

LLM for Japanese Text OCR: Automating Kuzushiji Recognition Using Decoder-Only OCR Architecture[†]

Damir Kim ^{1,*} , Natalia Bogach ¹ and Evgeny Pyshkin ² 

¹ Institute of Computer Science and Technology, Peter the Great St. Petersburg Polytechnic University, St. Petersburg 195251, Russia; bogach_nv@spbstu.ru

² School of Computer Science and Engineering, The University of Aizu, Aizu-Wakamatsu 965-8580, Japan; pyshe@u-aizu.ac.jp

* Correspondence: kanearest@mail.ru

[†] Presented at the 8th International Global Conference Series on ICT Integration in Technical Education & Smart Society, Aizuwakamatsu City, Japan, 20–26 January 2026.

Abstract

This study presents an application of DTrOCR approach to Japanese text optical recognition, particularly addressing the possibility of processing sophisticated writing styles such as kuzushiji or handwritten calligraphy. This approach utilizes the recently developed method which applies decoder-only transformer to the process of optical character recognition (OCR). Specifically, the current research examines the adaption of this method to the task of Japanese character recognition under the constraints of limited training and computational resources. We experimented with training the model both with and without fine-tuning using the corpora with the synthesized printed texts as well as the texts from kuzushiji—traditional Japanese cursive writing. Our findings suggest that a fine-tuned DTrOCR-based approach is promising for kuzushiji automatic recognition and can retain its accuracy on other types of Japanese texts using cursive writing.

Keywords: optical character recognition; large language models; kuzushiji; Japanese

1. Background

OCR (optical character recognition) is one of the constituent domains in computer vision. OCR method application areas include digitizing archived documents and books, automating document workflows in accounting systems, translating texts in photographs and images, as well as integration in language learning applications.

Neural networks and machine learning algorithms have significantly transformed the landscape of text recognition, especially for languages utilizing a large number of complex symbols, varying in form and writing styles, for instance, Japanese and Chinese.

Most of the known traditional OCR methods are based on autoencoder architecture [1–3], some leveraging a pretrained language model as a decoder [4]. LLM-based decoders ensure more accurate predictions, which leverage not only the visual features of particular characters but also the context derived from the previously predicted words and characters. Fujitake suggested the DTrOCR (Decoder-only Transformer for Optical Character Recognition) method [5], where the model employs a pretrained LLM as a decoder. The method has fewer parameters compared to many established traditional methods, while producing better accuracy [5].



Academic Editors: Debopriyo Roy,
George Fragulis and Peter Ilic

Published: 16 July 2026

Copyright: © 2026 by the authors.
Licensee MDPI, Basel, Switzerland.
This article is an open access article
distributed under the terms and
conditions of the [Creative Commons
Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

This paper explores the potential of using DTrOCR for Japanese text recognition, specifically addressing two major use cases:

- Automatic printed text recognition based on a synthesized Japanese text corpus;
- Recognition of kuzushiji texts.

2. Introduction to Kuzushiji

Kuzushiji (literally meaning “deformed characters”) is a highly cursive Japanese writing style used in many Japanese historical documents for over a thousand years prior to 1900 (see for example, [6]). Figure 1 shows an example of an Edo-period illustrated book printed in *ukiyo-e*, the style of Japanese woodblock prints and paintings that emerged and developed mostly during the Edo period (1603–1868) but still has a significant impact on modern and contemporary arts and cultural studies [7,8], where kuzushiji writing can often be seen. The illustrated explanations about the technology of making such books can be found in [9], pp. 26–27.

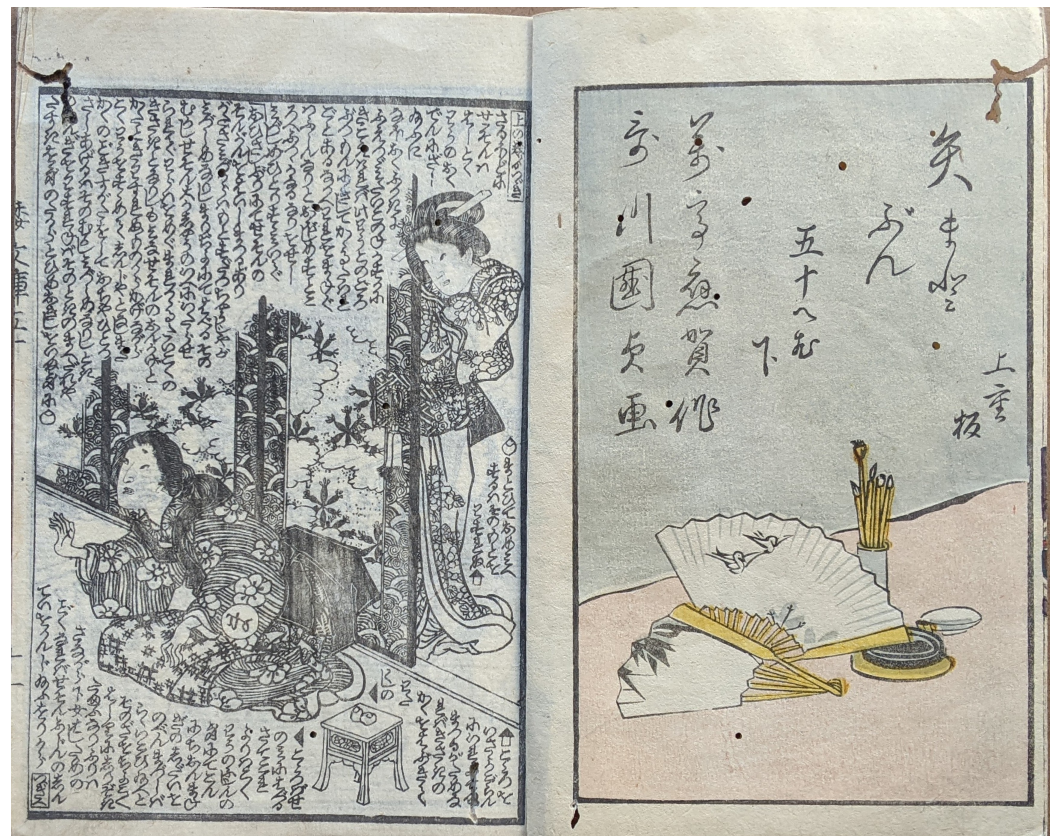


Figure 1. Example of the Edo period illustrated book: ŌGA Mantei (author), UTAGAWA Kunisada (illustrator), Yamato Bunko, vol. 50, gōkan. Late Edo period. Author’s collection.

Different characters can be distorted in kuzushiji by varying degrees. Figure 2 shows the cover and fragment from the typical dictionary illustrating such distortions. Interestingly, more frequently used characters can be even more deformed compared to others [10]. This imposes significant challenges in reading and understanding such documents. Kuzushiji is considered so challenging to learn that students enrolled in programs on Japanese history and philology often take specialized advanced courses to learn how to read kuzushiji [11,12].

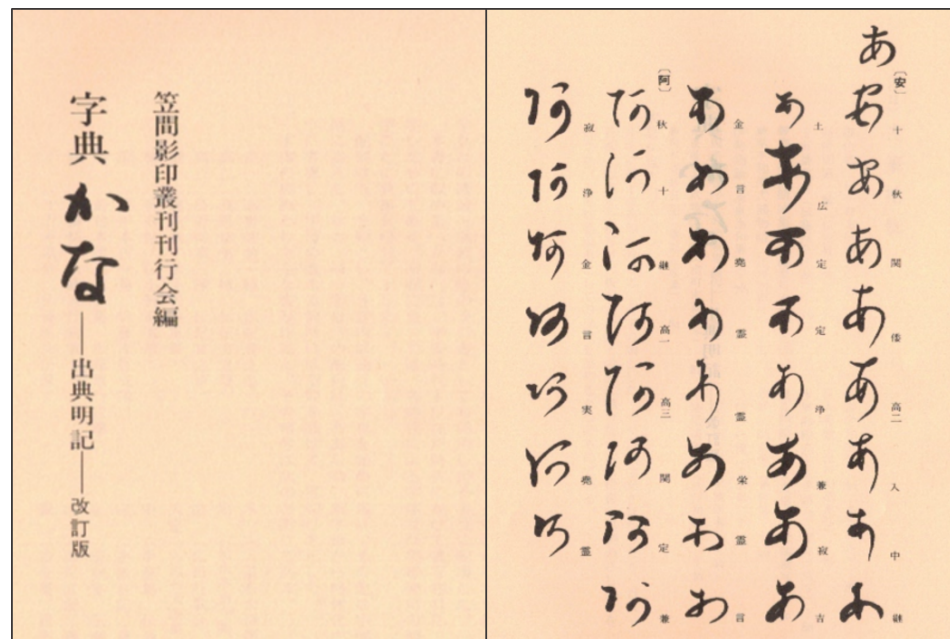


Figure 2. Fragments of typical vocabulary displaying kuzushiji character distortions [13].

3. Related Work

3.1. OCR for Japanese Texts

Numerous recent research works on Japanese OCR pertain to very specific tasks [14–16] varying from the recognition of documents belonging to specific time periods [15] and old cursive writings [14] to constructing text recognition systems for rather unique use cases, such as the recognition of mathematical formulae alongside text [16]. Some of them use transformer-based decoders [15] and even utilize multimodal LLMs [17].

However, only a few existing approaches aim to create a text recognition model that would perform well on a variety of text types. Such models exist for languages like English and Chinese [4,5], but to the best of our knowledge, they have not been applied to Japanese text recognition.

The majority of high-quality data currently available for LLM training is in English and Chinese [18], while the data for common text types such as scene and handwritten text in Japanese is much more limited. That is why Japanese is often considered as one of the examples of so-called “impossible” languages [19]. Specifically for the purposes of OCR, the above-mentioned limitation can be overcome through the use of a methodology that would be highly specific to the task [15] or through the use of synthesized data [17] along with oversampling techniques [14].

3.2. Modern Multimodal OCR Methods

Most modern multimodal OCR approaches mainly feature an encoder–decoder structure. Before the introduction of transformers, most methods used a combination of CNNs and RNNs. The encoder aims to extract visual representations of text images, while the decoder maps the representations to text. In recent years, transformers have been leveraged for OCR tasks [1,4,20–22]. For example, TrOCR [4] uses a combination of a vision transformer model as an encoder and a transformer language model as a decoder. MaskOCR project delivers another example [22], where a pair of transformers are used as components of an encoder–decoder architecture. This project also employs masked modeling to train both the encoder and decoder. Both of the mentioned methods achieved state-of-the-art results at the time of their publication, showcasing the potential of transformer architecture for OCR.

3.3. DTrOCR Original Method

DTrOCR [5] is one of the most recent OCR methods. The key peculiarity of this method is that it does not utilize an encoder, while a pretrained language model serves as a decoder; the latter is additionally trained for the OCR task. According to the reported results, while keeping the training costs at the same level, such a retrained model significantly improves performance compared to training the decoder from scratch.

Particularly, according to the results reported in [5], the DTrOCR method outperforms existing state-of-the-art methods on various benchmarks for a selection of text recognition tasks, including scene, handwritten, and printed text recognition. It performs nearly equally well in recognizing both English and Chinese texts, suggesting the generality of the method.

4. Method

4.1. Original Architecture

As mentioned previously, the pretrained language model is trained only to process a language and possesses the language knowledge; however, that knowledge is not connected to the image features the decoder receives. To teach the model to catch the connections between known tokens and specific image features, additional training is performed.

Before being passed to the language model, the input image undergoes patch embedding. The patch embedding module converts an image into a sequence of small, fixed-size 1D vectors (patches) by dividing, flattening, and linearly projecting them. This enables transformers to process images as sequences of tokens and preserves the local structure while reducing computational complexity. Those sequences are then passed into the decoder with positional embeddings. After receiving a special token [SEP], the decoder predicts the first-word token. Then, it begins to predict subsequent tokens auto-regressively until [EOS]—another special token—is passed to it. The diagram in Figure 3 illustrates the overall workflow.

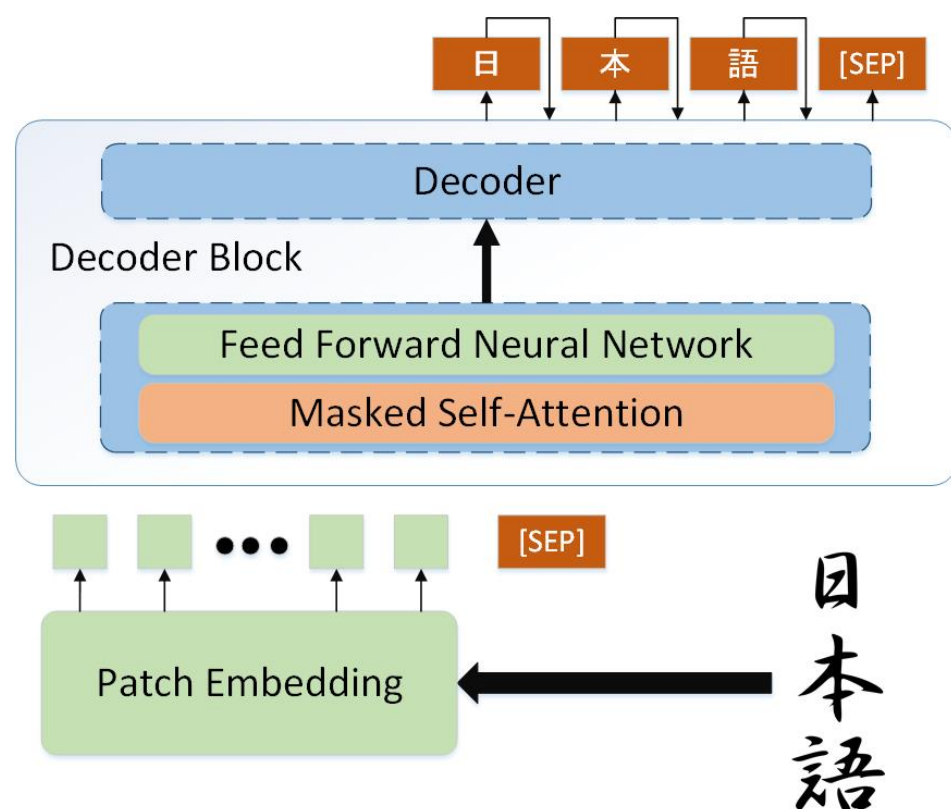


Figure 3. DTrOCR pipeline (adapted from [5]).

4.2. Model Pipeline

For the pretrained language model that serves as a decoder, similarly to [5], we use the Japanese-GPT2-small model from Rinna [23], the latter being a 12-layer, 768-hidden-size transformer-based language model [24] trained on the Japanese CC-100 dataset [25] and Japanese Wikipedia. This model is chosen due to its similarity in parameters to the GPT2 version used in DTrOCR project [5]. For patch embedding, we utilize ViT (Vision Transformer) [26] patch embedding. Both models are available on Hugging Face [27]—a leading present-day open-source AI platform and community.

4.3. Pretraining with Synthesized Datasets

For pretraining, 4,000,000 images were generated: half contain examples of Japanese texts in a modern layout organized in horizontal lines (*yokogaki*, in Japanese), while the remaining 2,000,000 samples contain text in a traditional Japanese writing style in vertical columns (*tategaki*). Text Renderer [28] was used for synthesizing this corpus. Figure 4 shows a few examples of generated images of both types. We used Japanese CC-100 [25] as a basis for synthesizing the data corpus because it was used to train the original GPT2-small language model [24]. Specifically, to ensure the model's resistance to noise, we used 41 different Japanese fonts at font sizes ranging from 10 to 30, while the number of characters ranged between 1 and 8 in different colors.

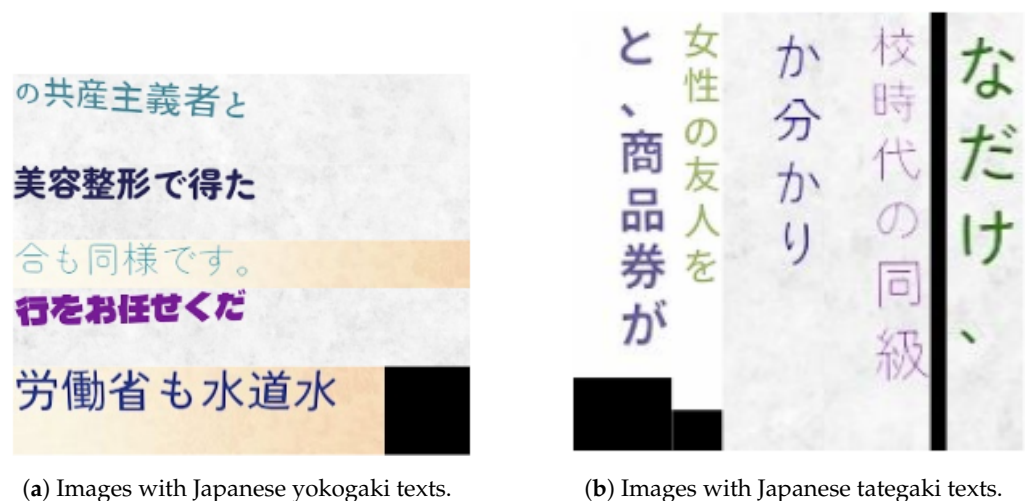


Figure 4. Examples of synthesized images.

Additionally, the original implementation of Text Renderer crops the generated images to the size of the text so that the text is centered and fills the entire image. We decided to disable that cropping with the expectation of improving the robustness of the final model and adjusting it to the different text placements in actual images without additional image processing.

We shuffled and split both the tategaki and yokogaki samples into training and testing sets in a proportion of 90% training data and 10% testing data. Finally, we applied data augmentation to all the synthesized dataset elements. For each image, we generated three additional augmented versions. We applied the RandAugment [29] three-layer policy and the magnitude of three, corresponding to low-intensity transformations. We also added random inversion with a probability of 0.2.

Thus, the overall training was performed using 15,200,000 synthesized images split into 475,000 batches, where each batch contained 32 images.

4.4. Fine-Tuning on the Kuzushiji-Kanji Dataset

Our primary interest is to test the method's ability to be applied to difficult text types that pose serious challenges for reading, comprehension, and automatic recognition. Kuzushiji is a perfect fit. Improving methods for performing automatic recognition of kuzushiji writing is an important current scientific and technological task that has an impact on the development of OCR instruments, cultural studies, and education.

The dataset used in this study is represented by the Kuzushiji-Kanji (KKanji) dataset by ROIS-DS [30], an imbalanced dataset with a total of 3832 Kanji characters (64×64 grayscale, 140,424 images), ranging from 1766 examples to only a single example per class.

One of our research questions was to examine to what extent model performance is affected by additional fine-tuning and training after the initial training. For that purpose, we decided to test it with kuzushiji data. The fine-tuning was conducted following the original research methodology [5]. Before that, we had split the data into training and testing sets with the same ratio as the synthesized data. Similar to the training, all data was passed for fine-tuning in random order.

5. Results

The training and fine-tuning were performed on an RTX 3060. Training on synthesized data with augmentation took approximately 50 h, while fine-tuning on kuzushiji with augmentation took 2 h. We used accuracy as the main metric to assess the model's performance. Additionally, we measured balanced accuracy on KKanji to better account for its imbalanced nature.

The tested scenarios for model training were as follows:

- Model A: Training only on the synthesized corpus;
- Model B: Model A + augmentation;
- Model C: Model B + fine-tuning on kuzushiji;
- Model D: Model C with extra augmentation;
- Model E: Training only on augmented kuzushiji;
- Model F: Model B + extra training on augmented kuzushiji.

Table 1 shows the efficiency of augmentation, which demonstrated a significant impact on the model performance. On synthesized images, the model gained over 10% accuracy in comparison to models A and B. A similar effect was observed when augmentation was applied to the training set of kuzushiji characters used for fine-tuning. Furthermore, fine-tuning has a noticeable negative impact on the accuracy of the original synthesized corpus: the more data used, the more loss of accuracy was observed.

Table 1. Influence of augmentation and fine-tuning on accuracy.

Model	Synthesized	Kuzushiji	Balanced Accuracy
A (synthesized)	78.195%	49.208%	49.231%
B (synthesized+augmented)	87.947%	50.703%	49.289%
C (model B + fine-tune)	86.978%	78.318%	63.547%
D (model C + augmentation)	83.527%	88.923%	73.392%
E (only training on kuzushiji)	12.277%	91.794%	79.408%
F (model B + training on kuzushiji)	28.133%	96.439%	89.832%

Models E and F were trained on kuzushiji, both with the use of augmentation. Model E was trained only on kuzushiji, while model F is based on model B and additionally trained on kuzushiji. As one might expect, model E shows high accuracy on kuzushiji as it was trained specifically for it. Model F shows even better results. However, although

the accuracy of model F is high on kuzushiji tests, it drops significantly on the synthesized corpus. Therefore, in a scenario with an already trained text recognition model, fine-tuning it for a specific task would be a more efficient choice than training it from scratch.

We also included the balanced accuracy values to Table 1, since it is calculated as the average of model sensitivity (true-positive rate) and specificity (true-negative rate) and therefore offers a more reliable performance measure compared to standard accuracy. For most models (except model A) the balanced accuracy values are lower than standard accuracy. This is to be expected, as there are classes weighted the same as others but having very few samples (even just one in multiple cases—Figure 5 shows a few examples). Such classes contribute to producing a negative impact on the overall accuracy. The balanced accuracy decrease is more significant for the models fine-tuned on kuzushiji, since a lower learning rate makes it difficult for the model to correctly learn the features of the less populated classes.

Table 2 illustrates the sensitivity of the model F (without augmentation) to filtering out the classes with a number of training samples fewer than the given threshold. As one can see, both accuracy and balanced accuracy continue to increase as the classes with fewer training samples are filtered out.

Table 2. Influence of sample quantity on balanced accuracy.

Threshold	Accuracy	Balanced Accuracy
5	96.896%	92.314%
10	97.209%	93.524%
20	97.552%	94.388%
50	98.037%	94.865%



Figure 5. Examples of characters represented by just one sample each in the dataset.

6. Case Study: Recognition of Handwritten Calligraphy-Style Writing

Based on the promising results of our experiments with kuzushiji writing, we ran a preliminary experiment with an attempt to use the model trained on kuzushiji to recognize the handwritten calligraphy style text, which is actually a fragment from an anonymized handwritten private letter. Figure 6 draws on the fragments of the document used for our case study experiment.

Such texts often pose difficulties even for readers with a good Japanese language background. We asked a Japanese expert to translate and mark up the input text in order to highlight the characters, which required some more time and context exploration even for an expert to decipher.

Since the handwritten letters contain Japanese alphabetical characters not present in the KKanji dataset, for this case study, model F (see Table 1) was enhanced with additional training on the K-49 dataset by ROIS-DS [30], which is an imbalanced dataset consisting of 49 characters including 48 Hiragana symbols and one Hiragana iteration mark.

Figure 7 depicts the marked-up translation along with the recognition output produced by the OCR system.

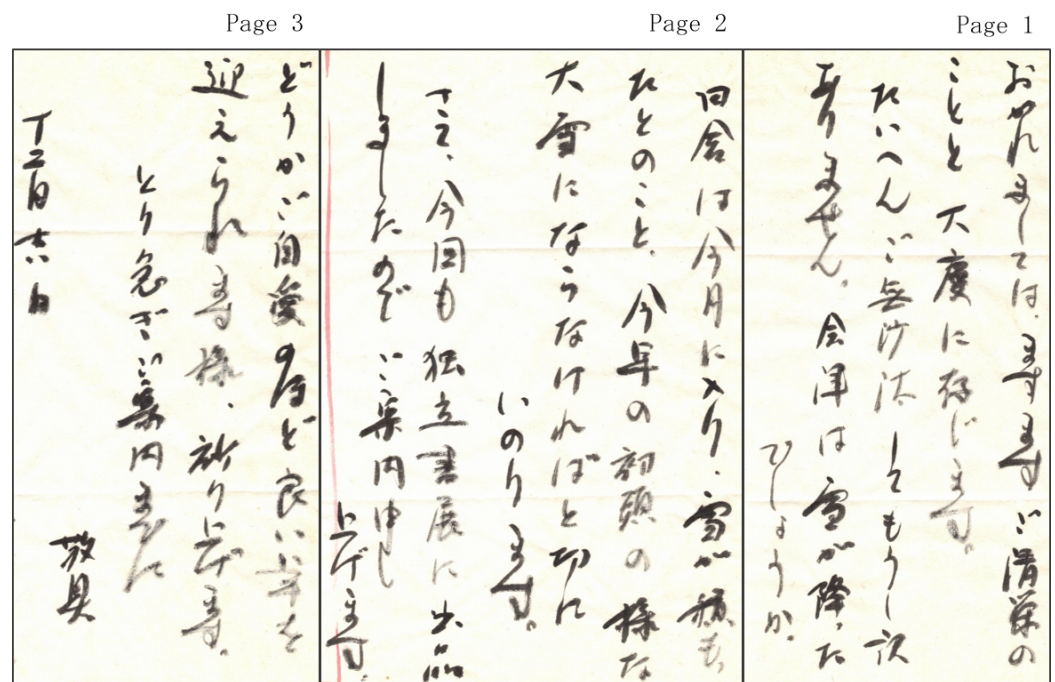


Figure 6. Calligraphy-style handwritten letter fragments. Author's archive.

Page 1	Page 2	Page 3
おかれましては、ますますご清栄のことと大慶に存じます。たいへんご無沙汰してもし訳ありません。会津は雪が降ったでしょうか。	田舎は今月に入り、雪が積もったとのこと。今年の初頭の様な大雪にならなければと切にいのります。さて、今回も独立書展に出品しましたのでご案内申し上げます	どうかご自愛のほど良い年を迎えられます様、祈り上げます。とり急ぎご案内までに。敬具 十二月 吉日
めれましては、ますますと情深のことと大慶に存じます。たいへんご無沙汰してもし訳ありません。合耳は雪が降ったでしょうか	日舎は入月に入り、雪が様もったとのこと。介年の相頭の様な火やになうなればと切にいのります。てわ八回も独立す展に出品しきしたので忘つ申し上げます	とかかこ自煮の住とらいさを逆えられます様、祈り上げます。とか色まに妾肉まかに賞具 十に日 吉

Figure 7. Transcribed letter and OCR results achieved by kuzushiji-trained model enhanced with K-49 [30] dataset training. The red circled characters are those that produced difficulties even for a Japanese expert, while the characters shown in red have not been correctly recognized by the model.

We must acknowledge that the recognition accuracy of this fragment of the handwritten text is just around 68%, which is lower compared to kuzushiji recognition accuracy, and the model struggles with the characters having too few samples in the Kuzushiji dataset. Also, we need to note that for this case study, in our prediction correctness assessment, we ignored the diacritic marks such as dakuten (*Dakuten*—a small diacritic mark placed at the top-right of Japanese hiragana and katakana characters to modify their pronunciation so that the voiced consonants are produced) and handakuten (*Handakuten*—a small diacritic mark placed at the top-right of Japanese hiragana and katakana characters to modify their pronunciation so that the semi-voiced consonants are produced) since the original dataset does not differentiate between them.

Though the results achieved in this case study show that the kuzushiji-trained model applied to the handwritten contemporary calligraphy is not as straightforwardly efficient as it is for predicting the kuzushiji characters themselves, the task suggests an interesting use case for the future work.

7. Discussion

7.1. Observations of the Major Outcomes

The performance of the model on kuzushiji, especially after training, suggests that it has great potential to be used for handwritten-text processing. Model fine-tuning must be enough to reap most of the benefits of training without losing its ability to perform on the other types of texts.

These observations are generally consistent with the findings in DTrOCR [5]. However, while the results reported in [5] suggest that the fine-tuning leads to improvement of recognition quality on all types of data, in our experiments, the accuracy decreased on the synthesized images compared to kuzushiji recognition quality, mainly due to the model being fine-tuned just on kuzushiji data.

Compared to the results reported in [5], our experiments with applying the model to Japanese texts do not reach the same level of performance: lower accuracy values can be explained by a significantly smaller amount of training and fine-tuning data, as well as their insufficient variability. To improve the accuracy of kuzushiji recognition, more samples are required for augmentation along with using oversampling techniques, the latter having been successfully applied to the same dataset as reported in [14].

7.2. Further Steps

After confirming the method's potential for universal Japanese text recognition, the next step would require a more accurate replication of the original research, which would involve a significant increase in the training data, adding synthesized scene and handwriting texts, as well as increasing the amount and variability of data samples.

The text generation software used in the original research does not support Japanese handwritten text; however, there are other models and generation methods available. Additionally, while the real data available for Japanese texts is not as varied and abundant as that for English and Chinese, there are still various available datasets that could be utilized for fine-tuning and training the model, for example:

- ICDAR 2019 multilingual dataset (scene text) [31];
- Manga109 (Japanese comics) [32,33];
- Kuzushiji Recognition (full pages of kuzushiji text) [34].

7.3. Prospective Application Areas

Various domains of digital humanities and education suggest plenty of promising cross-disciplinary applications of the method.

The results achieved during the experiments on kuzushiji recognition and case study on calligraphy-like handwritten letters unveil the possibilities to use similar models for transcribing and searching historical documents, reviving the antique woodcut printed books for contemporary readers, unscrambling the inscriptions in ukiyo-e pictures like those that can be seen in Figure 8, accompanying museum artifacts on exhibition display, and even unraveling the engravings on old memorial stones and monuments. The latter case could be an insightful opportunity to enhance traveling and history learning experiences to make them genuinely meaningful and truly memorable [35].

Language learning applications (specifically, in the domain of iCALL) can benefit from applying the method to the common types of texts, such as printed and scene texts, but also to developing technology-enhanced educational resources on learning kuzushiji characters or other palaeographic [36] writing styles used in ancient Japanese artistic and classical literature.



Figure 8. Portrait of the actor NAKAMURA Shikan by UTAGAWA Yoshiiku (also known as OCHIAI Yoshiiku) with elements of kuzushiji writing. Late Edo period. Author's collection.

Author Contributions: The authors wrote the paper in tight collaboration. Specifically, D.K. led the project, developed the software and ran experiments; N.B. contributed to software design and paper presentation; and E.P. extended the recognition scenarios, designed the paper presentation, extended the manuscript, and prepared its final version for publication. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The sources, including ipynb files for training and evaluation of the models described in the article, along with the modeling results are available in the project repository at <https://github.com/KimDamir/LLM-for-Japanese-Text-OCR-Automating-Kuzushiji-Recognition-Using-Decoder-only-OCR-Architecture> (accessed on 15 July 2025).

Acknowledgments: The authors sincerely thank Yohei Nishidate, Sachiko Ukawa, and Kazuya Ito for their helpful suggestions, which significantly improved our understanding of the particularities of Japanese handwritten and printed cursive writing. We cordially thank Akemi Kusakari for her help in translating the historical titles of old documents used in this paper. We are grateful to John Blake for proofreading the article.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CNN	Convolutional Neural Network
DTrOCR	Decoder-only Transformer for Optical Character Recognition
KKanji	Kuzushiji-Kanji (dataset)
KNN	K-Nearest Neighbors
LLM	Large Language Model
iCALL	Intelligent Computer-Assisted Language Learning
OCR	Optical Character Recognition
ViT	Vision Transformer

References

1. Bautista, D.; Atienza, R. Scene text recognition with permuted autoregressive sequence models. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 178–196.
2. Fang, S.; Xie, H.; Wang, Y.; Mao, Z.; Zhang, Y. Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Virtual, 19–25 June 2021; pp. 7098–7107.
3. Shi, B.; Bai, X.; Yao, C. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 2298–2304. [PubMed]
4. Li, M.; Lv, T.; Chen, J.; Cui, L.; Lu, Y.; Florencio, D.; Zhang, C.; Li, Z.; Wei, F. Trocr: Transformer-based optical character recognition with pre-trained models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Washington, DC, USA, 7–14 February 2023; Volume 37, pp. 13094–13102.
5. Fujitake, M. DTrOCR: Decoder-only transformer for optical character recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, HI, USA, 3–8 January 2024; pp. 8025–8035.
6. Mugikura, S. Kuzushiji Recognition with AI. 2022. Available online: <https://news.fontstand.com/essays/kuzushiji-recognition-with-ai/> (accessed on 17 February 2026).
7. Galimberti, S. The Influence of Ukiyo-e on Modern Architecture: From Art to Structure. 2024. Available online: https://www.researchgate.net/publication/393660258_The_Influence_of_Ukiyo-e_on_Modern_Architecture_From_Art_to_Structure?channel=doi&linkId=6874c02b0d8ed26a9d5ad392&showFulltext=true (accessed on 17 February 2026).
8. Jackson, D. Ukiyo-e Inspiring It All. *Agora Red Deer Polytech. Undergrad. J.* **2024**, *15*, 35–41. [CrossRef]
9. Harris, F. *Ukiyo-e: The Art of the Japanese Print*; Tuttle Publishing: North Clarendon, VT, USA, 2012.
10. Naruhodo. Introduction to Kuzushiji. 2012. Available online: <https://naruhodo.weebly.com/blog/introduction-to-kuzushiji> (accessed on 18 December 2025).
11. Columbia University East Asian Doctoral Program. Available online: <https://history.columbia.edu/east-asia/> (accessed on 17 February 2026).
12. Japanese Calligraphy and Manuscript Materials Course, University of Warsaw. Available online: https://informatorects.uw.edu.pl/en/courses/view?prz_kod=3600-JA-KJKMR-OW (accessed on 17 February 2026).
13. Matsuo, S. (Ed.) *Jiten Kana: Shutten Meiki. Kaichōban (Kana Dictionary: Sources Specified. Revised Edition)*; Kasama Shoin: Tokyo, Japan, 1972. (In Japanese)
14. Lippert, C.; Junger, A.; Rasul, M.G.; Yakub, M.; Shahriar, M.H.; Chowdhury, M. Kuzushiji (Japanese Text) Classification. 2022. Available online: https://www.researchgate.net/publication/364336274_Kuzushiji_Japanese_Text_Classification (accessed on 17 February 2026).
15. Le, A.; Kitamoto, A. Training Kindai OCR with parallel textline images and self-attention feature distance-based loss. *arXiv* **2025**, arXiv:2508.08537.
16. Kobayashi, Y.; Mimuro, S.; Suzuki, S.; Iijima, Y.; Okada, A. Basic research on a handwritten note image recognition system that combines two OCRs. *Procedia Comput. Sci.* **2021**, *192*, 2596–2605. [CrossRef]
17. Sasagawa, K.; Kurita, S.; Kawahara, D. Evaluating multimodal large language models on vertically written Japanese text. *arXiv* **2025**, arXiv:2511.15059.
18. Gregory, J.G. How Learning Works; Can LLMs learn “impossible” languages? *Commun. ACM* **2026**, *68*, 14–16.
19. Kallini, J.; Papadimitriou, I.; Futrell, R.; Mahowald, K.; Potts, C. Mission: Impossible language models. *arXiv* **2024**, arXiv:2401.06416.
20. Lu, N.; Yu, W.; Qi, X.; Chen, Y.; Gong, P.; Xiao, R.; Bai, X. MASTER: Multi-aspect non-local network for scene text recognition. *Pattern Recognit.* **2021**, *117*, 107980. [CrossRef]

21. Yu, D.; Li, X.; Zhang, C.; Liu, T.; Han, J.; Liu, J.; Ding, E. Towards accurate scene text recognition with semantic reasoning networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 12113–12122.
22. Lyu, P.; Zhang, C.; Liu, S.; Qiao, M.; Xu, Y.; Wu, L.; Yao, K.; Han, J.; Ding, E.; Wang, J. MaskOCR: Scene Text Recognition with Masked Vision-Language Pre-Training. *Trans. Mach. Learn. Res.* **2024**. Available online: <https://openreview.net/forum?id=KNAWoKKpi3> (accessed on 17 February 2026).
23. Zhao, T.; Sawada, K. Rinna/Japanese-Gpt2-Small-Pre-Trained Models for Japanese Language. Available online: <https://huggingface.co/rinna/japanese-gpt2-small> (accessed on 17 February 2026).
24. Sawada, K.; Zhao, T.; Shing, M.; Mitsui, K.; Kaga, A.; Hono, Y.; Wakatsuki, T.; Mitsuda, K. Release of pre-trained models for the Japanese language. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), Torino, Italy, 20–25 May 2024; pp. 13898–13905.
25. Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; Stoyanov, V. Unsupervised cross-lingual representation learning at scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 8440–8451.
26. Wu, B.; Xu, C.; Dai, X.; Wan, A.; Zhang, P.; Yan, Z.; Tomizuka, M.; Gonzalez, J.; Keutzer, K.; Vajda, P. Visual transformers: Token-based image representation and processing for computer vision. *arXiv* **2020**, arXiv:2006.03677.
27. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Passos, A.; Koukpoulos, F.; Moisé, A.; Cistac, P.; Polak, O.; et al. HuggingFace’s transformers: State-of-the-art natural language processing. *arXiv* **2019**, arXiv:1910.03771.
28. Text Renderer. 2021. Available online: https://github.com/Sanster/text_renderer (accessed on 17 February 2026).
29. Cubuk, E.D.; Zoph, B.; Shlens, J.; Le, Q.V. Randaugment: Practical automated data augmentation with a reduced search space. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 702–703.
30. Clanuwat, T.; Bober-Irizar, M.; Kitamoto, A.; Lamb, A.; Yamamoto, K.; Ha, D. Deep Learning for Classical Japanese Literature. *arXiv* **2018**, arXiv:1812.01718.
31. Nayef, N.; Patel, Y.; Busta, M.; Chowdhury, P.N.; Karatzas, D.; Khelif, W.; Matas, J.; Pal, U.; Burie, J.C.; Liu, C.I.; et al. ICDAR2019 Robust reading challenge on multi-lingual scene text detection and recognition – RRC-MLT-2019. In Proceedings of the 2019 International Conference on Document Analysis and Recognition (ICDAR), Sydney, Australia, 20–25 September 2019; pp. 1582–1587. [CrossRef]
32. Aizawa, K.; Fujimoto, A.; Otsubo, A.; Ogawa, T.; Matsui, Y.; Tsubota, K.; Ikuta, H. Building a Manga dataset “Manga109” with annotations for multimedia applications. *IEEE MultiMedia* **2020**, *27*, 8–18. [CrossRef]
33. Matsui, Y.; Ito, K.; Aramaki, Y.; Fujimoto, A.; Ogawa, T.; Yamasaki, T.; Aizawa, K. Sketch-based Manga retrieval using Manga109 dataset. *Multimed. Tools Appl.* **2017**, *76*, 21811–21838. [CrossRef]
34. anokas; Kitamoto, A.; Park, E.; Dane, S.; TheNuttYNetter; tkasasagi; Kan, W. Kuzushiji Recognition. 2019. Available online: <https://kaggle.com/competitions/kuzushiji-recognition> (accessed on 17 February 2026).
35. Pyshkin, E. Designing human-centric applications: Transdisciplinary connections with examples. In Proceedings of the 2017 3rd IEEE International Conference on Cybernetics (CYBCONF), Exeter, UK, 21–23 June 2017; pp. 1–6. [CrossRef]
36. Morris, J.H. Digital Resources for Japanese Palaeography. 2020. Available online: <https://digitalorientalist.com/2020/10/31/digital-resources-for-japanese-palaeography/> (accessed on 12 February 2026).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.