multiple-choice exam–style benchmarks remain important for evaluating core domain knowledge before addressing more complex tasks. In pharmaceutics, where terminology, regulations, and practical contexts differ from those in medicine, such controlled tasks serve as reliable diagnostics of foundational understanding. Still, these benchmarks alone are insufficient to assess a model's practical applicability. To bridge this gap between academic evaluation and real-world utility, we constructed three complementary datasets: one emphasizing academic coverage and two reflecting practical, real-world use cases.

Mid-training of LLMs. Recent studies have explored mid-training as an efficient approach to further develop the capability of a pre-trained LLM (Abdin et al., 2024). Specifically, continual pre-training (CPT) is a promising method of domain adaptation (Fujii et al., 2024). Instead of training from scratch, which is computationally prohibitive, CPT continues the pre-training process using high-quality, domain-specific corpora. This approach allows the model to maintain broad linguistic and reasoning capabilities while incrementally acquiring specialized knowledge. CPT particularly suits our setting because the Japanese pharmaceutical domain requires precise terminology understanding and document-style fluency. This domain is poorly represented in general web data, yet lacks the massive domain corpora needed for full-scale training. CPT therefore enables effective specialization under limited compute and data availability.

3 JPharmaBench

To evaluate language models in the Japanese pharmaceutical domain, we constructed three novel benchmarks. Each of them reflects 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 processed by LLMs and human experts. In addition, they are structured as QA tasks, allowing automated evaluation.

3.1 YakugakuQA: National Licensing Exam

YakugakuQA is a QA dataset based on the Japanese National License Examination for Pharmacists administered by the Ministry of Health, Labour and Welfare. While most questions follow a five-choice single-answer format, some questions allow multiple selections or include more than five options (Figure 3).

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 License Examination for Pharmacists.** The model is required to output "4" in this case.

To construct YakugakuQA, we collected the exam data (i.e., questions, answers and commentaries) from 2012 to 2024, available at *yakugaku lab*[3]. These samples were then cleansed using HTML parsing, space removal, and word normalization.

Further, to ensure the authenticity of the benchmark and facilitate the future use, we manually added metadata indicating whether a question is valid, based on the announcement issued after the examination from the Ministry of Health, Labour and Welfare. An invalid question may include contradiction in itself or allow multiple answers even for a single-choice question. Another metadata field manually added contains the category of a problem, spanning across nine related areas: pharmacy, pharmacology, chemistry, pathology, hygiene, physics, practice, law, and biology (Table 4 in Appendix).

3.2 NayoseQA: Synonym and Terminology Normalization

NayoseQA evaluates 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, brand name ↔ generic name, and chemical name ↔ common name.

This type of task, conventionally referred to as "nayose" in Japanese, is routinely performed in the pharmaceutical industry. It involves linguistic and domain-specific reasoning to recognize synonymous terms for pharmaceutical compounds. In

[3] https://yakugakulab.info/

Benchmark	Format	Main Skill	Source	#Examples	Language(s)
YakugakuQA	5-or-more-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.

real-world pharmaceutical practices 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.

NayoseQA is also a multiple-choice benchmark. To construct NayoseQA, we first collected headline entries and the corresponding different names from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database[4]. Next, we used a Japanese LLM, DeepSeek-R1-Distill-Qwen-32B-Japanese (Ishigami, 2025), to generate incorrect choices. These choices were further verified by Qwen2.5-72B and human participants to ensure authenticity.

3.3 SogoCheck: Inconsistency Detection in Paired Documents

SogoCheck is a multiple-choice, multiple-answer benchmark assessing how accurately LLMs can detect logical or factual inconsistencies[5] between two pieces of pharmaceutical text. Unlike factual benchmarks asking whether a synthetic document contains 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. This benchmark is particularly valuable because detecting inconsistencies is crucial in practical workflows such as regulatory review, where conflicting information can lead to severe medical or legal consequences.

In this task, we provide a model with a pair of Japanese documents. The model is then asked to check their explicit and implicit consistency. Some inconsistencies are trivial (e.g., numerical mistakes, see Figure 4), while others require pharmacological reasoning or recognition of subtle semantic contradictions (Figure 8 in Appendix). Note that a pair may even contain multiple types of inconsistencies. The model also needs to classify each inconsistency into a predefined set of labels (Appendix B.3). The evaluation metric is the accuracy of the detected line and predicted label.

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, Change in humidity

Figure 4: **One of the simplest examples from SogoCheck.** The numbers are inconsistent across two inputs. Note that most of the questions involve longer text pairs (1858 Japanese letters on average) which contain multiple discrepancies with different labels.

To construct SogoCheck, we used the Japanese Pharmacopoeia[6] as the primary source. It provides detailed information on individual pharmaceuticals circulating in Japan, such as a general description, chemical information (e.g., the chemical formula), an identification test to confirm the identity of the substance, and a purity test to detect impurities. We then randomly sampled 200 medicines and extracted the identification test section, which we used as one of a pair for each sample. Then, we used the same LLM as we used for NayoseQA to generate inconsistencies based on the candidate labels, before undergoing the same validation process as NayoseQA. The labels and the specific prompt we used are shown in Appendix B.3.

3.4 Statements on Data Contamination

The original sources used for creating our proposed benchmarks, particularly YakugakuQA and NayoseQA, may be contained in the pre-training data of some of the models we evaluate, such as GPT-4o and Qwen2.5, which is an acknowledged

[4] https://www.genome.jp/kegg/drug/
[5] Referred to as "sogo" in the industry.
[6] https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/0000066530.html

challenge in the literature. However, our experiments did not show any explicit signs of contamination, such as 100% accuracy. This indicates that the utility of our benchmarks is not impaired. Since the current literature lacks even simple, multiple-choice benchmarks for the pharmaceutical domain, we believe that our benchmark will be an important contribution to the community.

Moreover, our SogoCheck benchmark consists of synthetic examples that are novel and highly unlikely to be present in any model's training data. The low overall performance on this task shown in Section 5.2, even by GPT-4o, suggests that it poses an unsolved challenge that cannot be overcome simply by memorization.

4 JPHARMATRON

We developed JPHARMATRON through CPT based on Qwen2.5-7B (Yang et al., 2024), a multilingual open-source LLM that also supports Japanese, and evolutionary merging. This base model was chosen for its strong general performance, multilingual capacity, and availability under a commercially permissible license. See Appendix C.4 for further details.

We curated the training corpus from publicly available sources chosen to enhance the model's capability to solve practical tasks using professional knowledge. See Appendix C for details on the data curation process. We prepared three variations of the training corpus, resulting in these three models:

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 (Appendix C.1).

9B tokens (deduped): Based on the 10B-token corpus, further augmented with 1.2B tokens from the CC100 multilingual dataset (Conneau et al., 2020; Wenzek et al., 2020). After removing duplicates, the number of tokens was finally 9B tokens. This corpus is expected to be of higher quality than the 10B variant.

We emphasize that our goal was not to outperform proprietary LLMs such as GPT-4o, but to develop a practically deployable model as the 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).

5 Evaluation

5.1 Experimental Setups

We evaluated our domain-specific model against (1) general-purpose Japanese LLMs (viz., Swallow series or equivalent), (2) a medical LLM (viz., Meditron)[7], and (3) GPT-4o via the OpenAI API. The evaluation was conducted across JPHARMABENCH, toghether with two existing Japanese medical benchmarks: IgakuQA (Kasai et al., 2023) and the pharmaceutical subset of JMMLU. This setup enables direct comparison with prior work.

Each model was prompted once with three-shot examples (Appendix B) using identical hyperparameters (temperature = 0.8, top-p = 0.95, max tokens = 4096). Accuracy was computed by exact match, with models selecting one or more answers as appropriate.

Note that some questions in YakugakuQA require visual input, e.g., identifying a chemical reaction depicted in the image. Such image-based questions are outside the scope of our research and were therefore excluded from the experiments. Finally, see Table 4 for the number of questions by year and category used in our experiments.

5.2 Quantitative Results

Table 2 shows the models' accuracy on each benchmark. While GPT-4o achieved the highest overall accuracy, as expected from a frontier commercial LLM, our domain-specific model consistently outperformed both Meditron and the general-purpose Japanese models across all tasks. This highlights the effectiveness of domain-specific CPT in Japanese, and establishes our model as the strongest open baseline for Japanese pharmaceutical NLP tasks.

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 CPT, even without training from scratch. In addition, specialization in the medical domain alone

[7]We used 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 fair evaluation.

may be insufficient for handling pharmaceutical tasks effectively. The accuracy results by categories are listed in Table 3, together with additional larger models for reference: 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 (Llama-3.1-Swallow) was surprisingly small. This suggests that the task primarily requires lexical and semantic matching capabilities rather than deep domain-specific pharmaceutical knowledge. Although domain adaptation improved performance modestly, it appears that general-purpose LLMs with strong multilingual and synonym handling capabilities can already perform well on such terminology normalization tasks. We thus argue 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

To identify common failure patterns and inform future improvements in domain-specific LLMs, we analyze the 16.4% of questions in YakugakuQA that were answered incorrectly by GPT-4o.

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, it chose option "1" more times than there were questions with the correct option "1" (Figure 5a). This, combined with the highest error rate of the option "1" (Figure 5b), indicates a positional bias toward the first option.

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 stem from the prevalence of general knowledge, single-answer questions in these domains. This pattern is commonly observed across various LLMs, as shown in Table 3, and also in JMMLU as shown in Table 6.

Complex questions. Based on the previous observation, we employed Qwen2.5-72B-Instruct (Yang et al., 2024) to annotate questions that require complex reasoning or calculations, following the LLM-as-a-Judge framework (Li et al., 2024a). Although such questions accounted for fewer than 500 of the 3,021 total questions, 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 capabilities required for pharmaceutical NLP. While existing benchmarks such as IgakuQA and JMMLU primarily focus on factual recall, our benchmarks better reflect the demands of real-world pharmaceutical decision-making, as discussed in §3.

The evaluation results confirm that this broader scope offers meaningful insights. On YakugakuQA and NayoseQA, JPHARMATRON showed consistent improvements over the baselines, suggesting that domain-specific CPT effectively enhances factual recall and term-level understanding. In con-

	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 (deduped)	**62.0**	60.9	**26.7**	**64.7**	53.2

Table 2: **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 (deduped)	**80.5**	45.7	57.9	63.8	**73.8**	58.4	54.9	51.6	61.3	62.0

Table 3: **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.

trast, SogoCheck presented a more difficult challenge: the 2B-tokens variant even failed to improve. Moreover, the surprisingly low accuracy of GPT-4o indicates that current LLMs, even the state-of-the-art ones, struggle with subtle consistency checks in Japanese pharmaceutical contexts.

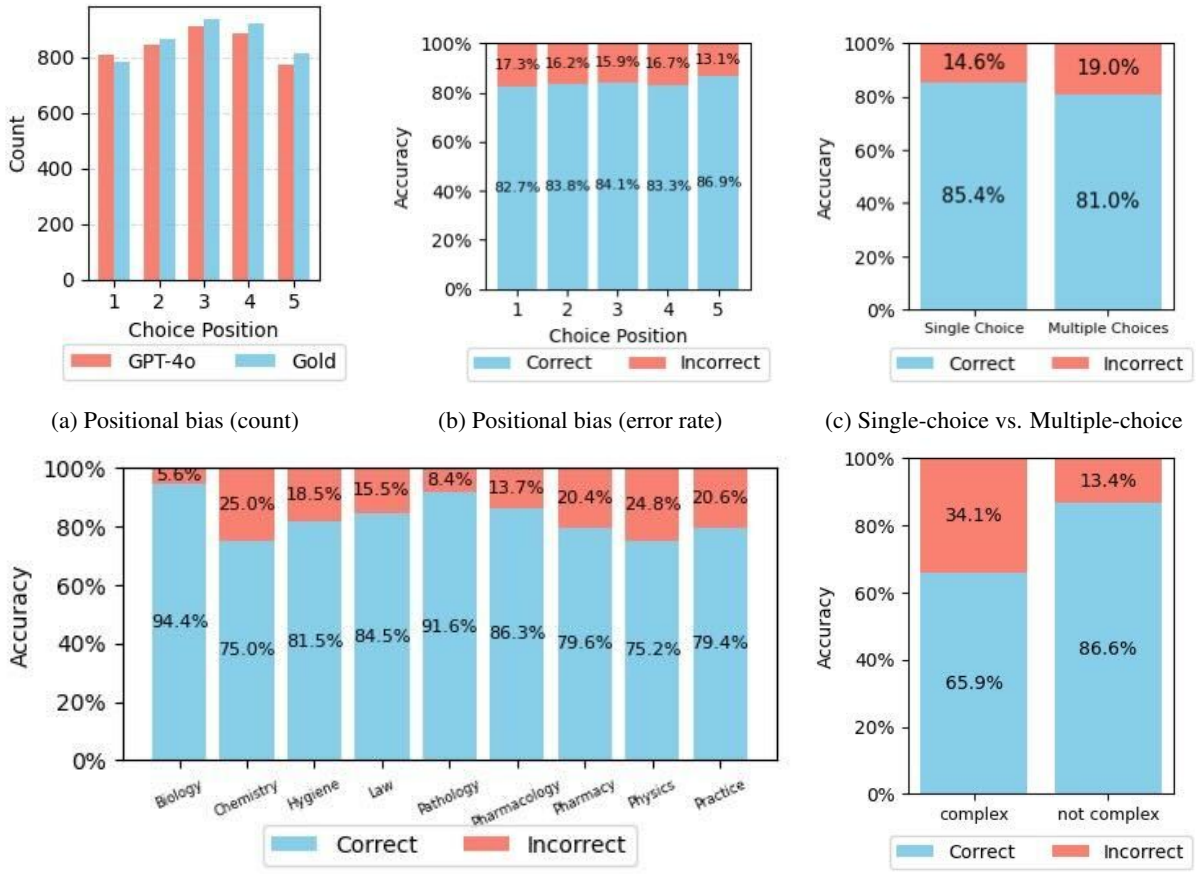
These findings highlight the diagnostic value of SogoCheck. Rather than being a standard QA task, it probes semantic understanding capabilities that go beyond surface-level knowledge. This suggests that inconsistency detection, especially in high-stakes domains such as pharmacovigilance, requires capabilities not captured well by general-purpose 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 choice. Our open-source setup offers a replicable framework for enterprises and researchers seeking to prepare specialized models within secure environments. Moreover, our benchmark suite lays the groundwork for more practical evaluations of LLMs 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 the one hand, 7B-parameter models such as JPHARMATRON are relatively feasible to deploy using a small cluster of GPUs. However, 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



(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.**

strong and useful pharmaceutical LLM. To this 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 their commercial counterparts (Table 3). 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, simply by incorporating a substantial amount of relevant training data. For the lower-performing subjects, including 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 narrowing the application scope.

7 Conclusion

We presented **JPHARMATRON**, a Japanese domain-specific LLM for the pharmaceutical field, trained via CPT on a bilingual pharmaceutical corpus. Alongside the model, we introduced **JPHARMABENCH**, the first benchmark suite covering a variety of pharmaceutical NLP tasks. Our model outperforms existing open medical LLMs across diverse pharmaceutical tasks, highlighting that general medical specialization alone is not sufficient for pharmaceutical applications. In particular, our 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 the 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 the model outputs by extracting only the selected choice and discarding any extra text.

Limitations of YakugakuQA

First, questions with images need to 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 LLMs. 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 multiple-choice QA task, which may not be sufficient for practical implementation, although it could serve as a minimum requirement.

Finally, the prompting strategy can also be improved. In our work, we used a simple setup as an initial step in this field. Note that in-context learning of LLMs has the potential to boost performance, as demonstrated by Medprompt (Nori et al., 2023) in medical QA 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 the 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. Developing more authentic and diverse inconsistency examples remains an open challenge for future work. Overall, however, we believe that our proposed benchmark serves as a valuable first step toward evaluating practical reasoning skills not covered by the existing benchmarks.

References

Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, and Ari S Morcos. 2023. Semdedup: Data-efficient learning at web-scale through semantic deduplication. In *Proceedings of the ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models (ME-FoMo)*.

Marah I Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio CT Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. 2024. Phi-4 technical report. Technical Report MSR-TR-2024-57, Microsoft.

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, Agustín Piqueres Lajarín, Hynek Kydlíček, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Ben Burtenshaw, Clémentine Fourrier, Haojun Zhao, Hugo Larcher,

Mathieu Morlon, Cyril Zakka, Colin Raffel, Leandro von Werra, and Thomas Wolf. 2025. SmolLM2: When smol goes big—data-centric training of a fully open small language model. In *Proceedings of the 2nd Conference on Language Modeling (COLM 2025)*, Montréal, Canada. To appear.

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.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, and Benyou Wang. 2025. Towards medical complex reasoning with LLMs through medical verifiable problems. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14552–14573, Vienna, Austria. Association for Computational Linguistics.

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*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

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)*.

Ryosuke Ishigami. 2025. Deepseek-r1-distill-qwen-32b-japanese.

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*.

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*.

Jianquan Li, Xidong Wang, Xiangbo Wu, Zhiyi Zhang, Xiaolong Xu, Jie Fu, Prayag Tiwari, Xiang Wan, and Benyou Wang. 2023. Huatuo-26m, a large-scale chinese medical qa dataset. *Preprint*, arXiv:2305.01526.

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*.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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, 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.

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*.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

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, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

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 real tasks in pharmaceutical companies, it is not yet confirmed to accomplish those real tasks with a professionally acceptable level of 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 further fine-tuning of 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 JPHARMABENCH and JMMLU

This section describes additional details on our evaluation method.

We conducted our experiments in a three-shot manner. The evaluation prompt for each benchmark was constructed by concatenating (1) general benchmark description, (2) three-shot examples and (3) a question and its choices.

B.1 YakugakuQA

As discussed in §3.1, we tallied the number of questions in YakugakuQA per category. See Table 4 for details.

General benchmark description. Here is the benchmark description text used for the YakugakuQA evaluation, translated into English:

> Solve the Japanese National License Examination for Pharmacists. Choose all of the correct answers.

Three-shot examples. 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

Year	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 4: **The number of questions used in our experiments by year and category.** Questions with visual inputs have been excluded from this table and our experiments.

> 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

Sample question from YakugakuQA. Figure 6 is the original Japanese format corresponding to Figure 3 in the main text.

次のうち、希薄溶液の理想的性質ではないものはどれですか？１つ選んでください。
1. 蒸気圧降下
2. 凝固点降下
3. 沸点上昇
4. 表面張力の低下
5. 浸透圧

Figure 6: **A sample question from YakugakuQA**, in its original Japanese text.

B.2 NayoseQA

General benchmark description. Here is the general benchmark description used for NayoseQA, translated in English:

> Please select one of the following options. Do not output your reasoning, only provide the answer.

Sample questions from NayoseQA. Figure 7 shows three sample questions from NayoseQA. The first question is about a chemical compound, while the remaining two involve translation tasks between English and Japanese.

常水の化学式は？
1. H2O
2. NH3
3. C6FeN6
4. Homogentisate

常水の英語表記は？
1. Fibronectin
2. Water
3. Taurine
4. Homogentisate

Water の日本語表記は？
1. 二酸化炭素
2. プトレシン
3. L-オルチニン
4. 常水

Figure 7: **Three sample questions from NayoseQA.** The first question asks the chemical formula for water. The second and third questions are about Japanese/English translation for water.

B.3 SogoCheck

General benchmark description. Below is the general benchmark description used for SogoCheck, translated into English.

> The provided [Test Method Document] is highly likely to contain discrepancies such as those listed in the [List of Potential Discrepancy Aspects]. Please extract and point out any discrepancies in the [Test Method Document] by referring to the [Reference Document], and present them in bullet-point format.
>
> [List of Potential Discrepancy Aspects]:
>
> (Omitted. See below.)

Inconsistency labels. These are the inconsistency labels we employ to generate synthetic documents and to evaluate models on SogoCheck:

- Typographical errors or omissions
- Changes in numerical values
- Changes in materials
- Changes in ingredient names
- Changes in time or temperature
- Changes in evaluation criteria
- Changes in expiration dates
- Changes in process control
- Changes in raw material control
- Changes in operating procedures
- Changes in specifications and test methods
- Changes in factory names
- Changes in parameters
- Deletion of substances
- Changes in the calculation of the starting date of the expiration period
- Changes in sampling timing
- Deletion or modification of test items, test methods, or test outsourcing parties

As for the distribution of labels, 54.21% out of 2,841 labels are marked as "Numerical changes". Other frequently appearing labels are "Expiration date changes" (3.41%), "Temperature changes" (3.34%), "Judgment criteria changes" (2.04%), and "Component name changes" (1.87%). We observe the clear tendency that changes in terms of numbers are predominant, which we suppose is reasonable in light of real-world discrepancies.

Sample task from SogoCheck. In addition to the example shown in Figure 4, Figure 8 presents a longer sample of the original Japanese text along with its English translation.

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

B.4 Pharmaceutical-related subset of JMMLU

The number of questions included in each category of JMMLU in our evaluation experiments is listed in Table 5. The category-wise accuracy is shown in Table 6. Consistent with the results for YakugakuQA (Table 3), we can observe the overall trend that biology tends to score higher than chemistry and physics.

C Model & Training

C.1 Training Data Curation

The CPT 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

Text A (ja): 型肝炎ワクチン Adsorbed Hepatitis B Vaccine 本品は B 型肝炎ウイルスの表面抗原を含む液にアルミニウム塩を加えて B 型肝炎ウイルスの表面抗原を不溶性とした液状の注射剤である．本品は生物学的製剤基準の沈降 B 型肝炎ワクチンの条に適合する．性状本品は振り混ぜるとき，均等に白濁する．

Text A (en): Adsorbed Hepatitis B Vaccine This product is a liquid injectable preparation in which aluminum salts are added to a solution containing the surface antigen of the hepatitis B virus, rendering the antigen insoluble. It complies with the standards for the“Adsorbed Hepatitis B Vaccine” in the Standards for Biological Preparations. When shaken, the product becomes uniformly turbid.

Text B (ja): 型肝炎ワクチン Adsorbed Hepatitis B Vaccine 本品は B 型肝炎ウイルスの表面抗原を含む液にアルミニウム塩を加えて B 型肝炎ウイルスの表面抗原を不溶性とした液状の注射剤である．本品は生物学的製剤基準の沈降 B 型肝炎ワクチンの条に適合する．性状本品は振り混ぜるとき，均等に白濁する．主成分本品の主成分は，B 型肝炎ウイルス表面抗原 (HBsAg) である．HBsAg は，B 型肝炎ウイルスの表面抗原である．HBsAg は，10〜30 μ g/0.5mL に含まれる．副成分本品の副成分は，アルミニウム塩である．アルミニウム塩は，0.3〜1.5 mg/0.5mL に含まれる．製造方法本品は，B 型肝炎ウイルス表面抗原を含有する液体をアルミニウム塩を加えて不溶化し，ろ過，充填，封栓，包装，検査を行い，製造する．試験方法 HBsAg 含量の測定 HBsAg 含量は，ELISA 法にて測定する．HBsAg 含量の基準は，5 μ g/0.5mL 以上である．

Text B (en): Adsorbed Hepatitis B Vaccine This product is a liquid injectable preparation in which aluminum salts are added to a solution containing the surface antigen of the hepatitis B virus (HBV), rendering the antigen insoluble. It conforms to the standards for Adsorbed Hepatitis B Vaccine as specified in the Standards for Biological Preparations. Description When shaken, the product becomes uniformly turbid. Active Ingredient The active ingredient of this product is the hepatitis B virus surface antigen (HBsAg). HBsAg is the surface antigen of the hepatitis B virus. The HBsAg content is between 10 and 30 μg per 0.5 mL. Inactive Ingredient The inactive ingredient of this product is aluminum salt. The aluminum salt content is between 0.3 and 1.5 mg per 0.5 mL. Manufacturing Method This product is manufactured by adding aluminum salts to a solution containing hepatitis B surface antigen to render it insoluble, followed by filtration, filling, sealing, packaging, and inspection. Test Method: Measurement of HBsAg Content The HBsAg content is measured using the ELISA method. The standard for HBsAg content is not less than 5 μg per 0.5 mL.

Labels (ja): 数値の誤り, 判定基準の変更

Labels (en): Numerical error, Change in evaluation criteria

Figure 8: **Sample task from SogoCheck.**

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

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. This includes practical business data such as guidelines on the general manufacturing process or risk management plan.

General-domain corpus. A part of FineWeb (Penedo et al., 2024) and Swallow Dataset[8].

This data composition was a deliberate choice to ensure that the trained model would be capable of solving practical problems in daily situations while acquiring professional academic knowledge of pharmaceuticals. For practical reasons, the primary component of the corpus came from academic sources.

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 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[9] 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[10]).

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 was reduced from 10B to 9B.

C.4 Base Model Selection

Discussing industrial applications often leads to the cost perspective. Different from development for research-only purposes, the operational cost at the inference time also should be taken into account, otherwise no institution can afford to utilize the trained model. Therefore, we restricted the model

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

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

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

Training Settings	
Method	CPT
Base model	Qwen2.5-7B
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 7: **Details of model training settings.**

size to around 7B for better usability considering the training cost and inference cost.

Secondly, we prioritized the use of a pre-trained model that had already been trained on Japanese data, since training an English-only model to learn Japanese might be a difficult challenge. For example, our evaluation experiments demonstrated that the improvement from Qwen2.5-7B to Meditron3-Qwen2.5-7B on YakugakuQA was only marginal. This is possibly because Meditron3-Qwen2.5-7B is intended for use in English, so its capability of QA in Japanese may be limited.

Finally, we also sought a model with a commercially viable license that would facilitate its adoption within the pharmaceutical industry.

Based on these criteria, we chose Qwen2.5-7B as the base model.

Comparison with Qwen3 (Yang et al., 2025). We empirically observed significant degradation in the instruction following capability when we switched the base model to Qwen3. This is possibly because the Qwen3 series is post-trained with reinforcement learning to enhance the reasoning ability, which we suspect makes the instruction following ability elastic and vulnerable.

C.5 Training Settings

Training was conducted using standard autoregressive language modeling objectives with the original tokenizer of Qwen2.5. Table 7 provides an overview of the training configuration.

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 8: **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.

C.6 Enhancing Instruction Following via Model Merging

Our domain-specific model trained through CPT 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, 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 capabilities 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, pre-trained 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 8 shows the superiority of EvoLLM (Akiba et al., 2025) coupled with DARE TIES merging.