Evaluating Language Models in Location Referring Expression Extraction from Early Modern and Contemporary Japanese Texts

Ayuki Katayama[♣] **Shohei Higashiyama**[♠,♣,◇] **Hiroki Ouchi**[♣,‡,◇] **Yusuke Sakai**[♣]
Ayano Takeuchi[♡] **Ryo Bando**[◇] **Yuta Hashimoto**[†] **Toshinobu Ogiso**[◇] **Taro Watanabe**[♣]
[♣] NAIST [♠] NICT [◇] NINJAL [♡] NIJL [†] National Museum of Ethnology [‡] RIKEN
{katayama.ayuki.kc1,hiroki.ouchi,sakai.yusuke.sr9,taro}@is.naist.jp,
shohei.higashiyama@nict.go.jp, takeuchi.ayano@nijl.ac.jp,
yhashimoto@rekihaku.ac.jp, {r-bando,togiso}@ninjal.ac.jp

Abstract

Automatic extraction of geographic information, including Location Referring Expressions (LREs), can aid humanities research in analyzing large collections of historical texts. In this study, we investigated how accurate pretrained Transformer language models (LMs) can extract LREs from historical texts. In particular, we evaluated two representative types of LMs, namely, masked language model and causal language model, using early modern and contemporary Japanese datasets. Our experimental results demonstrated the potential of contemporary LMs for historical texts, but also suggest the need for further model enhancement, such as pretraining on historical texts.

1 Introduction

Historical texts are crucial for a better understanding human and natural history because they record various events and activities of their time. From a *geographic* perspective, historical texts often include Location Referring Expressions (LREs), such as historical place and facility names, along with objects and events related to those locations. As representative examples of such texts, travelogues describe the experiences of the writer in the places they visited, and disaster records describe the affected regions, the scale of the damage, and peoples' situation. Automatic extraction and structurization of such geography-related information by computers can support humanities scholars in analyzing large collections of historical texts.

As a fundamental step for computer-aided geographic text analysis, this study addresses LRE extraction from historical Japanese texts. For an example sentence "名取川渡りて仙台に入る,"[1] an LRE system is required to extract two LREs, "名取川 (Natorigawa)" and "仙台 (Sendai)." Specifically, we investigates the LRE accuracy of Transformer (Vaswani et al., 2017) language models (LMs), which have achieved remarkable success in various natural language processing tasks (Devlin et al., 2019; Brown et al., 2020; Raffel et al., 2020). We focus on two representative types of LMs: Masked Language Model (MLM) and Causal Language Model (CLM).

For model evaluation experiments, we use three datasets: an early modern Japanese travelogue to which we added LRE annotations, Oku no Hosomichi (HOSOMICHI),[2] early modern Japanese disaster records, the Minna de Honkoku dataset (Hashimoto, 2023) (MINNA), and contemporary Japanese travelogues, the Arukikata Travelogue Dataset (Arukikata. Co., Ltd., 2022; Ouchi et al., 2023) (ARUKIKATA). The reasons for using contemporary texts alongside historical texts are twofold: (i) comparing model performance across texts from different eras, and (ii) investigating whether contemporary texts can enhance model performance on historical texts.

Our experiments demonstrated the following results:

- In all settings, an MLM with 3.4M parameters, BERT (Devlin et al., 2019), consistently outperformed a CLM with 7B parameters, Swallow (Fujii et al., 2024).
- The LMs that had been pretrained with contemporary texts achieved high accuracy on the contemporary dataset (F1 scores of up to 0.856 on ARUKIKATA), but yielded low to moderate accuracy on the historical datasets (up to 0.425 on HOSOMICHI and 0.687 on MINNA).
- Models fine-tuned with both contemporary and historical labeled texts achieved the best accuracy for the two historical datasets.

[1] The English translation is 'Crossed the Natori River and entered Sendai.'

[2] For reproducing our results, we will publish our HOSOMICHI annotation dataset at https://github.com/naist-nlp/historical-travelogues.

Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities, pages 331–338
November 16, 2024.

2 Background and Related Work

LRE extraction, also known as geotagging or toponym recognition, is a special case of named entity recognition (NER) (Nadeau and Sekine, 2007). LRE extraction has often been addressed within the task of geoparsing (Gritta et al., 2020), which aims to estimate the geographic coordinates or geographic database entries that correspond to the locations referenced by LREs.

Resources Previous studies have constructed location-annotated historical corpora and evaluated the performance of machine learning systems in LRE extraction using, for example, English news articles (Coll Ardanuy et al., 2022), English travel writings (Rayson et al., 2017; Sprugnoli et al., 2018), French literary texts (Kogkitsidou and Gambette, 2020), and Chinese historical books (Tang et al., 2024). For Japanese, some researchers have attempted to manually annotate LREs and their geographical coordinates in texts within historical disaster record databases, such as the Database of Materials for the History of Japanese Earthquakes (Kano and Ohmua, 2023) and Minna de Honkoku (Hashimoto, 2023).

System Evaluation Many studies have investigated various methods for recognizing named entities, including locations, in historical texts (Ehrmann et al., 2023). In particular, some recent studies have focused on pretrained Transformer LMs. Labusch et al. (2019) investigated training strategies for BERT suitable for NER on historical German newspaper texts. They showed that a contemporary BERT model achieved the highest accuracy when both pretraining on large unlabeled historical texts and labeled contemporary texts were performed prior to fine-tuning on target labeled historical texts. Tang et al. (2024) evaluated NER accuracy on ancient Chinese historical documents using MLMs pretrained on historical texts, and open and closed contemporary CLMs, with MLMs achieving higher accuracy.

3 Experiments

3.1 Training Scenarios

The purpose of our experiments in this study is to investigate how LMs pretrained with large contemporary texts can be adapted to historical texts. Thus, we employed the following training/evaluation scenarios with three datasets, explained later: (1) fine-tuning on contemporary texts only, (2) fine-tuning on historical texts only, (3) fine-tuning on both contemporary and historical texts, and then: (a) evaluating on contemporary texts or (b) evaluating on historical texts. Through these scenarios, we compare the accuracy of an MLM and a CLM.

Dataset	Register	#Sentences	#LREs
Arukikata-Train	Travelogue	6,516	3,102
Arukikata-Dev	Travelogue	601	260
Arukikata-Test	Travelogue	5,156	2,166
Minna-Train	Disaster	1,901	9,690
Minna-Test	Disaster	476	2,392
Hosomichi	Travelogue	523	242

Table 1: Dataset categories and statistics.

3.2 Datasets

We curated three datasets in Table 1: a contemporary text dataset (ARUKIKATA) and two early modern text datasets (MINNA and HOSOMICHI).

ARUKIKATA As contemporary Japanese texts, we used the ATD-MCL (Higashiyama et al., 2024), a dataset of travelogues with manually annotated LREs. We treated only `LOC-NAME` (location name) and `FAC-NAME` (facility name) mentions as LREs with `LOCATION` type, and ignored the other mentions. We followed the official train/dev/test split.

MINNA As one of the two historical text datasets, we used the annotation dataset[3] from the Minna de Honkoku database (Hashimoto, 2023).[4] The database comprises records of early modern Japanese disasters from around the 1800s, with manually annotated expressions with "date," "location," "damage," and "person" types. As preprocessing, we divided the single entire document into 50-character segments, and treated each segment as a sentence. Then, we extracted sentences with one or more LREs (i.e., "location" type expressions) and split these sentences into training and test sets at a ratio of 8:2.[5] Texts in this dataset are typically written in a style where locations and damages at the locations are enumerated, for example, "小石川御門内よりするが臺 小川丁 筋違御門迄少〻破損." [6]

[3] https://github.com/yuta1984/honkoku-data

[4] https://honkoku.org/index_en.html

[5] We treated LREs across segment boundaries as non-LREs.

[6] LREs are written with underlines. English translation is 'From inside the Koishikawa Gate to Suruga-dai, Ogawa-cho, and Sujikai Gate—some damage.'

HOSOMICHI As another historical text, we newly created an annotation dataset using the Oku no Hosomichi Wikisource text[7] (the dataset will be published as mentioned in §1). Oku no Hosomichi is one of the most famous and representative historical Japanese travelogues written by Matsuo Basho in the 1700s. We selected this source because it is a literary text focused on geographic human movement, unlike MINNA, which consists of practical records of geographic events; thus, the two datasets were written in similar periods of time but have different characteristics. Two of the authors manually assigned `LOCATION` type to the LRE spans within the text. Note that we use this dataset only for evaluation as unseen-domain early modern text because of its limited data size. Texts in Oku no Hosomichi and their English translations can be viewed, for example, on Wikipedia.[8]

3.3 Language Models

We evaluated two types of LMs, MLM and CLM, both of which were pretrained with large contemporary Japanese texts.

MLM We used a character-level Japanese pretrained model[9] of BERT (Devlin et al., 2019) by fine-tuning it with an additional fully-connected layer for label classification, following the settings in Appendix A. The model is trained to assign one of three labels (`B-LOCATION`, `I-LOCATION`, and `O`) to each character token with the softmax cross-entropy loss.

CLM We used Swallow-7b-hf (Fujii et al., 2024),[10] which has undergone continual pretraining from a Llama 2 (Touvron et al., 2023) model with Japanese language data. We fine-tuned the Swallow model with QLoRA (Dettmers et al., 2023) following the settings in Appendix A. Since CLM generates text in an autoregressive manner, we applied the prompt template shown in Figure 1 to each input sentence and fed the filled prompt into the model. The model is then trained to generate the text following "Answer:". We adopted this simple prompt based on the previous study (Kito et al., 2024), which demonstrated the minimal effect of prompt differences in NER when fine-tuning LLMs.

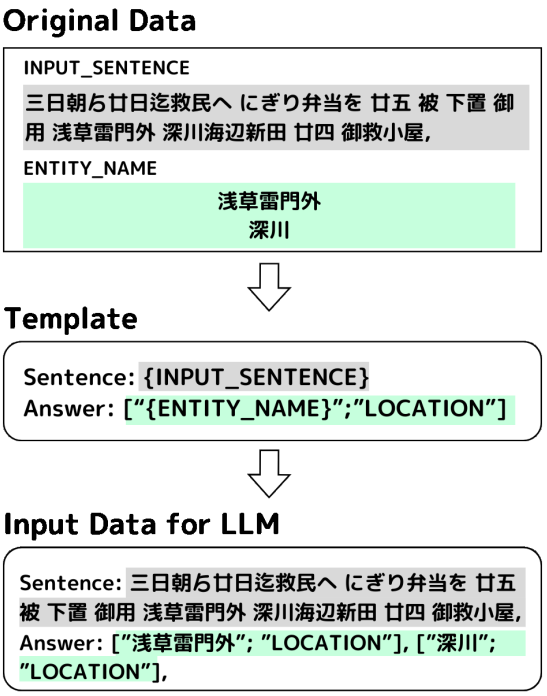

Figure 1: Example of input and output text for Swallow.

4 Results and Discussion

Table 2 shows the mean F1 scores of three model runs with different random seeds for each training setting and each evaluation dataset. For each run, the model checkpoint with the best F1 score on the development data was selected.[11] We will focus on important aspects in the following sections, and additional discussion is provided in Appendix B.

4.1 On Contemporary Travelogues

On the ARUKIKATA evaluation data, both LMs trained on the ARUKIKATA training data achieved the best F1 scores (0.856 by BERT and 0.797 by Swallow). The models trained on the MINNA training data showed poor accuracy (0.269 and 0.162), and the models trained on the mixed training data did not show any improvement over those trained only on the ARUKIKATA training data. The main reason of these results is the large discrepancy in characteristics between the two datasets; there are differences not only in the

[7] https://ja.wikisource.org/wiki/%E3%81%8A%E3%81%8F%E3%81%AE%E3%81%BB%E3%81%9D%E9%81%93

[8] https://en.wikipedia.org/wiki/Oku_no_Hosomichi

[9] https://huggingface.co/tohoku-nlp/bert-large-japanese-char-v2

[10] https://huggingface.co/tokyotech-llm/Swallow-7b-hf

[11] For the experiments using the MINNA data, we used a random 5% of the training sentences as the development data for training Swallow and the entire training data as the development data for training BERT.

Model	Training data	Evaluation data		
		ARUKIKATA	MINNA	HOSOMICHI
BERT-Large	ARUKIKATA	**0.856**	0.224	0.345
	MINNA	0.269	0.657	0.361
	ARUKIKATA+ MINNA	0.832	**0.687**	**0.425**
Swallow-7b-hf	ARUKIKATA	**0.797**	0.029	0.244
	MINNA	0.162	0.174	0.257
	ARUKIKATA+ MINNA	0.753	**0.267**	**0.411**

Table 2: F1 scores of two LMs on each eavaluation data.

era of the texts but also in writing style due to the text register (ARUKIKATA comprising travelogues and MINNA comprising disaster records).

4.2 On Early Modern Disaster Records

On the MINNA evaluation data, we observed the following three findings.

First, both LMs trained on the mixed training data achieved the best F1 scores. These results are somewhat surpising: adding MINNA training data was not effective for evaluation on ARUKIKATA (§4.1), but adding ARUKIKATA training data was effective for evaluation on MINNA. A possible reason is that knowledge of a wide variety of place names may have been useful for MINNA evaluation data; whereas ARUKIKATA data includes a variety of place names from across Japan, MINNA data is biased towards locations around Edo (present-day Tokyo).

Second, the absolute F1 scores for MINNA evaluation data were overall lower than those for ARUKIKATA evaluation data. This would be because the LMs were pretrained on contemporary Japanese texts. LMs pretrained on historical texts can improve the downstream task performance, as demonstrated by Labusch et al. (2019), which is an interesting future direction.

Third, the performance of Swallow (up to 0.267) was substantially lower than that of BERT (up to 0.687). A possible reason is the difference between the training methods, full-parameter tuning for BERT and QLoRA tuning for Swallow. During the QLoRA tuning that we used for Swallow, only a small number of newly added parameters were updated, and the original parameters were fixed. Thus, the model may not be able to fit the training data sufficiently. However, additional evaluation is needed to verify this: tuning BERT with QLoRA. Other possible reasons are the differences in the approach to the extraction task, namely, classification by BERT and language generation by Swallow, as well as differences in pretraining tasks, namely, MLM and CLM. These could impact the differences in the knowledge acquired during pretraining, as well as the number of examples necessary for downstream task training.

4.3 On Early Modern Travelogue

On the HOSOMICHI evaluation data, both LM achieved close F1 scores when trained on the ARUKIKATA training data and when trained on MINNA training data. Moreover, both LMs trained on the mixed training data achieved the best F1 scores. These results indicate that the two training data were both effective and complementary in the extraction of LREs from the HOSOMICHI data. Probable reasons are as follows. Although the text registers of MINNA and HOSOMICHI are different (disaster records vs a travelogue), the time period of both is relatively close. Altough the time period of ARUKIKATA and HOSOMICHI are different (contemporary vs early modern), both travelogue data may be similar in that they include wide range of place names in Japan and describe the writer's experiences at each location.

Because of the cross-domain scenario, the absolute F1 scores on the HOSOMICHI evaluation data is not high: up to approximately 0.4. Straightfoward apporaches to imporve extraction accuracy for this dataset include pretraining on similar domain texts and fine-tuning with similar domain labeled examples.

4.4 Qualitative Analysis

Table 3 shows LRE examples predicted by the LMs trained on the mixed training data for HOSOMICHI dataset. Although "室の八島 (Muro no Yashima)" is a single LRE, both LMs only rec-

Gold	室の八島 (Muro no Yashima)
BERT	八島 (Yashima)
Swallow	八島 (Yashima)

Table 3: Example LREs predicted by LMs fine-tuned with ARUKIKATA+MINNA data for the sentence "室の八島に詣す (Visiting Muro no Yashima)" in the HOSOMICHI evaluation data.

ognized "八島 (Yashima)" as an LRE. One possible reason is that "の (no)" was misinterpreted as Japanese particle that indicates possession or belonging, leading the LMs to understand it as "Yashima of Muro" or "Muro's Yashima," although the entire span is a single phrase. This failure suggests that the LMs lack knowledge about historical place names.

5 Conclusion

This study investigated the extraction accuracy of representative pretrained Japanese LMs using early modern and contemporary LRE datasets. One of main findings from our experiments is the effectiveness of fine-tuning with both contemporary and historical labeled texts. Possible future work includes (i) expanding the evaluation to cover a broader range of eras and registers, and (ii) investigating pretraining strategies using unlabeled historical texts effective for downstream tasks, including LRE extraction and others.

Limitations

In this study, we selected one representative Japanese LM for both MLM and CLM. It is unclear whether similar trends would be observed with other Japanese LMs. Therefore, it is desirable to evaluate a more diverse LMs for a comprehensive analysis in the future. However, considering that most current Japanese LMs are based on the Transformer architecture, we believe that the choice of models is appropriate as a first step in identifying the potential challenges that Japanese LMs may face in extracting LREs from historical Japanese texts.

Ethics Statement

The evaluation datasets present no licensing issues, as ARUKIKATA is under the MIT License, MINNA is under the CC-BY-SA 4.0 License, and HOSOMICHI is sourced from Wikisource under the same CC-BY-SA 4.0 License. Furthermore, since the original text of "Oku no Hosomichi" is in the public domain, there are no copyright issues related to its distribution. Additionally, the annotation data only involves tagging the original text, which means it does not contain any harmful content in our artifacts.

Acknowledgements

We would like to thank the anonymous reviewers and meta reviewers for their constructive comments. This study was supported by JSPS KAKENHI Grant Number JP23K24904 and NIJAL Joint Resource-use Projects (B) "Geoparsing for Historical Japanese Text."

References

Arukikata. Co., Ltd. 2022. Arukikata travelogue dataset. Informatics Research Data Repository, National Institute of Informatics. https://doi.org/10.32130/idr.18.1.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Mariona Coll Ardanuy, David Beavan, Kaspar Beelen, Kasra Hosseini, Jon Lawrence, Katherine McDonough, Federico Nanni, Daniel van Strien, and Daniel C. S. Wilson. 2022. A dataset for toponym resolution in nineteenth-century english newspapers. *Journal of Open Humanities Data*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2023. Named entity recognition and classification in historical documents: A survey. *ACM Comput. Surv.*, 56(2).

Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. 2024. Continual pre-training for cross-lingual LLM adaptation: Enhancing Japanese language capabilities. In *Proceedings of the First Conference on Language Modeling*, COLM, page (to appear), University of Pennsylvania, USA.

Milan Gritta, Mohammad Taher Pilehvar, and Nigel Collier. 2020. A pragmatic guide to geoparsing evaluation: Toponyms, named entity recognition and pragmatics. *Language Resources and Evaluation*, 54:683–712.

Yuta Hashimoto. 2023. Prototype development of a markup system for historical disaster records. *IPSJ SIG Computers and the Humanities Technical Report*, 2023-CH-131(2):1–6.

Shohei Higashiyama, Hiroki Ouchi, Hiroki Teranishi, Hiroyuki Otomo, Yusuke Ide, Aitaro Yamamoto, Hiroyuki Shindo, Yuki Matsuda, Shoko Wakamiya, Naoya Inoue, Ikuya Yamada, and Taro Watanabe. 2024. Arukikata travelogue dataset with geographic entity mention, coreference, and link annotation. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 513–532, St. Julian's, Malta. Association for Computational Linguistics.

Yasuyuki Kano and Junzo Ohmua. 2023. Integration of geographical information into the data of materials for the history of Japanese earthquakes. *IPSJ SIG Computers and the Humanities Technical Report*, 2023-CH-131(3):1–3.

Taisei Kito, Kohei Makino, Makoto Miwa, and Yutaka Sasaki. 2024. *Koyū hyōgen chūshutsu ni-okeru daikibo gengo* model *no* LoRA fine-tuning *no gakushū settei no chōsa* (Investigating LoRA fine-tuning training settings of large language models in named entity recognition). In *Proceedings of the 30th Annual Conference of the Association for Natural Language Processing*.

Eleni Kogkitsidou and Philippe Gambette. 2020. Normalisation of 16th and 17th century texts in French and geographical named entity recognition. In *Proceedings of the 4th ACM SIGSPATIAL Workshop on Geospatial Humanities*, GeoHumanities '20, page 28–34, New York, NY, USA. Association for Computing Machinery.

Kai Labusch, Preußischer Kulturbesitz, Clemens Neudecker, and David Zellhöfer. 2019. BERT for named entity recognition in contemporary and historical German. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)*, pages 9–11.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Hiroki Ouchi, Hiroyuki Shindo, Shoko Wakamiya, Yuki Matsuda, Naoya Inoue, Shohei Higashiyama, Satoshi Nakamura, and Taro Watanabe. 2023. Arukikata travelogue dataset. arXiv:2305.11444.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(1).

Paul Rayson, Alex Reinhold, James Butler, Chris Donaldson, Ian Gregory, and Joanna Taylor. 2017. A deeply annotated testbed for geographical text analysis: The corpus of lake district writing. In *Proceedings of the 1st ACM SIGSPATIAL Workshop on Geospatial Humanities*, GeoHumanities'17, page 9–15, New York, NY, USA. Association for Computing Machinery.

Rachele Sprugnoli et al. 2018. Arretium or arezzo? A neural approach to the identification of place names in historical texts. In *Proceedings of the Fifth Italian Conference on Computational Linguistics*. CEUR-WS.

Xuemei Tang, Qi Su, Jun Wang, and Zekun Deng. 2024. CHisIEC: An information extraction corpus for Ancient Chinese history. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3192–3202, Torino, Italia. ELRA and ICCL.

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, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. arXiv:2307.09288.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz

Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

A Model Hyperparameters

Table 4 and Table 5 show the hyper-parameters used for BERT-Large and Swallow-7b-hf, respectively.

Hyper-parameter	Value
training epochs	20
batch size	32
learning rate	1e-5
lr scheduler type	linear
warmup ratio	0.1
gradient norm clipping threshold	1.0
optimizer	AdamW

Table 4: The hyper-parameters used for BERT-Large.

Hyper-parameter	Value
training epochs	10
batch size	8
learning rate	5e-5
lr scheduler type	linear
optimizer	paged_adamw_8bit
quant_method	BITS_AND_BYTES
load_in_4bit	True
bnb_4bit_use_double_quant	True
bnb_4bit_quant_type	nf4
bnb_4bit_compute_dtype	float16
lora_alpha	16
lora_dropout	0.1
bottleneck_r	64
torch_dtype	float16

Table 5: The hyper-parameters used for Swallow-7b-hf.

B Additional Experimental Results

Detailed Results for the Main Experiment Table 6 show precision and recall as well as F1 scores of the two LMs in the main experiment, which is shown in Table 2 in §4. For simplicity, the results of the models that achieved the best accuracy among training data settings are shown for each evaluation dataset. We observed that Swallow achieved moderate to high precision (0.533-0.895) for each evaluation dataset, which is not significantly lower than that of BERT (0.367–0.841) and is even higher in two out of three datasets. However, Swallow yielded consistently lower recall than BERT, particularly showing very low recall (0.172) for MINNA. This indicates that Swallow made conservative predictions and that improvements in learning methods or prompts are necessary to enhance coverage.

Effects of Instruction Language in CLM Prompt We conducted an additional experiment using another prompt for the CLM after the review, based on a reviewer's comment that suggested to use (contemporary and early modern) Japanese prompts. Specifically, we used a Japanese prompt, which replaces "Sentence:" and "Answer:" in the original English prompt with "入力文:" and "回答:", respectively.[12] As show in Table 7, compared to Swallow fine-tuned with the English prompt, the model fine-tuned with the Japanese prompt yielded significantly lower F1 scores in four out of nine settings, which are underlined in Table 7, while it achieved similar F1 scores in the other settings. A possible reason for this degradation is that the backbone model, Swallow-7b-hf, has not been instruction-tuned in the Japanese language. This result suggests the necessity of evaluating more diverse LMs, including instruction-tuned models, as well as investigating LMs' sensitivity to different prompts.

Model	Train	Eval	P	R	F1
BERT	A	A	0.841	0.872	0.856
	A+M	M	0.662	0.714	0.687
	A+M	H	0.367	0.506	0.425
Swallow	A	A	0.895	0.717	0.797
	A+M	M	0.594	0.172	0.267
	A+M	H	0.533	0.335	0.411

Table 6: Precision (P), Recall (R), and F1 scores of two LMs. "A," "M," and "H" represent ARUKIKATA, MINNA, and HOSOMICHI, respectively.

Prompt	Training data	Evaluation data A	M	H
En	A	0.797	0.029	0.244
	M	0.162	0.174	0.257
	A+M	0.753	0.267	0.411
Ja	A	0.304	0.032	0.095
	M	0.164	0.028	0.258
	A+M	0.564	0.272	0.418

Table 7: F1 scores of Swallow with English (En) and Japanese (Ja) prompts. "A," "M," and "H" represent ARUKIKATA, MINNA, and HOSOMICHI, respectively.

[12] Because the instruction text of our prompt template is concise, and the model was pretrained in English and contemporary Japanese, we only conducted an additional experiment with the contemporary Japanese prompt. We will investigate the effects of more sophisticated prompts, including those based on early modern Japanese, in the future.