JamC-QA: A Multiple-Choice Question Answering Benchmark for Japan-Specific Knowledge

Teruaki Oka, Tomohide Shibata and Nao Yoshida
SB Intuitions Corp.
{teruaki.oka, tomohide.shibata, nao.yoshida}@sbintuitions.co.jp

Abstract

We introduce JamC-QA, a multiple-choice question answering benchmark specifically designed to evaluate Japan-specific knowledge. Existing Japanese QA benchmarks largely consist of questions translated from English or derived from professional exams, primarily targeting academic or generally shared knowledge. Consequently, this limits the usefulness of distinguishing the performance of high-performing Large Language Models on local knowledge acquisition. To address this, JamC-QA serves as a robust resource for assessing the acquisition of Japan-specific knowledge. It comprises 2,309 challenging instances that were created entirely from scratch by human annotators across eight categories: culture, custom, regional identity, geography, history, government, law, and healthcare. Instances that were easily answerable by weak models were filtered out. Evaluation results highlight the critical distinction between model types: while multilingual models scored highly on general benchmarks like MMLU and JMMLU, the results on JamC-QA indicate that they do not fully capture Japan-specific knowledge. Japanese-language models outperform multilingual models, especially on culture- and region-related knowledge such as proverbs, traditional events, and local customs. Furthermore, we find a notable division within Japanese models: models further pretrained on Japanese text excel at administrative and legal questions, while models trained from scratch perform strongly on local and cultural aspects.

Keywords: multiple-choice QA, Japan-specific knowledge, evaluation benchmark, Japanese language models, weak LLM filtering

1. Introduction

Large language models (LLMs) have rapidly transformed a wide range of fields and domains. Research interest has expanded beyond English, leading to the development of multilingual and language-specific models.

Our research focuses on Japanese, aiming to examine how effectively LLMs have acquired Japan-specific knowledge. In Japan, giving *Ochūgen* and *Oseibo* gifts in summer and winter to express gratitude is a customary practice, which involves specific etiquette. Institutions such as the child allowance, which is paid in even-numbered months, are also distinctive. A central research question is the extent to which multilingual LLMs have acquired the Japan-specific knowledge exemplified above, and whether they encode a greater proportion of such knowledge compared to Japanese LLMs.

Despite the growing development of LLMs in multiple languages, most existing benchmarks are English-centric, leaving a gap in evaluating LLM performance between English and non-English languages. As one example of addressing the gap, English benchmarks such as MMLU (Hendrycks et al., 2021) and MMLU-Pro (Wang et al., 2024) are often translated into other languages, including Japanese (Lai et al., 2023, Fenogenova et al., 2024; Singh et al., 2025; Xuan et al., 2025). MMLU mostly covers universally shared topics, such as astronomy or international law, thus translating them changes the language but not the underlying content. In this setting, models are effectively only solving universally shared problems and are not being tested on Japan-specific knowledge. This is illustrated in Figure 1. The score rankings of MMLU and its Japanese translation, JMMLU (All) (Yin et al., 2024), are very similar. This suggests that translation alone does not provide insight into whether LLMs have acquired knowledge specific to the Japanese context. Some of the instances in JMMLU (*Japanese questions*) were created from scratch and are not globally universal, but they are too simple to clearly reveal the performance gap between multilingual and Japanese LLMs.

Alternatively some benchmarks are constructed in the languages of each country based on exams (Li et al., 2024; Yüksel et al., 2024; Poh et al., 2024; Verma et al., 2025; Son et al., 2025; Romanou et al., 2025). For example, when evaluated using the Japanese subset of INCLUDE (Romanou et al., 2025)[1], an exam-based benchmark derived from driving licenses, medical exams, and professional certification exams, multilingual models and Japanese continued pretrained models from LLaMA (Fujii et al., 2024) achieve high scores in the second group from the right in Figure 1. This shows a trend similar to that observed in MMLU and JMMLU.

This paper introduces **JamC-QA**, created from

[1] https://huggingface.co/datasets/CohereLabs/include-base-44

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 4536–4546
11–16 May 2026.

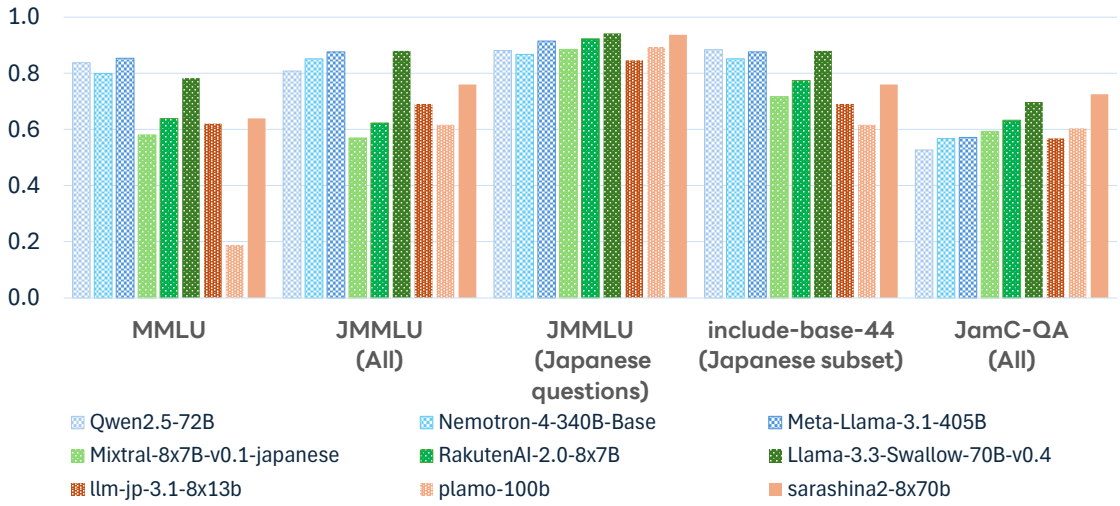

Figure 1: Comparison of model performance (accuracy) across five benchmarks: MMLU, JMMLU (All questions), JMMLU (Japanese questions), include-base-44 (Japanese subset), and JamC-QA (All questions). Qwen2.5-72B, Nemotron-4-340B-Base, and Meta-Llama-3.1-405B are multilingual models; Mixtral-8x7B-v0.1-japanese, RakutenAI-2.0-8x7B, and Llama-3.3-Swallow-70B-v0.4 are continued pretraining models; llm-jp-3.1-8x13b, plamo-100b, and sarashina2-70b are Japanese scratch models.

scratch to address the gap between benchmarks evaluating universally shared knowledge and those assessing Japan-specific knowledge. JamC-QA uses a multiple-choice question answering (**MCQ**) format, similar to MMLU. Each MCQ instance consists of a question, answer choices, and the correct choice. The benchmark comprises MCQ instances across eight categories: culture, custom, regional identity, geography, history, government, law, and healthcare. To ensure appropriate difficulty, instances that are easily answered by weak LLMs were automatically filtered out, whereas instances involving knowledge that is universally shared across languages or overlapping with other instances were manually excluded. These procedures yielded a challenging benchmark, as illustrated by the rightmost graph in Figure 1. Evaluation experiments demonstrate that Japanese scratch models excel in knowledge of proverbs and local events, whereas Japanese continued pretraining models perform better on administrative and legal knowledge.

Our contributions are as follows:

- Introduction of JamC-QA, a multiple-choice QA benchmark targeting Japan-specific knowledge.
- Provision of a high-quality benchmark of 2,309 MCQ instances, all created from scratch.
- Analysis of category-specific strengths and weaknesses of Japanese scratch models, Japanese continued pretraining models, and multilingual models.
- Facilitation of research on Japanese language models.

2. Related Work

A common approach to creating non-English benchmarks involves three methods: translating existing English benchmarks, using exam questions written in each target language, and constructing benchmarks entirely from scratch.

Translation-based benchmarks. The simplest approach to creating non-English benchmarks is to translate existing English benchmarks. MCQ English benchmarks such as MMLU (Hendrycks et al., 2021), MMLU-Pro (Wang et al., 2024), and MMLU-Redux (Gema et al., 2025) cover a wide range of academic and professional subjects, and have been translated into other languages to construct non-English benchmarks. In Lai et al. (2023), MMLU is translated into multiple languages fully automatically using ChatGPT. Global MMLU (Singh et al., 2025) and MMLU-ProX (Xuan et al., 2025) incorporate human corrections into the machine-translated output. MMMLU[2] relies entirely on professional human translations. Translation-based approaches have limitations: translated MCQ instances may be unnatural or contain errors, they tend to focus on universally shared knowledge, and they may implicitly carry English-speaking cultural

[2] https://huggingface.co/datasets/openai/MMMLU

norms, which can misalign with local contexts.

Exam-based benchmarks. To address these issues, some benchmarks are constructed from real exam instances from regions where each target language is used, without translation, such as KMMLU (Son et al., 2025), CMMLU (Li et al., 2024), ArabicMMLU (Koto et al., 2024), PersianMMLU (Ghahroodi et al., 2024), TurkishMMLU (Yüksel et al., 2024), MalayMMLU (Poh et al., 2024), and INCLUDE (Romanou et al., 2025). These benchmarks focus on the subjects covered in professional exams and rarely directly test region-specific knowledge. An analysis of the Japanese MCQ instances in INCLUDE, those from driving license, medical, and professional certification exams, reveals that they often contain knowledge shared with other countries, highlighting that exam-based benchmarks rarely assess region-specific culture. Evaluations using INCLUDE on Japanese knowledge also indicate that multilingual models achieve high scores, and continued pretraining models built upon them further enhance performance. An exception is MILU (Verma et al., 2025), which includes instances on regional culture, history, traditions, and festivals and similar benchmark for Japan is needed, but none currently exists.

From-scratch benchmarks. There are also efforts to create benchmarks from scratch. The closed test split of ruMMLU (Fenogenova et al., 2024) and some categories in JMMLU (Yin et al., 2024) include expert-authored MCQ instances following MMLU categories. In JMMLU, instances based on Japan-specific knowledge are included in four categories: japanese_idiom, japanese_history, japanese_geography, and japanese_civics. These benchmarks primarily assess academic knowledge and do not directly evaluate broader local customs or everyday cultural practices. Many of the MCQ instances can already be answered correctly by recent high-performing LLMs, which limits their usefulness in distinguishing model performance (see Figure 1).

3. Design of JamC-QA

This section describes JamC-QA, a benchmark consisting solely of instances that assess knowledge specific to Japan. Unlike existing benchmarks, which are translations of English benchmarks or based on exam questions and include knowledge common across regions, JamC-QA was created manually from scratch and its questions are generally more challenging. To avoid the wide range of possible answers associated with open-ended questions, JamC-QA adopts MCQ format, similar to MMLU.

Although the *Japanese questions* in JMMLU were created from scratch, their question categories still follow those of MMLU. The categories of MMLU were primarily designed for educational and academic domains and thus do not align with our goal of evaluating knowledge specific to Japan. To measure the extent to which models acquire knowledge specific to Japan, we newly define eight categories for JamC-QA: culture, custom, regional identity, geography, history, government, law, and healthcare. While MMLU also includes categories such as geography, history, and law, it focuses on the geography and history of English-speaking regions, as well as international law. In contrast, JamC-QA focuses specifically on Japanese history, Japanese geography, and Japanese law. Although Japanese history and geography are included in the Japanese questions of JMMLU, these can already be answered correctly by recent high-performing LLMs; thus JamC-QA aims to provide more challenging instances.

The categories are defined as follows:

Culture: Covers knowledge related to uniquely Japanese cultural aspects, including language, traditional culture, and subculture.

Custom: Focuses on Japanese customs, manners, etiquette, and proper behaviors in daily life.

Regional Identity: Assesses knowledge rooted in local regions, such as regional festivals, famous sites, and local legends.

Geography: Covers knowledge typically taught in school social studies or geography, but includes instances that are more difficult than standard academic exams.

History: Covers knowledge typically taught in school social studies or Japanese history classes, with instances that are more challenging, comparable to those in Geography.

Government: Focuses on administrative services and policies relevant and accessible to everyday life in Japan.

Law: Covers Japanese legal statutes and their enforcement. While other benchmarks may focus on international law, this category covers Japanese law, including the Constitution, Civil Code, and Penal Code.

Healthcare: Includes knowledge of medical care and nursing available in everyday life in Japan, including practical knowledge of medical facilities, drugstores, and emergency situations.

Examples of MCQ instances from each category are provided in Table 1 for clarity.

Culture	
Question:	小春日和が表す季節を選べ (Choose the season that "*Koharu-biyori*" refers to.)
Choices:	冬から春, 秋から冬, 夏から秋, 春から夏 (From winter to spring, From autumn to winter, From summer to autumn, From spring to summer)
Custom	
Question:	快気祝いに贈る品物として適切なものを選べ (Choose the appropriate item to give as a gift for "*Kaiki-iwai*" (a celebration of recovery from illness).")
Choices:	鉢植え, 手鏡, 入浴剤, マグカップ (Potted plant, Hand mirror, Bath salts, Mug cup)
Regional Identity	
Question:	どじょう鍋の発祥の地はどれ? (Which region is the birthplace of "dojõ-nabe" (a Japanese loach hot pot dish) ?)
Choices:	京都府, 東京都, 大阪府, 福岡県 (Kyoto Prefecture, Tokyo Metropolis, Osaka Prefecture, Fukuoka Prefecture)
Geography	
Question:	海に接していない都道府県の数を選べ (Choose the number of Japanese prefectures that do not border the sea.)
Choices:	6 個, 5 個, 8 個, 7 個 (Six, Five, Eight, Seven)
History	
Question:	源氏と平氏の合戦の回数を選択肢の中から選べ (Choose the number of battles fought between the Genji and Taira clans.)
Choices:	7 回, 13 回, 17 回, 3 回 (7 times, 13 times, 17 times, 3 times)
Government	
Question:	2024 年現在、児童手当の支給月はいつか選べ (As of 2024, choose the months in which the "*jidō teate* (child allowance)" is paid.)
Choices:	年末のみ, 毎月, 奇数月, 偶数月 (Only at the end of the year, Every month, Odd-numbered months, Even-numbered months)
Law	
Question:	憲法第 16 条において、国民参政の一つの形態として権利が認められ、国会に提出されているものはどれ? (Under Article 16 of the Constitution, which of the following, recognized as a form of citizen participation, can be submitted to the National Diet?)
Choices:	陳情書, 意見書, 質問主意書, 請願書 (Appeal letter, Statement of opinion, Written question for the government, Petition)
Healthcare	
Question:	生後 6 ヵ月から接種できる定期接種ワクチンを選べ (Choose the routine vaccine that can be administered from six months of age.)
Choices:	インフルエンザ, 新型コロナ, 日本脳炎, おたふくかぜ (Influenza, COVID-19, Japanese encephalitis, Mumps)

Table 1: Example MCQ instances from each JamC-QA category, presented in Japanese along with English translations. Correct choices are indicated in red.

4. Construction of JamC-QA

The MCQ instances in JamC-QA were created by 21 in-house annotators. They received category definitions and examples, along with the following guidelines:

- The text of each question and choices should not be directly copied from Wikipedia, because LLMs may have been trained on it. Wikipedia may be consulted only to verify the MCQ instances.
- Geography and history instances should go beyond standard academic difficulty, as models are assumed to already possess textbook-level knowledge.
- Annotators should not use any LLMs (e.g., ChatGPT, Gemini) when creating instances, to prevent the instances from being biased against particular LLMs.

Although annotators were instructed to create challenging instances, the actual difficulty of the instances, as well as the performance achieved by the models, can only be assessed through evaluation with LLMs, revealing that some of the created instances were easily solved. Following Wang et al. (2024), a weak LLM filtering method was adopted to remove instances that were excessively easy. Any instances correctly answered by at least half

(4 or more) of the following eight weak LLMs were excluded:

- Qwen3-1.7B-Base (Yang et al., 2025)
- calm2-7b[3]
- llm-jp-3-3.7b (LLM-jp et al., 2024)
- Meta-Llama-3.1-8B (Grattafiori et al., 2024)
- Mistral-7B-v0.1 (Jiang et al., 2023)
- Minitron-8B-Base (Muralidharan et al., 2024)
- gemma-2-baku-2b[4]
- sarashina2.2-0.5b[5]

All 5,686 initial instances were created, and 793 of them were subsequently removed through weak-LLM filtering. For example, the following History instance was removed because it was too easy:

Question: 古事記に記載がある日本初代の天皇といえば誰か
(Who is considered the first emperor of Japan according to the *Kojiki*?)
Choices: 懿徳天皇, 神武天皇, 安寧天皇, 綏靖天皇
(*Emperor Itoku*, *Emperor Jimmu*, *Emperor Annei*, *Emperor Suizei*)

Following the filtering, we implemented a rigorous multi-stage verification to ensure benchmark reliability. Our proofreading team first corrected typographical errors and refined phrasing. Subsequently, two authors independently inspected all remaining candidates to verify factual accuracy, eliminate ambiguity, and remove unsuitable, non-Japan-specific, or duplicate instances. We also excluded misaligned items and added date information to time-sensitive questions. Any disagreements were resolved through formal adjudication. This comprehensive review resulted in a final benchmark of 2,309 instances, each with a single, indisputable correct answer.

Table 2 shows the distribution of instances across categories. The number of instances varies widely: the culture category has the most instances (640), while healthcare has the fewest (48) due to the specialized knowledge required.

5. Experiments

To assess how well language models acquire specific knowledge of Japanese, we conducted an evaluation using JamC-QA.

[3] https://huggingface.co/cyberagent/calm2-7b
[4] https://huggingface.co/rinna/gemma-2-baku-2b
[5] https://huggingface.co/sbintuitions/sarashina2.2-0.5b

Category	Number of Questions
Culture	640
Custom	200
Regional Identity	397
Geography	272
History	343
Government	110
Law	299
Healthcare	48
Total	**2,309**

Table 2: Number of instances per category in JamC-QA.

5.1. Experimental Settings

The evaluation was conducted using the following LLMs:

Closed-source Multilingual models:

- gpt-5_reasoning-high
- gpt-5_reasoning-medium

Open-source Multilingual models:

- gpt-oss-120b_reasoning-high(OpenAI, 2025)
- gpt-oss-120b_reasoning-medium (OpenAI, 2025)
- Meta-Llama-3.1-405B (Grattafiori et al., 2024)
- Nemotron-4-340B-Base (Nvidia et al., 2024)
- Qwen2.5-72B (Qwen et al., 2025)

Open-source Japanese continued pretraining models:

- Llama-3.3-Swallow-70B-v0.4(Fujii et al., 2024)
- RakutenAI-2.0-8x7B[6]
- Mixtral-8x7B-v0.1-japanese[7]

Open-source Japanese scratch models:

- sarashina2-8x70b[8]
- plamo-100b (Preferred Elements et al., 2024)
- llm-jp-3.1-8x13b (Nakamura et al., 2025)

Except for the OpenAI models, all models are base models. Japan-specific knowledge is extensive and cannot be fully acquired through additional supervised fine-tuning (SFT) or reinforcement learning (RL) alone. This study evaluates

[6] https://huggingface.co/Rakuten/RakutenAI-2.0-8x7B
[7] https://huggingface.co/abeja/Mixtral-8x7B-v0.1-japanese
[8] https://huggingface.co/sbintuitions/sarashina2-8x70b

the extent to which models acquire knowledge during pretraining on large-scale text corpora. The base models are the highest-scoring models from organizations, irrespective of their size, based on overall accuracy on JamC-QA. The gpt-5 models were evaluated on August 7, 2025.

All base models were evaluated under a few-shot learning setup. The 4-shot examples were kept consistent across evaluations to ensure reproducibility, avoiding random variations that could arise from the example selection. The template shown in Figure 2 is applied only in the culture category, with the same 4 examples used consistently for this category. Base models also adopt a direct string generation approach, in which they produce the full text of the selected choice rather than merely selecting a label. This ensures that base models are evaluated on their ability to select the correct choice itself, not merely produce a label.

OpenAI models were evaluated in a zero-shot setting. They used a label output setting (A, B, C, D) to avoid the complexity of extracting the selected choice string from the free-form chat text output.

The evaluation score was calculated as the percentage of instances for which the output produced by the model exactly matches the correct choice string or label. The reported accuracy for OpenAI models, which were executed using sampling, is averaged over three runs. Conversely generation for base models is performed using greedy search without any sampling to ensure reproducible and consistent output.

5.2. Results

Table 3 presents the evaluation results on JamC-QA. GPT-5 achieves the highest scores both overall (All) and across individual categories. GPT-5 has acquired more knowledge about Japan than other multilingual or Japanese models. The score gap between GPT-5 and gpt-oss shows that this result cannot be attributed solely to reasoning ability. Other multilingual models achieve lower scores, even if they are large in size. These scores are lower compared to Japanese scratch models or Japanese continued pretraining models.

Excluding GPT-5, the overall accuracy shows that the highest-scoring model among Japanese LLMs is sarashina2-8x70b. The LLMs with the next highest scores are those that underwent continued pretraining on Japanese text, including Llama-3.3-Swallow-70B-v0.4. Although sarashina2-8x70b is the top-performing model in most individual categories, Llama-3.3-Swallow-70B-v0.4 achieves the same score in the custom category. In the government and law categories, Llama-3.3-Swallow-70B-v0.4 outperforms sarashina2-8x70b. These results indicate that, even for Japan-specific knowledge, different models have distinct areas of strength. Compared with sarashina2-8x70b, Llama-3.3-Swallow-70B-v0.4 appears to be particularly strong on questions related to legal knowledge.

以下はタスクを説明する指示と、追加の背景情報を提供する入力の組み合わせです。要求を適切に満たす回答を書いてください。
(*The following is a combination of task instructions and input providing additional background information. Please provide an answer that appropriately fulfills the requirements.*)
指示:: 質問と回答の選択肢を入力として受け取り、選択肢から回答を選択してください。回答の他には何も含めないことを厳守してください。
(*Instruction:: Given a question and a set of answer choices, select the correct answer from the choices. Do not include anything other than the answer.*)

質問:: 秋の季語でないものを選択肢の中から選べ, 選択肢::
月見
紅葉
麗か
山粧う, 回答:: 麗か
質問:: 七段飾りの雛人形を飾る際、5 段目に並べるものはどれ?, 選択肢::
随身
御駕篭
箪笥
仕丁, 回答:: 仕丁
質問:: 萬歳楽・新年・東遊・越天楽のうち、催馬楽に当たるものはどれ, 選択肢::
萬歳楽
新年
越天楽
東遊, 回答:: 新年
質問:: 餅つきで杵を振る人はつき手と呼ぶが、もち米をひっくり返す人は何と呼ぶか選択肢から選べ, 選択肢::
返し手
押し手
こね手
もみ手, 回答:: 返し手
質問:: {{ question }}, 選択肢::
{{ choice0 }}
{{ choice1 }}
{{ choice2 }}
{{ choice3 }}, 回答::

Figure 2: Template for the 4-shot examples used in the culture category. The italicized English text has been added solely for clarity and is not part of the actual template.

6. Discussion

6.1. Acquisition of Japan-Specific Knowledge

To investigate the trends in the knowledge acquired by Japanese scratch models, Japanese

Model	All (2,309)	Cult. (640)	Cust. (200)	Reg.ID (397)	Geo. (272)	Hist. (343)	Govt. (110)	Law (229)	HC (48)
Closed-source Multilingual Models									
gpt-5_reasoning-medium	**0.847**	**0.864**	0.862	0.814	**0.835**	0.863	**0.839**	**0.835**	0.910
gpt-5_reasoning-high	0.845	0.862	**0.890**	**0.825**	0.809	**0.866**	0.830	0.807	**0.931**
Open-source Multilingual Models									
Meta-Llama-3.1-405B	**0.571**	0.558	0.545	0.484	0.500	**0.679**	0.646	**0.629**	**0.688**
Nemotron-4-340B-Base	0.567	**0.573**	**0.615**	**0.511**	0.467	0.595	**0.727**	0.582	0.667
Qwen2.5-72B	0.527	0.522	0.595	0.426	0.438	0.606	0.609	0.562	**0.688**
gpt-oss-120b_reasoning-high	0.523	0.506	0.527	0.392	0.539	0.603	0.536	0.593	**0.688**
gpt-oss-120b_reasoning-medium	0.517	0.502	0.557	0.410	**0.510**	0.570	0.585	0.556	**0.688**
Open-source Japanese Continued Pretraining Models									
Llama-3.3-Swallow-70B-v0.4	**0.697**	**0.689**	**0.775**	0.589	**0.566**	**0.776**	**0.773**	**0.783**	**0.854**
RakutenAI-2.0-8x7B	0.633	0.622	0.725	**0.617**	0.511	0.714	0.709	0.575	0.813
Mixtral-8x7B-v0.1-japanese	0.593	0.602	0.670	0.579	0.493	0.612	0.736	0.545	0.667
Open-source Japanese Scratch Models									
sarashina2-8x70b	**0.725**	**0.714**	**0.775**	**0.761**	**0.654**	**0.784**	**0.736**	**0.632**	**0.917**
plamo-100b	0.603	0.602	0.650	0.637	0.504	0.682	0.609	0.515	0.688
llm-jp-3.1-8x13b	0.568	0.595	0.635	0.582	0.449	0.589	0.627	0.502	0.625

Table 3: Evaluation results of various LLMs on JamC-QA. Values in parentheses in the header indicate the number of instances per category. The models are grouped into multilingual LLMs, continued pretraining models, and scratch models, and are sorted in descending order by their overall (All) scores.

continued pre-training models, and multilingual models. In this section, we compare these three categories of LLMs. Generally, the amount of Japanese text used during training increases in the order of multilingual < Japanese continued pretraining < Japanese scratch models. Our analysis confirms that LLMs trained on larger amounts of Japanese text are capable of solving more difficult instances, ranging from common proverbs and customs to advanced Japan-specific questions such as Japanese place names, the *Hyakunin Isshu* anthology, and slang. To demonstrate this, we selected the top-performing base models from each group: Meta-Llama-3.1-405B (hereafter, Llama), Llama-3.3-Swallow-70B-v0.4 (hereafter, Swallow), and sarashina2-8x70b (hereafter, Sarashina). The closed-source models GPT-5 are excluded from this comparison; despite their extremely high performance, their lack of details makes them unsuitable for analyzing output examples and tendencies.

First, MCQ instances that Sarashina and Swallow answered correctly, while the multilingual model Llama failed, are examined. These examples suggest that uniquely Japanese knowledge, such as proverbs, customs, and cultural events, is not sufficiently captured by multilingual models. The following are representative examples.

Question: ことわざ「夫婦喧嘩は○○も食わない」の○○に入る動物を選択肢から選べ
(Select the animal that completes the Japanese proverb "Even a ___ won't eat a quarrel between husband and wife" (A quarrel between husband and wife is nobody's business.).)
Choices: 猫, 虎, 蛇, 犬
(cat, tiger, snake, dog)
Llama: 猫

Question: 頭語「前略」の結語として正しいのはどれ?
(Which closing phrase is correct for the Japanese letter opening "*Zenryaku*" (without preamble)?)
Choices: 早々, 急迫, 敬白, 拝啓
(*Sōsō* (shortly/promptly), *Kyūhaku* (urgent), *Keihaku* (respectfully yours), *Haikei* (Dear Sir/Madam))
Llama: 敬白

Question: 選択肢の中で最も高い年齢のお祝いを選べ
(Select the celebration corresponding to the highest age among the options.)
Choices: 卒寿, 傘寿, 米寿, 白寿
(*Sotsuju* (90th birthday), *Sanju* (80th birthday), *Beiju* (88th birthday), *Hakuju* (99th birthday))
Llama: 卒寿

Next, instances that Sarashina answered correctly while both Swallow and Llama failed are examined. Although Swallow underwent continued pretraining on Japanese texts, it appears that knowledge of more specialized topics-such as obscure Japanese place names or poets included in the *Hyakunin Isshu anthology*—remains incompletely captured.

Question: 兵庫県の地名「安口」の正しい読み方は

どれ?
(What is the correct reading of the place name “安口” in Hyōgo Prefecture?)
Choices: はたくち, はだかす, あくち, あんこう
(*Hatakuchi*, *Hadakasu*, *Akuchi*, *Ankō*)
Swallow: はたくち
Llama: あくち

Question: 百人一首に歌が入っていない女流歌人を選択肢から選べ
(Select the female poet who does not have a poem included in the *Hyakunin Isshu anthology*.)
Choices: 清少納言, 額田王, 紫式部, 小野小町
(*Sei Shōnagon*, *Nukata no Ōkimi*, *Murasaki Shikibu*, *Ono no Komachi*)
Swallow: 清少納言
Llama: 清少納言

Finally, instances that none of the three models answered correctly are examined. These cases suggest that even Japanese-language models face difficulties with idiomatic expressions, slang, or visually grounded knowledge that is challenging to represent in text.

Question: 薄い酒という意味の隠語はどれ
(Which slang term refers to weak sake?)
Choices: むらさめ, あかうま, 水酉, 般若湯
(*Murasame*, *Akauma*, *Mizudori*, *Hannyatō*)
Sarashina: 水酉
Swallow: 般若湯
Llama: 水酉

Question: 交通標識のうち、四角形のものはどれ?
(Which of the following traffic signs is rectangular?)
Choices: 前方優先道路, 指定方向外進行禁止, 高さ制限, 安全地帯
(Priority Road, No Entry Except Specified Direction, Height Limit, Refuge island)
Sarashina: 指定方向外進行禁止
Swallow: 指定方向外進行禁止
Llama: 指定方向外進行禁止

The closed-source model gpt-5_reasoning-medium successfully solved the two instances. This result indicates that GPT-5 has acquired knowledge aspects not yet captured by open-source models.

Analysis of performance in each category shows that Swallow performed particularly well in the law category. This is likely because continued pretraining of Swallow selectively utilized high-quality academic texts, including legal documents, which contributed to strong performance in legal knowledge. The following is an example instance that Swallow answered correctly while Sarashina and Llama did not.

Question: 2024 年現在、労働基準法において、災害などの非常時にある労働者から請求があった場合、会社が労働者に支払うべきものとして定められているのはどれ?
(As of 2024, under the Labor Standards Act, which of the following is prescribed as the payment a company must provide to a worker upon request during emergencies such as disasters?)
Choices: 休業手当, すでに働いた分の賃金, 来月分の賃金, 平均賃金の 6 割以上の手当
(Leave allowance, Wages for work already performed, Wages for the following month, Allowance of at least 60% of the average wage)
Sarashina: 平均賃金の 6 割以上の手当
Llama: 休業手当

6.2. Analysis of Instance-Level Difficulty

We demonstrated in the experiment in Section 5 (Table 3) that our benchmark achieved an appropriate level of difficulty without excessively easy instances, by carefully designing the construction process and filtering out instances correctly answerable by weak LLMs. In this section, we investigate whether the difficulty of each instance is appropriately distributed and whether the benchmark effectively measures performance gaps between models. Specifically, we examine whether the benchmark contains a disproportionate number of either excessively easy instances or, conversely, overly difficult ones. Regarding the latter, we investigate whether these instances involve knowledge that remains to be acquired by future LLMs or instead stem from benchmark artifacts, such as question ambiguity or incorrect ground truth labels. Figure 3 illustrates the distribution of the number of models that correctly answered each instance. The horizontal axis represents the number of models that selected the correct choice, whereas the vertical axis indicates the number of instances. The instances are well-distributed across various difficulty levels, ranging from those correctly answered by only a few models to those answered by all. This distribution ensures that the benchmark maintains a sufficient variety of items without a disproportionate concentration of excessively easy or overly difficult instances.

A total of 436 instances were correctly answered by all models, corresponding to the rightmost bar in Figure 3. As detailed in Section 4, we employed several weak models to filter out excessively easy instances. Consequently, approximately 20% of the benchmark consists of instances that, while

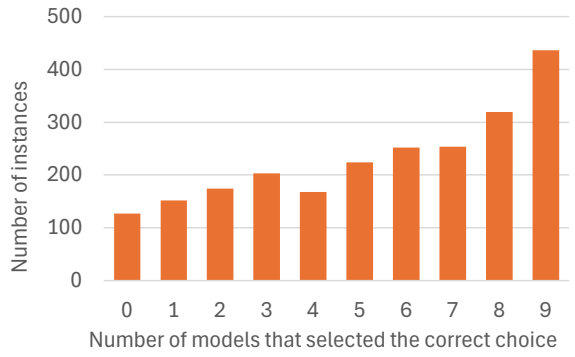

Figure 3: Distribution of correct answer counts. The horizontal axis represents the number of models that answered correctly, while the vertical axis indicates the total number of instances for each number of correct models.

challenging for weaker models, were correctly answered by the stronger models employed in our experiment, which possess larger parameter sizes and superior Japanese QA performance. This composition ensures that our benchmark remains demanding while providing sufficient ease of answering to evaluate even weaker models. A qualitative examination of these instances reveals that they include questions regarding the correct readings of kanji characters such as 百日紅 (*Sarusuberi*), 山茶花 (*Sazanka*), and 石楠花 (*Shakunage*), which are difficult to read yet frequently encountered in Japan. Additionally, the set contains instances focusing on widely recognized, uniquely Japanese business etiquette.

Question: 百日紅の読み方はどれ?
(Which is the correct reading of “百日紅”?)
Choices: さるすべり, ひゃくにちそう, さるばな, ひゃっかくれない
(*Sarusuberi*, *Hyakunichisou*, *Sarubana*, *Hyakkakurenai*)

Question: 会社宛に送る封筒で、宛名の企業名の後に書くものはどれ?
(Which honorific should follow the company name on an envelope addressed to a firm?)
Choices: 様, 御中, 殿, 各位
(*Sama*, *Onchu*, *Dono*, *Kakui*)

A total of 127 instances were not correctly answered by any of the models, corresponding to the leftmost bar in Figure 3. A qualitative examination of these instances confirms that the incorrect responses stem from the high difficulty of the questions, which involve regional aliases used only in certain parts of Japan or the specific contents of legal documents, rather than any flaws in the benchmark.

Question: 十三夜の異名として一部地域で呼ばれているものはどれ?
(Which of the following is an alias for Jusanya, the 13th night of the ninth lunar month, used in certain regions?)
Choices: 後の月, 栗の名月, 小麦の名月, 豆の名月
(*Nochi-no-tsuki*, *Kuri-no-meigetsu*, *Komugi-no-meigetsu*, *Mame-no-meigetsu*)

Question: 2024 年現在、食品衛生法第七章の見出しはどれ?
(As of 2024, what is the title of Chapter 7 of the Food Sanitation Act?)
Choices: 罰則, 検査, 監視指導, 食品及び添加物
(Penal Provisions, Inspections, Monitoring and Guidance, Food and Additives)

These observations confirm that JamC-QA encompasses a broad range of difficulty levels, ranging from knowledge-based questions that are challenging from an overseas perspective but well-known in Japan to highly specialized items requiring knowledge limited to specific regions or professions within the country.

7. Conclusion

This paper introduced JamC-QA, a benchmark focused on knowledge specific to Japan that presents a greater challenge than existing Japanese QA datasets. JamC-QA comprises 2,309 MCQ instances across eight categories, which cover knowledge not typically assessed in academic exams or standard tests. JamC-QA enables researchers to identify the strengths of Japanese-language models while highlighting areas where multilingual models exhibit insufficient coverage.

Future work includes:

- Increasing the number of instances in categories with relatively few items to reduce category imbalance.
- Subdividing the broad culture category into subcategories, for example, by referencing sections used in Japanese newspaper culture pages.
- Expanding the benchmark to include open-ended questions and other task formats such as translation and mathematics, that capture knowledge specific to Japan and not shared globally.

8. Data Availability

The benchmark developed in this study is publicly available on Hugging Face at https://huggingface.co/datasets/sbintuitions/JamC-QA under a Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0) license. By releasing this resource, we aim to facilitate further research and development of Japanese language models.

9. Bibliographical References

Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, and Kakeru Hattori et al.,. 2024. Continual pre-training for cross-lingual llm adaptation: Enhancing japanese language capabilities. Technical report.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, and Ahmad Al-Dahle et al.,. 2024. The llama 3 herd of models. Technical report.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, and Diego de las Casas et al.,. 2023. Mistral 7b. Technical report.

Viet Lai, Chien Nguyen, Nghia Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan Rossi, and Thien Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 318–327, Singapore. Association for Computational Linguistics.

LLM-jp, Akiko Aizawa, Eiji Aramaki, Bowen Chen, Fei Cheng, and Hiroyuki Deguchi et al.,. 2024. Llm-jp: A cross-organizational project for the research and development of fully open japanese llms. Technical report.

Saurav Muralidharan, Sharath Turuvekere Sreenivas, Raviraj Joshi, Marcin Chochowski, Mostofa Patwary, and Mohammad Shoeybi et al.,. 2024. Compact language models via pruning and knowledge distillation. Technical report.

Taishi Nakamura, Takuya Akiba, Kazuki Fujii, Yusuke Oda, Rio Yokota, and Jun Suzuki. 2025. Drop-upcycling: Training sparse mixture of experts with partial re-initialization. In *The Thirteenth International Conference on Learning Representations*.

Nvidia, Bo Adler, Niket Agarwal, Ashwath Aithal, Dong H. Anh, and Pallab Bhattacharya et al.,. 2024. Nemotron-4 340b technical report. Technical report.

OpenAI. 2025. gpt-oss-120b gpt-oss-20b model card. Technical report.

Preferred Elements, Kenshin Abe, Kaizaburo Chubachi, Yasuhiro Fujita, Yuta Hirokawa, and Kentaro Imajo et al.,. 2024. Plamo-100b: A ground-up language model designed for japanese proficiency. Technical report.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, and Bo Zheng et al.,. 2025. Qwen2.5 technical report. Technical report.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, and Bo Zheng et al.,. 2025. Qwen3 technical report. Technical report.

10. Language Resource References

Alena Fenogenova and Artem Chervyakov and Nikita Martynov and Anastasia Kozlova and Maria Tikhonova and Albina Akhmetgareeva et al.,. 2024. *MERA: A Comprehensive LLM Evaluation in Russian*. Association for Computational Linguistics.

Aryo Pradipta Gema and Joshua Ong Jun Leang and Giwon Hong and Alessio Devoto and Alberto Carlo Maria Mancino and Rohit Saxena et al.,. 2025. *Are We Done with MMLU?* Association for Computational Linguistics.

Omid Ghahroodi and Marzia Nouri and Mohammad Vali Sanian and Alireza Sahebi and Doratossadat Dastgheib and Ehsaneddin Asgari and Mahdieh Soleymani Baghshah and Mohammad Hossein Rohban. 2024. *Khayyam Challenge (Persian-MMLU): Is Your LLM Truly Wise to The Persian Language?*

Dan Hendrycks and Collin Burns and Steven Basart and Andy Zou and Mantas Mazeika and Dawn Song and Jacob Steinhardt. 2021. *Measuring Massive Multitask Language Understanding*.

Fajri Koto and Haonan Li and Sara Shatnawi and Jad Doughman and Abdelrahman Sadallah and Aisha Alraeesi et al.,. 2024. *ArabicMMLU: Assessing Massive Multitask Language Understanding in Arabic*. Association for Computational Linguistics.

Haonan Li and Yixuan Zhang and Fajri Koto and Yifei Yang and Hai Zhao and Yeyun Gong and Nan Duan and Timothy Baldwin. 2024. *CMMLU: Measuring massive multitask language understanding in Chinese*. Association for Computational Linguistics.

Soon Chang Poh and Sze Jue Yang and Jeraelyn Ming Li Tan and Lawrence Leroy Tze Yao Chieng and Jia Xuan Tan and Zhenyu Yu and Foong Chee Mun and Chee Seng Chan. 2024. *MalayMMLU: A Multitask Benchmark for the Low-Resource Malay Language*. Association for Computational Linguistics.

Angelika Romanou and Negar Foroutan and Anna Sotnikova and Sree Harsha Nelaturu and Shivalika Singh and Rishabh Maheshwary et al.,. 2025. *INCLUDE: Evaluating Multilingual Language Understanding with Regional Knowledge*.

Shivalika Singh and Angelika Romanou and Clémentine Fourrier and David Ifeoluwa Adelani and Jian Gang Ngui and Daniel Vila-Suero et al.,. 2025. *Global MMLU: Understanding and Addressing Cultural and Linguistic Biases in Multilingual Evaluation*. Association for Computational Linguistics.

Guijin Son and Hanwool Lee and Sungdong Kim and Seungone Kim and Niklas Muennighoff and Taekyoon Choi et al.,. 2025. *KMMLU: Measuring Massive Multitask Language Understanding in Korean*. Association for Computational Linguistics.

Sshubam Verma and Mohammed Safi Ur Rahman Khan and Vishwajeet Kumar and Rudra Murthy and Jaydeep Sen. 2025. *MILU: A Multi-task Indic Language Understanding Benchmark*. Association for Computational Linguistics.

Yubo Wang and Xueguang Ma and Ge Zhang and Yuansheng Ni and Abhranil Chandra and Shiguang Guo et al.,. 2024. *MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark*.

Weihao Xuan and Rui Yang and Heli Qi and Qingcheng Zeng and Yunze Xiao et al.,. 2025. *MMLU-ProX: A Multilingual Benchmark for Advanced Large Language Model Evaluation*.

Ziqi Yin and Hao Wang and Kaito Horio and Daisuke Kawahara and Satoshi Sekine. 2024. *Should We Respect LLMs? A Cross-Lingual Study on the Influence of Prompt Politeness on LLM Performance*. Association for Computational Linguistics.

Arda Yüksel and Abdullatif Köksal and Lütfi Kerem Senel and Anna Korhonen and Hinrich Schuetze. 2024. *TurkishMMLU: Measuring Massive Multitask Language Understanding in Turkish*. Association for Computational Linguistics.