Can Language Models Handle a Non-Gregorian Calendar? The Case of the Japanese *wareki*

Mutsumi Sasaki[1] **Go Kamoda**[2,3] **Ryosuke Takahashi**[1,4] **Kosuke Sato**[1]
Kentaro Inui[5,1,4] **Keisuke Sakaguchi**[1,4] **Benjamin Heinzerling**[4,1]
[1]Tohoku University [2]SOKENDAI [3]NINJAL [4]RIKEN [5]MBZUAI
Correspondence: mutsumi.sasaki@dc.tohoku.ac.jp

Abstract

Temporal reasoning and knowledge are essential capabilities for language models (LMs). While much prior work has analyzed and improved temporal reasoning in LMs, most studies have focused solely on the Gregorian calendar. However, many non-Gregorian systems, such as the Japanese, Hijri, and Hebrew calendars, are in active use and reflect culturally grounded conceptions of time. If and how well current LMs can accurately handle such non-Gregorian calendars has not been evaluated so far. Here, we present a systematic evaluation of how well language models handle one such non-Gregorian system: the Japanese *wareki*. We create datasets that require temporal knowledge and reasoning in using *wareki* dates. Evaluating open and closed LMs, we find that some models can perform calendar conversions, but GPT-4o, Deepseek V3, and even Japanese-centric models struggle with Japanese calendar arithmetic and knowledge involving *wareki* dates. Error analysis suggests corpus frequency of Japanese calendar expressions and a Gregorian bias in the model's knowledge as possible explanations. Our results show the importance of developing LMs that are better equipped for culture-specific tasks such as calendar understanding. [1]

1 Introduction

The training data of English-centric Language Models (LMs) predominantly assumes the Gregorian calendar as the default temporal framework. However, many cultures use non-Gregorian calendar systems such as the Islamic *hijri* calendar, the Hebrew calendar, or the Japanese calendar *wareki*. The Hijri calendar guides both religious observances and civil matters in Islamic countries, while the Hebrew calendar remains essential for Jewish religious traditions. The *wareki* system plays an important role in contemporary Japan, appearing in official documents, driver's licenses, and commemorative items. Hence, LMs need to be capable of handling such systems to achieve cultural competence. For example, since the *wareki* system is widely used in Japan, LMs encounter *wareki* dates in virtually all NLP tasks. For example, a Japanese information retrieval query like "retrieve all relevant documents from the last 10 years" might cross era boundaries (Fig. 1) and in cross-cultural settings involving different calendar systems, tasks like machine translation and cross-lingual information retrieval require translating dates across calendars.

Although the importance of incorporating cultural perspectives into language models is increasingly recognized (Shen et al., 2024; Pawar et al., 2024) and efforts have been made to build cultural commonsense benchmarks across languages and regions (Khairallah et al., 2024; Kim et al., 2024; Wang et al., 2024), little attention has been paid to culturally grounded temporal expressions such as calendars. Recent work has evaluated date arithmetic (Wang and Zhao, 2024; Gaere and Wangenheim, 2025; Chu et al., 2024), temporal reasoning (Chen et al., 2021), date format understanding (Bhatia et al., 2025a,b), and has analyzed the impact of tokenization (Bhatia et al., 2025a) and internal representations (Heinzerling and Inui, 2024; El-Shangiti et al., 2025) on calendar-based reasoning in LMs, but these efforts exclusively used the Gregorian calendar.

Here, we focus on *wareki* as a representative non-Gregorian system, motivated by its widespread use in Japan, its relative complexity, and the availability of both English- and Japanese-centric open models for cross-linguistic comparison. The *wareki* system divides time into eras, each starting from year 1. Since the end of an era is tied to major historical events such as imperial succession, their lengths are irregular; Meiji, Taisho, Showa, Heisei lasted 45, 15, 64, and 31 years, respectively, and the end of the

[1] github.com/cl-tohoku/Non-Gregorian-Calendar

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 444–463
December 20-24, 2025 ©2025 Association for Computational Linguistics

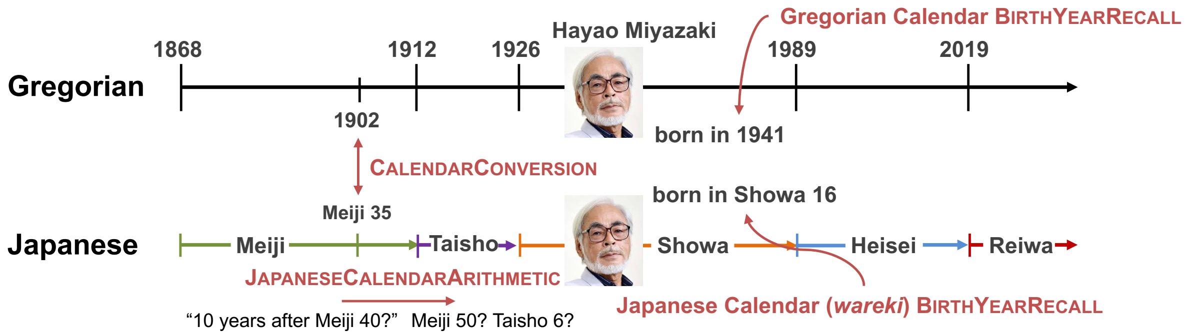

Figure 1: In the Japanese calendar (*wareki*), years are expressed using era names, which change irregularly according to historic events such as an emperor's accession. For example, the Reiwa era began on May 1, 2019, with the accession of Emperor Naruhito, so 2020 corresponds to Reiwa 2. In addition to showing the five eras of modern Japan (bottom) in relation to the Gregorian calendar (top), this figure illustrates three tasks designed to evaluate how LMs handle *wareki* system: (1) **CalendarConversion** between Gregorian calendar and *wareki*; (2) **JapaneseCalendarArithmetic** across era boundaries; (3) **BirthYearRecall** in both calendar systems.

current Reiwa is unknown. A further complication is that a Gregorian year can span two eras if an era transition occurs in that year. For example, the year 2019 corresponds to Heisei 31 until April 30 and then becomes Reiwa 1 from May 1. In non-transition years, a Gregorian year maps exactly to one *wareki* year, such as 2020 = Reiwa 2.

Given these properties of *wareki*, we designed three evaluation tasks in English and Japanese, focusing on both factual knowledge and reasoning (Fig. 1). Our evaluation of four English-centric open-source models, five Japanese-centric open-source models, and two frontier models shows that English-centric models face considerable difficulties with conversions and reasoning. In contrast, Japanese-centric models and the frontier models demonstrated relatively good performance in a simple format conversion task. Similarly, they struggled with more complex tasks, such as reasoning across era transitions and recalling birth years in the Japanese calendar. Furthermore, our error analysis revealed that the accuracy of the *wareki* reasoning task for each era is strongly correlated with the corpus frequency of *wareki* expressions, and that the performance in the *wareki* recall task is influenced by the Gregorian bias of the model's knowledge. Considering the widespread use of *wareki* by over 100 million people in Japan and Japanese being a high-resource language, our findings highlight a surprisingly low capability in Japanese-centric LMs. In a broader context, we hope to encourage further work aimed at evaluating and improving culture-specific temporal reasoning.

2 How well do English-centric and Japanese-centric LMs handle *wareki*?

To evaluate LMs' ability to handle the Japanese calendar *wareki*, we design three tasks that target distinct aspects of calendar reasoning: **CalendarConversion** (§ 2.1), **JapaneseCalendarArithmetic**(§ 2.2), and **BirthYearRecall**(§ 2.3). Our analysis, based on the synthetic data we created (see App. A for details), covers the five Japanese eras from 1868 to the present: Meiji, Taisho, Showa, Heisei, and Reiwa. We evaluate eleven language models in total, including five Japanese-centric models (llm-jp-3-13b (LLM-jp, 2024), sarashina2-13b, Swallow-13b (Fujii et al., 2024; Okazaki et al., 2024), Swallow-MS-7b, and Llama3-Swallow-8B), four English-centric models (Llama-2-7B, Llama-2-13B (Touvron et al., 2023), Mistral-7B (Jiang et al., 2023), and Llama3-8B (Grattafiori et al., 2024)), and two frontier models, GPT-4o (OpenAI et al., 2024) and DeepSeek V3 (DeepSeek-AI et al., 2025)(models details in App. C). Japanese-centric LMs are prompted in Japanese, while English-centric LMs are prompted in English, using few-shot prompts to encourage format adherence . For GPT-4o and DeepSeek V3, we use both Japanese and English few-shot prompts (prompt details in App. B).

We adopted greedy decoding with deterministic outputs. For CalendarConversion, we used a single prompt (since paraphrases did not appear to have any impact in preliminary experiments), while for JapaneseCalendarArithmetic, and BirthYearRecall, we used multiple prompt variants and report aggregated scores.

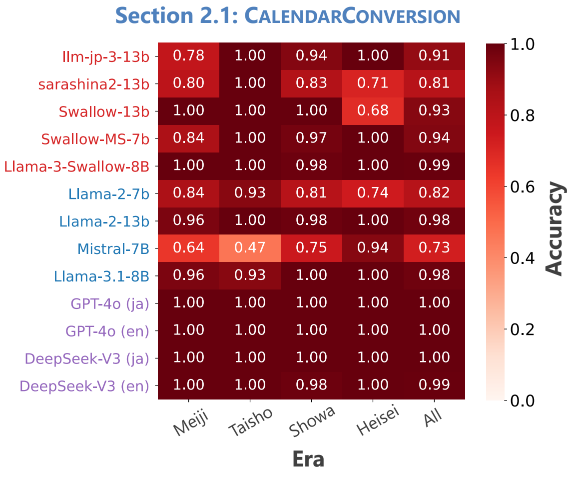

Figure 2: Performance on CALENDARCONVERSION (Gregorian→Japanese setting). Japanese-centric LMs (red labels) and frontier LMs (purple labels) perform nearly perfectly across all eras. Some English-centric LMs (blue labels) fail even at simple conversions.

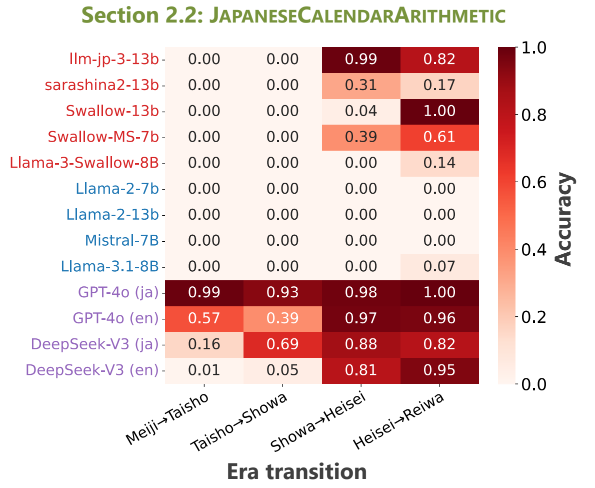

Figure 3: Performance on JAPANESECALENDARARITHMETIC. A large performance gap is observed between Japanese-centric LMs (red labels) and English-centric LMs (blue labels). Even frontier LMs (purple labels) struggle with this task.

2.1 CALENDARCONVERSION

Settings. This task evaluates the ability to convert years between the Gregorian calendar and *wareki*. We constructed a dataset of corresponding Gregorian and *wareki* years for the Meiji (1868–1912), Taisho (1912–1926), Showa (1926–1989), and Heisei (1989–2019) eras. In this task, an LM prompted "In the Japanese calendar, 1804 corresponds to Bunka 1. In the Japanese calendar, 1992 corresponds to" should output "Heisei 4".[2]

For each era, we measure conversion accuracy in both directions. For Gregorian targets, accuracy is defined as: $\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}(\hat{y}_i = y_i)$, where $\hat{y}_i$ and y_i are the predicted and target Gregorian year for instance i, N the number of instances, and $\mathbb{1}$ the indicator function. For *wareki* targets, accuracy is defined as: $\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}(\hat{E}_i = E_i \wedge \hat{x}_i = x_i)$, where E_i and x_i are the target era and year in the era (e.g. "Heisei" and "4" in "Heisei 4"), and $\hat{E}_i$ and $\hat{x}_i$ are the corresponding model predictions.

Results. Fig. 2 shows the accuracy of Gregorian-to-Japanese CALENDARCONVERSION. Japanese-centric models, GPT-4o, and DeepSeek V3 consistently achieved near-perfect accuracy across all eras, demonstrating strong conversion capability. In contrast, English-centric LMs show large variations in performance across models and eras. For example, 13B English LMs achieved over 90% accuracy across all eras, whereas Llama-2-7b and Mistral-7B achieved much lower accuracy.

[2]For fairness across eras, one-shot examples are from eras not included in our evaluation, e.g., Bunka, Koka, or Tenpo.

2.2 JAPANESECALENDARARITHMETIC

Settings. This task evaluates the ability to perform date arithmetic across *wareki* era boundaries. Specifically, we select a date within a five-year window before each era transition, as well as the date ten years after. We sampled 500 unique and non-overlapping dates for each era. In this task, LMs are required to answer the year that is ten years after the given input date. Each instance is constructed so that an era transition always occurs, from the input-side era (referred to as the pre-era) to the output-side era (referred to as the post-era). For example, when prompted "Ten years after March 8, Tenpo 14 is March 8, Koka 3. Ten years after September 19, Heisei 27 is" the LM should answer "September 19, Reiwa 7". We evaluate models using the accuracy. It is defined as the ratio of outputs that exactly match the correct date: $\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}(E_i = \hat{E}_i \wedge x_i = \hat{x}_i)$.

Results. Fig. 3 shows the results for JAPANESECALENDARARITHMETIC. All Japanese-centric and English-centric LMs showed low accuracy on early era transitions (Meiji→Taisho and Taisho→Showa). Even a frontier LM, DeepSeek V3, struggled with the era transitions of these era pairs. In contrast, for more recent transitions such as Heisei→Reiwa and Showa→Heisei, Japanese-centric LMs demonstrated notably better perfor-

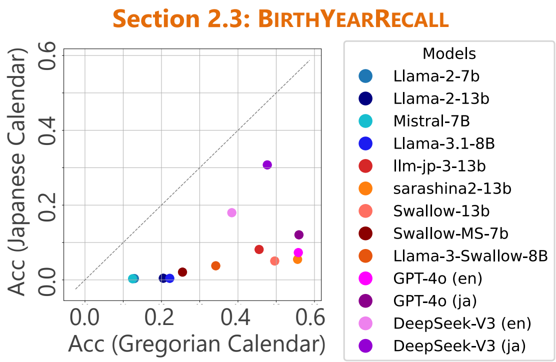

Figure 4: Comparison of BirthYearRecall accuracy in both Gregorian and *wareki* formats. The diagonal marks equal accuracy; below it indicates a Gregorian bias. Even Japanese-centric LMs and frontier LMs exhibit a strong bias towards the Gregorian calendar, and Japanese-centric LMs perform comparatively better with the Japanese calendar than English-centric LMs.

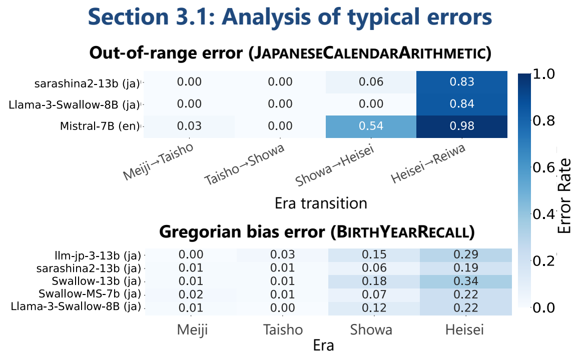

Figure 5: Analysis of why many models fail from the perspectives of typical error patterns. In JapaneseCalendarArithmetic, out-of-range errors (e.g., generating "Heisei 37") may contribute to failures in newer eras. In BirthYearRecall, Gregorian bias errors (responding in Gregorian years despite 3-shot *wareki* prompts) cause failures, especially in newer eras.

mance. For example, llm-jp-3-13b achieved accuracies of 0.99 and 0.82, respectively. On the other hand, English-centric LMs showed almost zero or very low accuracy even for recent eras.

Evaluation using a more lenient metric (App. E) revealed that the performance gap between Japanese and English models mainly stems from their handling of same-year transitions (e.g., Heisei 31 → Reiwa 1 in 2019). Specifically, English models often fail to recognize such transitions, whereas Japanese models tend to handle them correctly.

2.3 BirthYearRecall

Settings. This task measures the ability to recall the birth year of Japanese individuals. We extracted 300 Japanese individuals per era from Wikidata, filtering for entities with at least 20 relations. For example, given the three-shot prompt "According to the Japanese calendar, Ieyasu Tokugawa was born in Tenmon 11. $(\cdots)$. According to the Japanese calendar, Mao Asada was born in" the model should answer "Heisei 2". We evaluate models using accuracy, which is an exact match of the prediction and the target. For Gregorian output, accuracy is $\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}(\hat{y}_i = y_i)$. For *wareki* output, the prediction must match both the era and the year within the era: $\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}(E_i = \hat{E}_i \wedge x_i = \hat{x}_i)$.

Results. Fig. 4 compares the accuracy of BirthYearRecall in Gregorian (x-axis) versus Japanese calendar years (y-axis). The results indicate that all models exhibit a clear bias toward the Gregorian calendar, indicating that even Japanese-centric LMs mainly store birth years in the Gregorian format. While English-centric LMs perform poorly on *wareki* recall, Japanese-centric and frontier LMs show reasonable ability, though they still do better with the Gregorian calendar overall.

Moreover, evaluation using a more lenient metric (App. F) suggests that models may roughly recall the birth year at the era level or that minor shifts arise during internal conversions from the Gregorian to the Japanese calendar.

Furthermore, we also examined the consistency of BirthYearRecall across different calendar systems (Japanese and Gregorian) by measuring the percentage of individuals for whom the model also correctly recalled the Gregorian year, given that it had already correctly recalled the *wareki* birth year (App. G). As a result, while some Japanese-centric models achieved over 80% consistency, others remained around 50%, indicating that even among Japanese-centric models, there is substantial variation in how knowledge related to the Gregorian and Japanese calendars is recalled.

3 Discussion

3.1 Analysis of typical errors

To gain an understanding of model failure modes, we analyzed typical errors.

In the JapaneseCalendarArithmetic, we examined the proportion of out-of-range errors (e.g., generating "Heisei 37" even though the Heisei era ended at year 31). Fig. 5 (top) shows that such an error was particularly pronounced during

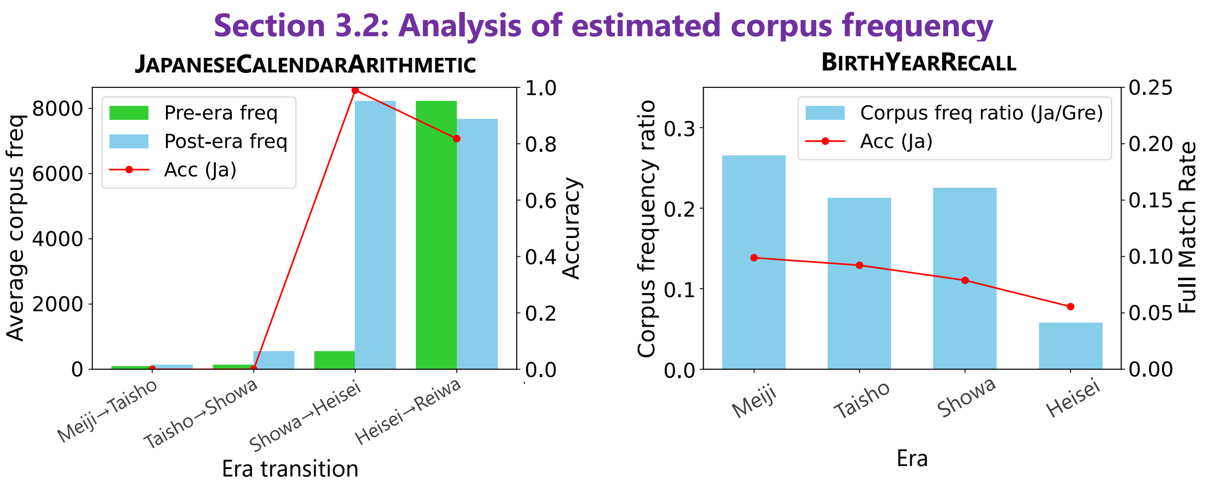

Figure 6: Analysis of why many models fail from the perspective of estimated corpus frequency. For llm-jp-3-13b, the only model with a public corpus, frequency analysis suggests that missing *wareki* expressions cause failures in JapaneseCalendarArithmetic, while Gregorian bias explains lower accuracy in BirthYearRecall.

the Heisei→Reiwa transition across several LMs. This result suggests that in later eras, out-of-range errors are one of the causes of failure.

For BirthYearRecall, we examined the proportion of Gregorian bias errors for each era. This analysis focused on cases where models, despite being given three-shot *wareki* examples designed to induce Japanese calendar responses, still produced answers in Gregorian years. As shown in Fig. 5 (bottom), this error is particularly pronounced in recent eras across all Japanese models, suggesting that Gregorian bias may hinder their ability to accurately recall years in *wareki* format.

3.2 Analysis of estimated corpus frequency

Among the models investigated in this study, we examined llm-jp-3-13b, the only Japanese model with a publicly available pretraining corpus, trained on the llm-jp-corpus-v3 (Enomoto et al., 2024). We used Infini-gram (Liu et al., 2024) to count year expressions in the pretraining corpus.

In JapaneseCalendarArithmetic, we analyzed the correlation between the frequency of *wareki* year occurrences and the accuracy of this task. We prepared Japanese expressions of *wareki* ranging from Meiji 1 to Reiwa 11 and measured their frequencies in the corpus. For each era, we calculated the average frequency across its constituent years. The results suggest a correlation between task accuracy and the corpus frequency of post-era Japanese calendar expressions (Fig. 6, left). In fact, the Pearson correlation coefficient between pre-era frequency and accuracy was 0.5086 ($p = 0.4914$), whereas that between post-era frequency and accuracy was much stronger, at 0.9959 ($p = 0.0041$). This indicates that the poor performance likely stems from the underrepresentation of post-era expressions in the pretraining data.

In contrast, no positive correlation was observed between the frequency of *wareki* year expressions and the accuracy in BirthYearRecall. To investigate further, we measured the frequencies of both *wareki* and corresponding Gregorian-year expressions in the corpus, averaged them within each era (from Meiji to Heisei), and calculated the ratio of Gregorian-year to *wareki* expression frequencies. This ratio showed a strong positive correlation with the accuracy of BirthYearRecall averaged over era (Fig. 6, right), and the Pearson correlation was 0.9367 ($p = 0.0633$). This suggests that in later eras, *wareki* expressions appear relatively less frequently in the corpus than Gregorian ones, which may underlie the models' failures.

4 Conclusions

This work analyzed whether LMs can handle the Japanese calendar, a non-Gregorian calendar. We evaluated models on tasks involving conversion, arithmetic, and factual recall. While Japanese-centric LMs and frontier LMs handle basic conversions, most models struggle with more complex tasks and show inconsistent behavior. Our findings highlight the need to understand the model limitations when dealing with non-Gregorian calendar systems and motivate future research investigating the causes of the revealed failures.

Limitations

While our work reveals LMs have difficulty in handling the Japanese calendar, it has several limitations.

First, our study focuses only on the Japanese calendar. Among our three tasks, CALENDARCONVERSION and BIRTHYEARRECALL can be extended to other calendars using date pairs or biographical data. On the other hand, JAPANESECALENDARARITHMETIC is specific to *wareki*, where eras change irregularly with imperial succession. We believe that, when extending this line of research to other calendar systems, it is crucial to design similarly system-specific tasks that reflect each calendar system's unique temporal structure. For example, with the Hijri calendar, one could ask which Gregorian season Ramadan falls in for a given year. The task will test whether models understand how lunar cycles shift the timing of Ramadan across seasons.

Second, our evaluation relies on prompt-based testing and subsequent error analysis based on typical error patterns and corpus frequency, rather than directly examining the internal representations of calendrical knowledge. While this analysis provides valuable insights into potential sources of error, it has limitations in identifying their causes in a direct, detailed manner. Future work could employ probing or mechanistic interpretability methods to more precisely identify error sources and propose remedies.

Third, among currently available Japanese-centric models, llm-jp-3-13b is the only one with publicly released training data. Therefore, the distribution of calendar-related expressions in other models' training data remains unknown. The release of pretraining data from more Japanese-centric models may help explain performance differences across models.

Ethical Considerations

All data created and/or used in this work was synthetically generated and/or derived from Wikidata, a public knowledge base released under the CC0 1.0 Universal license. As such, we do not foresee any ethical concerns regarding personally identifying information or offensive content. Also, the llm-jp-corpus-v3 (Enomoto et al., 2024) used in § 3 is publicly available.

All language models used in this study are publicly available. We strictly adhered to the terms and conditions of each model's license, including, but not limited to, those released under the Meta Llama and Mistral licensing terms.

During the development of code and the writing of this paper, we made use of AI assistants, including large language models. All code snippets and textual content generated with the assistance of such tools were carefully reviewed and revised by the authors to ensure scientific integrity, accuracy, and ethical compliance.

Acknowledgements

This work was supported by JST BOOST Grant Numbers JPMJBS2412, JPMJBS2421, JPMJBY24F9; JST CREST Grant Number JPMJCR20D2; JSPS KAKENHI Grant Number JP25K03175; AMED Grant Number JP25wm0625405, and the Nakajima Foundation.

References

Gagan Bhatia, Maxime Peyrard, and Wei Zhao. 2025a. Date fragments: A hidden bottleneck of tokenization for temporal reasoning. *Preprint*, arXiv:2505.16088.

Gagan Bhatia, Ming Ze Tang, Cristina Mahanta, and Madiha Kazi. 2025b. DateLogicQA: Benchmarking temporal biases in large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 321–332, Albuquerque, USA. Association for Computational Linguistics.

Wenhu Chen, Xinyi Wang, William Yang Wang, and William Yang Wang. 2021. A dataset for answering time-sensitive questions. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Haotian Wang, Ming Liu, and Bing Qin. 2024. TimeBench: A comprehensive evaluation of temporal reasoning abilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1204–1228, Bangkok, Thailand. Association for Computational Linguistics.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Ahmed Oumar El-Shangiti, Tatsuya Hiraoka, Hilal AlQuabeh, Benjamin Heinzerling, and Kentaro Inui.

2025. The geometry of numerical reasoning: Language models compare numeric properties in linear subspaces. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 550–561, Albuquerque, New Mexico. Association for Computational Linguistics.

Rintaro Enomoto, Arseny Tolmachev, Takuro Niitsuma, Shuhei Kurita, and Daisuke Kawahara. 2024. Investigating web corpus filtering methods for language model development in Japanese. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 154–160, Mexico City, Mexico. Association for Computational Linguistics.

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)*.

Edward Gaere and Florian Wangenheim. 2025. Datetime: A new benchmark to measure llm translation and reasoning capabilities. *Preprint*, arXiv:2504.16155.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Benjamin Heinzerling and Kentaro Inui. 2024. Monotonic representation of numeric attributes in language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 175–195, Bangkok, Thailand. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Christian Khairallah, Salam Khalifa, Reham Marzouk, Mayar Nassar, and Nizar Habash. 2024. Camel morph MSA: A large-scale open-source morphological analyzer for Modern Standard Arabic. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2683–2691, Torino, Italia. ELRA and ICCL.

Eunsu Kim, Juyoung Suk, Philhoon Oh, Haneul Yoo, James Thorne, and Alice Oh. 2024. CLIcK: A benchmark dataset of cultural and linguistic intelligence in Korean. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3335–3346, Torino, Italia. ELRA and ICCL.

Jiacheng Liu, Sewon Min, Luke Zettlemoyer, Yejin Choi, and Hannaneh Hajishirzi. 2024. Infini-gram: Scaling unbounded n-gram language models to a trillion tokens. In *First Conference on Language Modeling*.

LLM-jp. 2024. Llm-jp: A cross-organizational project for the research and development of fully open japanese llms. *Preprint*, arXiv:2407.03963.

Naoaki Okazaki, Kakeru Hattori, Hirai Shota, Hiroki Iida, Masanari Ohi, Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Rio Yokota, and Sakae Mizuki. 2024. Building a Large Japanese Web Corpus for Large Language Models. In *Proceedings of the First Conference on Language Modeling (COLM)*.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2024. Survey of cultural awareness in language models: Text and beyond. *Preprint*, arXiv:2411.00860.

Siqi Shen, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, Soujanya Poria, and Rada Mihalcea. 2024. Understanding the capabilities and limitations of large language models for cultural commonsense. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5668–5680, Mexico City, Mexico. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Xiaonan Wang, Jinyoung Yeo, Joon-Ho Lim, and Hansaem Kim. 2024. KULTURE bench: A benchmark for assessing language model in Korean cultural context. In *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, pages 914–927, Tokyo, Japan. Tokyo University of Foreign Studies.

Yuqing Wang and Yun Zhao. 2024. TRAM: Benchmarking temporal reasoning for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6389–6415, Bangkok, Thailand. Association for Computational Linguistics.

A Datasets

The datasets used in each task are presented in Tbl. 1 for CALENDARCONVERSION, Tbl. 2 and Tbl. 3 for JAPANESECALENDARARITHMETIC, and Tbl. 4 for BIRTHYEARRECALL.

For CALENDARCONVERSION, we constructed a dataset of corresponding Gregorian and *wareki* years for the Meiji (1868–1912), Taisho (1912–1926), Showa (1926–1989), and Heisei (1989–2019) eras.

For JAPANESECALENDARARITHMETIC, we sampled 500 unique and non-overlapping dates for each era. In the after-ten-years setting, dates were randomly selected from the last five years of the Meiji, Taisho, Showa, and Heisei eras. In the before-ten-years setting, dates were taken from the first five years of the Taisho, Showa, Heisei, and Reiwa eras. To ensure temporal precision at era boundaries, we carefully constructed the dataset such that, for example, Heisei dates end on April 30, Heisei 31, and Reiwa dates begin on May 1, Reiwa 1. This ensures that the model must reason across era boundaries to generate a correct answer.

For BIRTHYEARRECALL, we extracted 300 Japanese individuals per era from Wikidata, filtering for entities with at least 20 relations. In total, 1,200 individuals were sampled across the Meiji, Taisho, Showa, and Heisei eras. The distribution of birth year data for the individuals is shown in Fig. 7. Each entry includes the individual's name and birth year in both Japanese and English formats to accommodate prompts in both languages.

B Prompts

The prompts used for the four tasks in our experiments are summarized in Tbl. 5. For each task, we prepared both English and Japanese versions of the prompts. English prompts were used with English-centric models, and Japanese prompts were used with Japanese-centric models. In frontier models (GPT-4o and DeepSeekV3), since these models can handle both Japanese and English prompts, we use both Japanese and English prompts. To ensure consistency, we used the same user prompts as for the other models. Also, to induce responses in the intended format, we set the system prompt as follows:

Japanese: "あなたは和暦とグレゴリオ暦の専門家です。以下に続くように文章を答えのみ生成してください。"

English: "You are an expert in the Japanese and Gregorian calendars. Please generate only the answer that continues from the text below."

For CALENDARCONVERSION, the prompts request conversion between Gregorian and *wareki* dates, in both directions. A one-shot example was provided before the prompt to ensure that the model outputs the answer in the correct format, either as a four-digit year for Gregorian dates or as a combination of an era name and a year for Japanese dates.

For JAPANESECALENDARARITHMETIC, the prompts ask about the year ten years before or after a given date, often across era boundaries. As with the previous task, a one-shot example was shown in advance to guide the model toward the correct output format (e.g., “August 29, Reiwa 10”).

For BIRTHYEARRECALL, the prompts asked for the birth year of a Japanese individual in either the Gregorian calendar or the *wareki*. We provided three-shot examples before the prompt to help the model produce answers in the correct format, either as a four-digit number for Gregorian dates or as an era name followed by a year for Japanese dates.

C Models

We use four English-centric models: Llama-2-7B, Llama-2-13B (Touvron et al., 2023), Mistral-7B (Jiang et al., 2023), and Llama-3.1-8B (Grattafiori et al., 2024). For Japanese-centric models, we include two trained from scratch in Japanese (llm-jp-3-13b (LLM-jp, 2024) and sarashina2-13b), and three models continued pretraining in Japanese: Swallow-13b (Fujii et al., 2024; Okazaki et al., 2024), Swallow-MS-7b, and Llama3-Swallow-8B. All experiments were conducted using a single RTX 6000 Ada (48GB) GPU. Also, we used GPT-4o (OpenAI et al., 2024) and DeepSeek V3 (DeepSeek-AI et al., 2025) as comparative baselines for the frontier models.

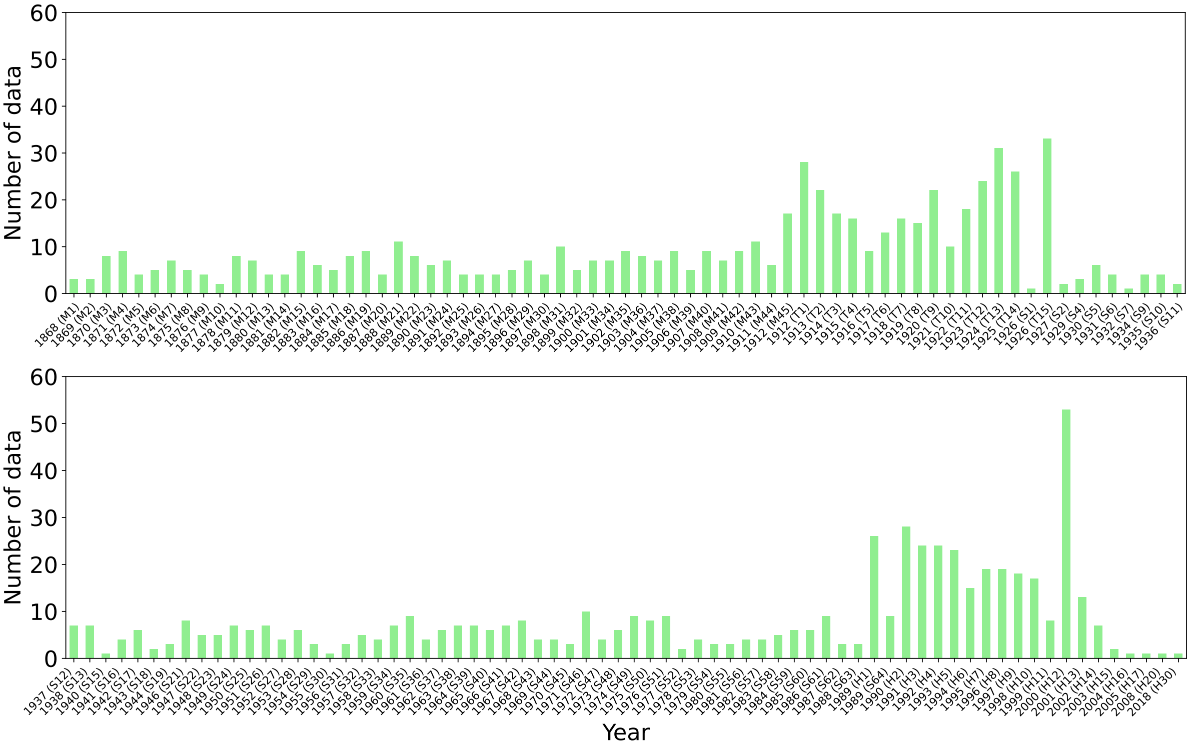

Figure 7: The distribution of birth-year data used in BIRTHYEARRECALL. The horizontal axis represents the year, and the vertical axis represents the number of data samples. The letters M, T, S, and H on the horizontal axis correspond to Meiji, Taisho, Showa, and Heisei, respectively. Years without bars indicate that there are zero samples for individuals born in that year. As mentioned in § 2.3, a total of 300 individuals were sampled for each era, resulting in 1,200 data samples in total.

Era	Lang	#	Gregorian	Japanese
Meiji	en	1	1868	Meiji 1
Meiji	en	2	1869	Meiji 2
⋮	⋮	⋮	⋮	⋮
Meiji	en	45	1912	Meiji 45
Meiji	ja	1	1868年	明治1年
Meiji	ja	2	1869年	明治2年
⋮	⋮	⋮	⋮	⋮
Meiji	ja	45	1912年	明治45年
Taisho	en	1	1912	Taisho 1
Taisho	en	2	1913	Taisho 2
⋮	⋮	⋮	⋮	⋮
Taisho	en	15	1926	Taisho 15
Taisho	ja	1	1912年	大正1年
Taisho	ja	2	1913年	大正2年
⋮	⋮	⋮	⋮	⋮
Taisho	en	15	1926年	大正15年
Showa	en	1	1926	Showa 1
Showa	en	2	1927	Showa 2
⋮	⋮	⋮	⋮	⋮
Showa	en	64	1989	Showa 64
Showa	ja	1	1926年	昭和1年
Showa	ja	2	1927年	昭和2年
⋮	⋮	⋮	⋮	⋮
Showa	ja	64	1989年	昭和64年
Heisei	en	1	1989	Heisei 1
Heisei	en	2	1990	Heisei 2
⋮	⋮	⋮	⋮	⋮
Heisei	en	31	2019	Heisei 31
Heisei	ja	1	1989年	平成1年
Heisei	ja	2	1990年	平成2年
⋮	⋮	⋮	⋮	⋮
Heisei	ja	31	2019年	平成31年

Table 1: Dataset used for CalendarConversion

Era	Lang	#	Date (Japanese calendar)	Gold date (Japanese calendar)
Meiji	en	1	August 24, Meiji 41	August 24, Taisho 7
Meiji	en	2	April 30, Meiji 42	April 30, Taisho 8
⋮	⋮	⋮	⋮	⋮
Meiji	en	500	August 1, Meiji 42	August 1, Taisho 8
Meiji	ja	1	明治41年8月24日	大正7年8月24日
Meiji	ja	2	明治42年4月30日	大正8年4月30日
⋮	⋮	⋮	⋮	⋮
Meiji	ja	500	明治42年8月1日	大正8年8月1日
Taisho	en	1	November 13, Taisho 12	November 13, Showa 8
Taisho	en	2	February 5, Taisho 12	February 5, Showa 8
⋮	⋮	⋮	⋮	⋮
Taisho	en	500	November 23, Taisho 15	November 23, Showa 11
Taisho	ja	1	大正12年11月13日	昭和8年11月13日
Taisho	ja	2	大正12年2月5日	昭和8年2月5日
⋮	⋮	⋮	⋮	⋮
Taisho	ja	500	大正15年11月23日	昭和11年11月23日
Showa	en	1	December 11, Showa 63	December 11, Heisei 10
Showa	en	2	September 22, Showa 62	September 22, Heisei 9
⋮	⋮	⋮	⋮	⋮
Showa	en	500	April 7, Showa 63	April 7, Heisei 10
Showa	ja	1	昭和63年12月11日	平成10年12月11日
Showa	ja	2	昭和62年9月22日	平成9年9月22日
⋮	⋮	⋮	⋮	⋮
Showa	ja	500	昭和63年4月7日	平成10年4月7日
Heisei	en	1	November 9, Heisei 28	November 9, Reiwa 8
Heisei	en	2	December 24, Heisei 30	December 24, Reiwa 10
⋮	⋮	⋮	⋮	⋮
Heisei	en	500	February 27, Heisei 31	February 27, Reiwa 11
Heisei	ja	1	平成28年11月9日	令和8年11月9日
Heisei	ja	2	平成30年12月24日	令和10年12月24日
⋮	⋮	⋮	⋮	⋮
Heisei	ja	500	平成31年2月27日	令和11年2月27日

Table 2: Dataset used for JapaneseCalendarArithmetic (add ten years)

Era	Lang	#	Date (Japanese calendar)	Gold date (Japanese calendar)
Taisho	en	1	August 28, Taisho 5	August 28, Meiji 39
Taisho	en	2	December 22, Taisho 4	December 22, Meiji 38
⋮	⋮	⋮	⋮	⋮
Taisho	en	500	September 20, Taisho 4	September 20, Meiji 38
Taisho	ja	1	大正5年8月28日	明治39年8月28日
Taisho	ja	2	大正4年12月22日	明治38年12月22日
⋮	⋮	⋮	⋮	⋮
Taisho	ja	500	大正4年9月20日	明治38年9月20日
Showa	en	1	January 21, Showa 6	January 21, Taisho 10
Showa	en	2	October 29, Showa 6	October 29, Taisho 10
⋮	⋮	⋮	⋮	⋮
Showa	en	500	February 19, Showa 2	February 19, Taisho 6
Showa	ja	1	昭和6年1月21日	大正10年1月21日
Showa	ja	2	昭和6年10月29日	大正10年10月29日
⋮	⋮	⋮	⋮	⋮
Showa	ja	500	昭和2年2月19日	大正6年2月19日
Heisei	en	1	January 25, Heisei 1	January 25, Showa 54
Heisei	en	2	September 4, Heisei 1	September 4, Showa 54
⋮	⋮	⋮	⋮	⋮
Heisei	en	500	May 29, Heisei 1	May 29, Showa 54
Heisei	ja	1	平成1年1月25日	昭和54年1月25日
Heisei	ja	2	平成1年9月4日	昭和54年9月4日
...	...	...	...	...
Heisei	ja	500	平成1年5月29日	昭和54年5月29日
Reiwa	en	1	July 7, Reiwa 4	July 7, Heisei 24
Reiwa	en	2	September 19, Reiwa 3	September 19, Heisei 23
⋮	⋮	⋮	⋮	⋮
Reiwa	en	500	July 4, Reiwa 5	July 4, Heisei 25
Reiwa	ja	1	令和4年7月7日	平成24年7月7日
Reiwa	ja	2	令和3年9月19日	平成23年9月19日
⋮	⋮	⋮	⋮	⋮
Reiwa	ja	500	令和5年7月4日	平成25年7月4日

Table 3: Dataset used for JapaneseCalendarArithmetic (subtract ten years)

Era	Lang	#	Name	Birth year (Gregorian)	Birth year (Japanese)
Meiji	en	1	Bunji Tsushima	1898	Meiji 31
Meiji	en	2	Heinosuke Gosho	1902	Meiji 35
⋮	⋮	⋮	⋮	⋮	⋮
Meiji	en	300	Kikuko, Princess Takamatsu	1911	Meiji 44
Meiji	ja	1	津島文治	1898年	明治31年
Meiji	ja	2	五所平之助	1902年	明治35年
⋮	⋮	⋮	⋮	⋮	⋮
Meiji	ja	300	宣仁親王妃喜久子	1911年	明治44年
Taisho	en	1	Tetsuo Takaha	1926	Taisho 15
Taisho	en	2	Kiyoshi Ito	1915	Taisho 4
⋮	⋮	⋮	⋮	⋮	⋮
Taisho	en	300	Yozo Matsushima	1921	Taisho 10
Taisho	ja	1	高羽哲夫	1926年	大正15年
Taisho	ja	2	伊藤清	1915年	大正4年
⋮	⋮	⋮	⋮	⋮	⋮
Taisho	ja	300	松島与三	1921年	大正10年
Showa	en	1	Hiroshi Katsuno	1949	Showa 24
Showa	en	2	Homare Sawa	1978	Showa 53
⋮	⋮	⋮	⋮	⋮	⋮
Showa	en	300	Hideki Matsui	1974	Showa 49
Showa	ja	1	勝野洋	1949年	昭和24年
Showa	ja	2	澤穂希	1978年	昭和53年
⋮	⋮	⋮	⋮	⋮	⋮
Showa	ja	300	松井秀喜	1974年	昭和49年
Heisei	en	1	Miyuri Shimabukuro	1994	Heisei 6
Heisei	en	2	Sakura Miyawaki	1998	Heisei 10
⋮	⋮	⋮	⋮	⋮	⋮
Heisei	en	300	Maimi Yajima	1992	Heisei 4
Heisei	ja	1	島袋美由利	1994年	平成6年
Heisei	ja	2	宮脇咲良	1998年	平成10年
⋮	⋮	⋮	⋮	⋮	⋮
Heisei	ja	300	矢島舞美	1992年	平成4年

Table 4: Dataset used for BIRTHYEARRECALL

Task Type	Lang	Option	#	Prompt(example)
CalendarConversion	en	GtoJ	1	In the Japanese calendar, the year 1992 corresponds to
CalendarConversion	en	JtoG	1	In the Gregorian calendar, Heisei 4 corresponds to the year
CalendarConversion	ja	GtoJ	1	平成4年を西暦に変換すると、
CalendarConversion	ja	JtoG	1	1992年を和暦に変換すると、
JapaneseCalendarArithmetic	en	+10yr	1	Ten years after August 29, Heisei 30 is
JapaneseCalendarArithmetic	en	+10yr	2	If you go forward 10 years from August 29, Heisei 30, you get
JapaneseCalendarArithmetic	en	+10yr	3	The date 10 years after August 29, Heisei 30 is
JapaneseCalendarArithmetic	en	-10yr	1	Ten years before April 27, Heisei 3 is
JapaneseCalendarArithmetic	en	-10yr	2	If you go back 10 years from April 27, Heisei 3, you get
JapaneseCalendarArithmetic	en	-10yr	3	The date 10 years prior to April 27, Heisei 3 is
JapaneseCalendarArithmetic	ja	+10yr	1	平成30年8月29日の10年後は
JapaneseCalendarArithmetic	ja	+10yr	2	平成30年8月29日から10年経つと
JapaneseCalendarArithmetic	ja	+10yr	3	平成30年8月29日に対する10年後の日付は
JapaneseCalendarArithmetic	ja	-10yr	1	平成3年4月27日の10年前は
JapaneseCalendarArithmetic	ja	-10yr	2	平成3年4月27日から10年さかのぼると
JapaneseCalendarArithmetic	ja	-10yr	3	平成3年4月27日に至る10年前の日付は
BirthYearRecall	en	G	1	According to the Gregorian calendar, Hideki Matsui was born in
BirthYearRecall	en	G	2	The Gregorian calendar states that Hideki Matsui was born in
BirthYearRecall	en	G	3	The Gregorian calendar dates Hideki Matsui's birth to
BirthYearRecall	en	J	1	According to the Japanese calendar, Hideki Matsui was born in
BirthYearRecall	en	J	2	The Japanese calendar states that Hideki Matsui was born in
BirthYearRecall	en	J	3	The Japanese calendar dates Hideki Matsui's birth to
BirthYearRecall	ja	G	1	西暦で松井秀喜が生まれたのは
BirthYearRecall	ja	G	2	松井秀喜の誕生年は
BirthYearRecall	ja	G	3	松井秀喜の生まれ年は
BirthYearRecall	ja	J	1	和暦で松井秀喜が生まれたのは
BirthYearRecall	ja	J	2	松井秀喜の誕生年は
BirthYearRecall	ja	J	3	松井秀喜の生まれ年は

Table 5: Prompts used for each task. In the Option column, GtoJ and JtoG denote Gregorian to Japanese calendar and Japanese to Gregorian CalendarConversion, respectively. +10yr and -10yr indicate addition and subtraction of ten years in JapaneseCalendarArithmetic. G and J represent prompts requiring output in the Gregorian and Japanese calendars, respectively. Few-shot examples were added before each prompt to guide the model to respond in the correct format.

Model name	Lang	paper	Repo name on Huggingface
Llama-2-7b	en	Touvron et al. (2023)	meta-llama/Llama-2-7b
Llama-2-13b	en	Touvron et al. (2023)	meta-llama/Llama-2-13b
Mistral-7B	en	Jiang et al. (2023)	mistralai/Mistral-7B-v0.1
Llama3.1-8B	en	Grattafiori et al. (2024)	meta-llama/Llama-3.1-8B
llm-jp-3-13b	ja	LLM-jp (2024)	llm-jp/llm-jp-3-13b
sarashina2-13b	ja	-	sbintuitions/sarashina2-13b
Swallow-13b	ja	Fujii et al. (2024); Okazaki et al. (2024)	tokyotech-llm/Swallow-13b-hf
Swallow-MS-7b	ja	-	tokyotech-llm/Swallow-MS-7b-v0.1
Llama3-Swallow-8B	ja	-	tokyotech-llm/Llama-3-Swallow-8B-v0.1

Table 6: List of Japanese-centric and English-centric models used in the experiments.

D CALENDARCONVERSION

Fig. 8 shows the results of CALENDARCONVERSION: from Gregorian to *wareki* (left) and from *wareki* to Gregorian calendar (right). While Japanese-centric models, GPT-4o, and DeepSeek V3 perform near-perfect conversions in both directions, English-centric models exhibit greater variance across models and eras, generally showing inferior performance. Nevertheless, some English models, such as Llama3.1-8B, demonstrate high accuracy in simple conversions. In certain cases, such as Mistral-7B for the Taishō era, models succeed in only one conversion direction, highlighting asymmetries in their learned representations.

E JAPANESECALENDARARITHMETIC

To show the details of the results in JAPANESECALENDARARITHMETIC, we introduce three metrics: Era Match, Near Match, and Full Match.

Era Match captures a coarse understanding of Japanese calendar eras and is defined as the ratio of outputs containing the correct era: $\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}(\hat{E}_i = E_i)$. In the above example, "September 19, Heisei 37" would be incorrect, as the Heisei era ended before reaching year 37, while "September 20, Reiwa 6" is a correct era match.

Near Match accounts for the difficulty of conversions involving transition years and is defined as the ratio of predictions that are off by at most one year: $\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}(|\mathrm{G}(E_i, x_i) - \mathrm{G}(\hat{E}_i, \hat{x}_i)| \leq 1)$, where $\mathrm{G}(E, x)$ converts a *wareki* era and year to its Gregorian equivalent.

Full Match jointly measures knowledge of era transitions and year arithmetic. It is defined as the ratio of outputs that exactly match the correct date: $\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}(E_i = \hat{E}_i \wedge x_i = \hat{x}_i)$. This metric is the same as the accuracy used in § 2.2.

The results are shown in Fig. 9. The top half shows the results for adding ten years, and the bottom half shows the results for subtracting ten years, across Japanese era boundaries.

From the Era Match accuracy, we can see that most models seem to understand the basic order of the eras, with a few exceptions. However, as discussed in § 2.2, the Full Match accuracy shows that even Japanese-centric models and frontier models consistently fail to reason correctly across transitions between older eras, such as Meiji to Taisho or Taisho to Showa. In contrast, for more recent transitions, such as Heisei to Reiwa, Japanese-centric models perform much better than English-centric models.

Japanese models and frontier models usually get higher Full Match accuracies for newer eras, but English models still show low accuracy even in those cases. Many models, especially English ones, show a big gap between their Near Match accuracy and Full Match accuracy. This means they often give answers that are just one year off and cannot handle era transitions exactly. One main reason for these mistakes is that the models do not take into account that era transitions often happen in the same Gregorian year (e.g., Heisei 31 and Reiwa 1 both correspond to 2019).

To sum up, most Japanese-centric models seem to understand the timeline of recent eras well enough to reason correctly across era boundaries, whereas English-centric models still struggle with this.

F BIRTHYEARRECALL

To show the details of the results in BIRTHYEARRECALL, we introduce two metrics: Full match and Within ± 3 years Match. **Full Match** requires an exact match of the prediction and the target. For Gregorian output, accuracy is $\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}(\hat{y}_i = y_i)$. For *wareki* output, the prediction must match both the era and the year within the era: $\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}(E_i = \hat{E}_i \wedge x_i = \hat{x}_i)$. This metric is the same as the accuracy used in § 2.3

Within $\pm$3 years Match allows a deviation of ± 3 years. For *wareki*, this is defined as: $\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}(E_i = \hat{E}_i \wedge |x_i - \hat{x}_i| \leq 3)$ meaning that the prediction must be in the same era and within a 3-year range. In the Gregorian setting, we convert the predicted and target year into *wareki* values (E, x) and then apply the tolerance match condition.

Fig. 10 shows the results of the two metrics in BIRTHYEARRECALL. The x-axis indicates accuracy in the Gregorian calendar, and the y-axis indicates accuracy in *wareki*.

English-centric models perform poorly in recalling birth years in *wareki*. Japanese-centric models and frontier LMs show moderate success when prompted in Japanese, but still underperform compared to their accuracy in Gregorian date recall. Although Japanese-centric models are trained on Japanese corpora, they mainly store birth years in the Gregorian format.

For all Japanese LMs, the Within ±3-years match accuracy for *wareki* recall is more than three times higher than the Full Match accuracy. This suggests that models may either retrieve era-based years with some inaccuracy or rely on internal Gregorian-to-Japanese conversions that result in small shifts.

G CrossCalendarConsistency

Settings. This task evaluates the consistency of BirthYearRecall across calendars. Specifically, it measures the ratio of individuals for whom the model correctly predicts the birth year both in *wareki* and in the Gregorian calendar. We define **Full Match Consistency** as: $\frac{1}{M}\sum_{i=1}^{M}\mathbb{1}(\hat{y}_i = y_i)$, where M is the number of individuals for whom the model's *wareki* prediction is exactly correct (i.e., $\hat{E}_i = E_i$ and $\hat{x}_i = x_i$).

We also report consistency under **Within ±3-years Match**, which allows for a 3-year deviation in the Gregorian prediction. It is defined as: $\frac{1}{M}\sum_{i=1}^{M}\mathbb{1}(|\hat{y}_i - y_i| \leq 3)$.

Results. Fig. 11 shows the consistency of Japanese-centric models in BirthYearRecall, measuring the ratio of cases where the model correctly answers both in *wareki* and Gregorian format. Results are reported using both exact match and 3-year tolerance criteria. English-centric models are omitted because they rarely produced correct *wareki* output in this task.

While some Japanese models, such as sarashina2-13b and Swallow-13b, achieve over 80% consistency, others like Swallow-MS-7b show lower consistency around 50%. These results show that even in Japanese-centric LMs, there is considerable variation in how knowledge relating to Gregorian and Japanese calendars is recalled.

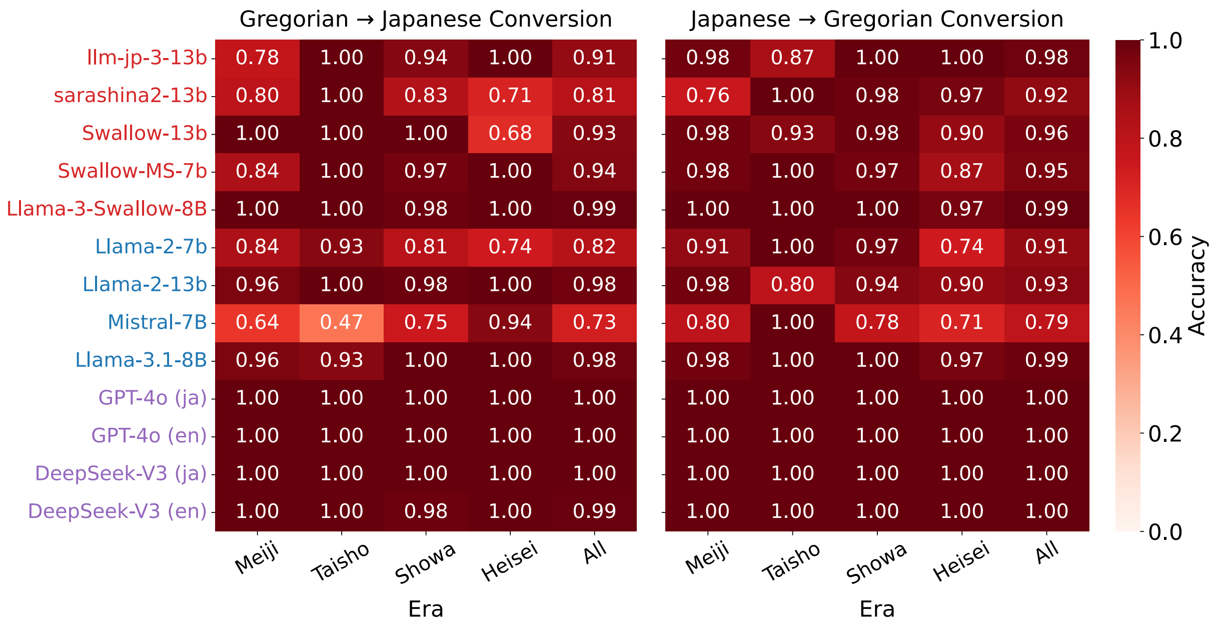

Figure 8: Accuracy of Gregorian-to-Japanese (left) and Japanese-to-Gregorian (right) CALENDARCONVERSION. Japanese-centric models and frontier models achieved near-perfect accuracy in both directions, while English-centric models showed notable variation across models and eras.

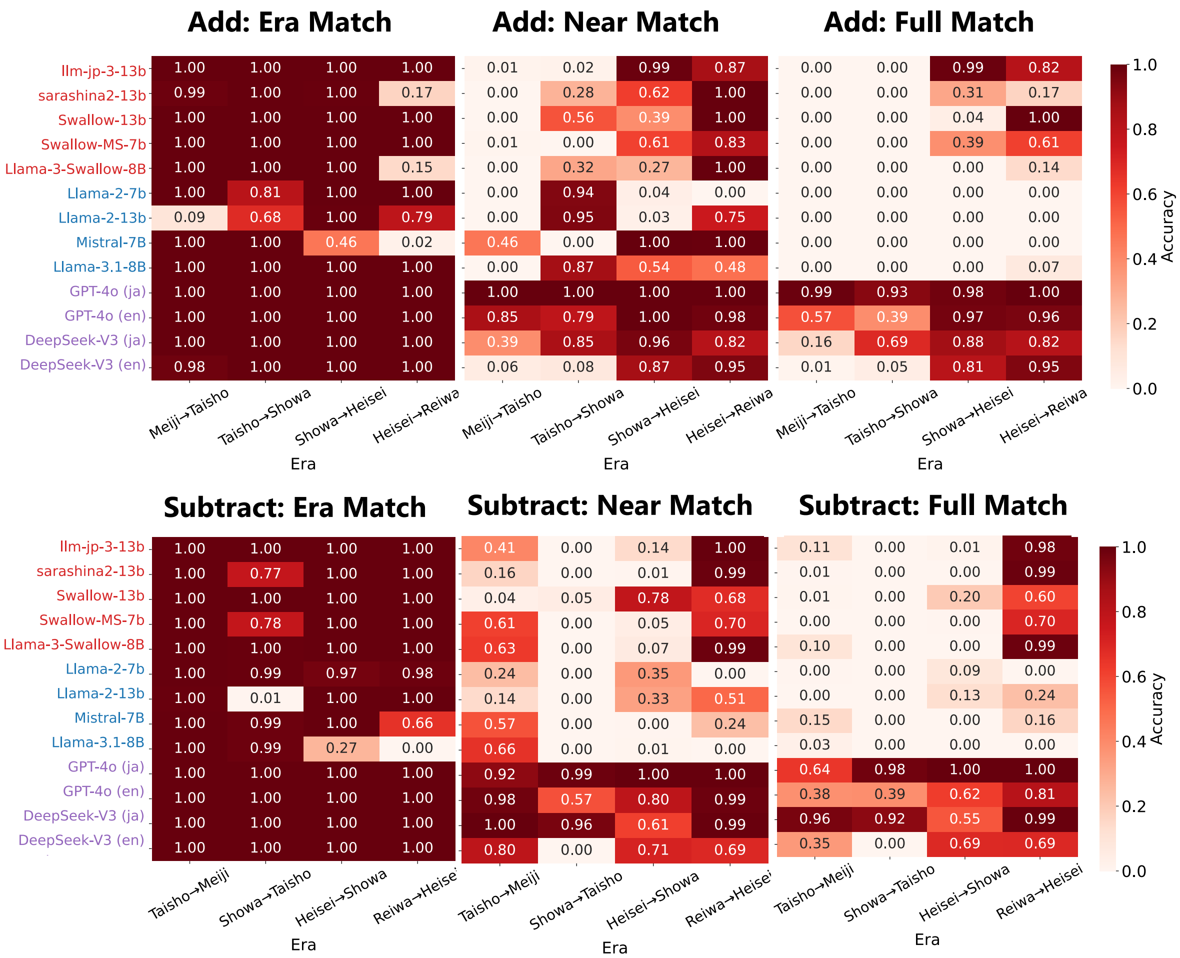

Figure 9: Evaluation results for three metrics(Era match Accuracy, Near match Accuracy, and Full match Accuracy) in JAPANESECALENDARARITHMETIC: 10-year addition (top) and 10-year subtraction (bottom). Both Japanese-centric and English-centric models achieve high Era match accuracies for most transitions, indicating that they generally understand the chronological order of era. However, for the Full match accuracy, only Japanese models show high performance on recent transitions, suggesting that they capture the timeline discontinuity at era boundaries. English models, by contrast, fail to do so, likely because they do not recognize that era transitions (e.g., Heisei 31 to Reiwa 1) can occur within the same Gregorian year.

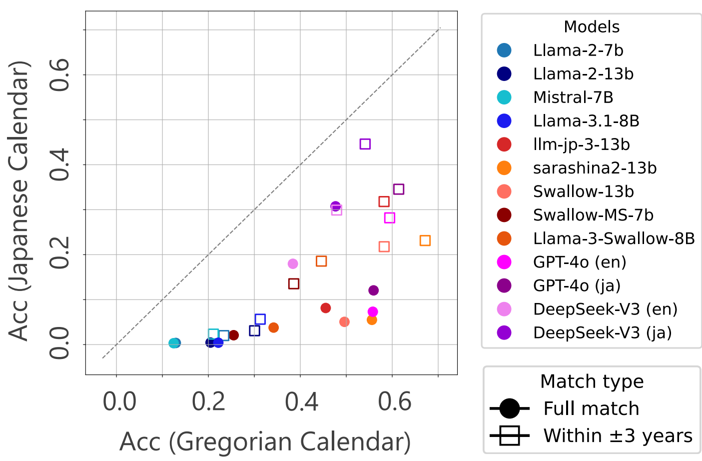

Figure 10: Results of BIRTHYEARRECALL using both Japanese and English prompts. The accuracy of both Japanese-centric and English-centric models varies significantly depending on whether the person's name is presented in Japanese or English. For Japanese calendar outputs, English models fail to produce correct answers regardless of the name format, while Japanese models only succeed when the name is presented in Japanese.

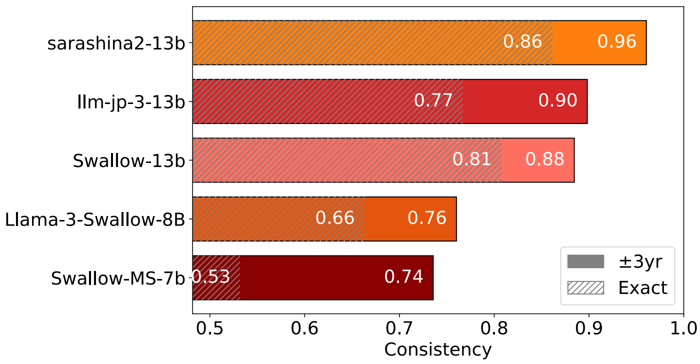

Figure 11: CROSSCALENDARCONSISTENCY: proportion of items correctly answered in the Japanese calendar that are also correctly answered in the Gregorian calendar under exact match and within a ±3-year tolerance. For exact matches, some Japanese-centric models, such as sarashina2-13b and Swallow-13b, exhibit high consistency above 80%, while others, like Swallow-MS-7b, remain around 50%.