Leveraging External Knowledge for Historical Document Restoration via Retrieval-Augmented Large Language Models

Gabeen Kim[1] **Kyeongpil Kang**[2]*

[1]Department of AI Convergence, Kangwon National University
[2]Department of Computer Science and Engineering, Kangwon National University
gokong0516@kangwon.ac.kr rudvlf0413@kangwon.ac.kr

Abstract

Historical documents act as invaluable knowledge archives but often suffer from illegibility due to physical deterioration and damage. While existing restoration methods based on masked language modeling effectively utilize local context, they struggle to restore named entities that require external historical knowledge. To address this limitation, we introduce a novel framework for historical document restoration that leverages large language models with retrieval-augmented generation (RAG). By combining the implicit knowledge of pretrained LLMs with explicitly retrieved external context, our model ARI effectively mitigates the challenge of inferring context-dependent proper nouns. Extensive experiments on Korean historical documents demonstrate that our approach significantly outperforms baselines, achieving substantial gains in restoring both general characters and named entities. Furthermore, comprehensive evaluations including expert assessments confirm that **ARI** serves as a practical tool for domain experts, promising to accelerate the analysis of historical records.

1 Introduction

Historical records are repositories of vast information spanning centuries or even millennia. Recognizing the global importance of these records, national governments and researchers are working not only to preserve these ancient documents but also to uncover and analyze the knowledge they contain. For instance, the Annals of the Joseon Dynasty (**AJD**)[1] and the Journal of the Royal Secretariat (**JRS**)[2], which span 500 years of history and contain not only historical facts but also natural events, are inscribed on the UNESCO Memory of the World Register. These archives are extensively analyzed by researchers across diverse domains to derive valuable insights.

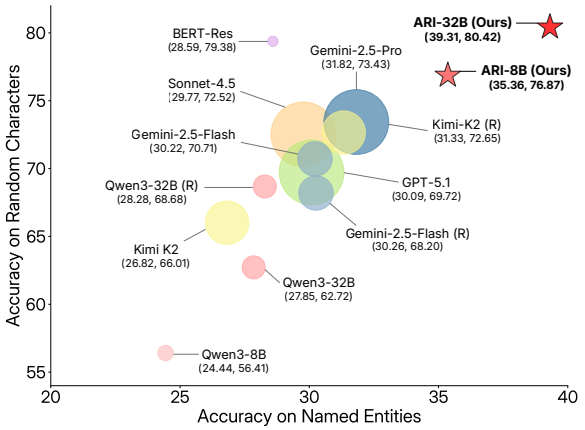
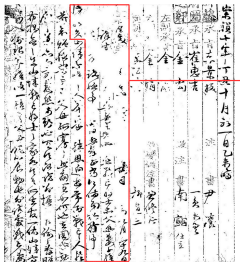


Figure 1: Performance of the proposed model compared with the baselines. The x- and y-axes represent the restoration accuracies for named entities and randomly masked characters, respectively.

While these historical archives possess immense value, preserving them and extracting knowledge present significant challenges. A primary issue is the inherent vulnerability of historical manuscripts to damage and degradation. These documents were typically inscribed on materials such as wood or plant fibers, which are far less durable than modern paper. Consequently, they are highly sensitive to environmental conditions, prone to discoloration and structural deterioration upon exposure to light, heat, or moisture. Furthermore, prior to the advent of metal movable type, texts were transcribed manually using brushes or pens. Accordingly, legibility was highly contingent on the scribe's penmanship, which inevitably introduced transcription errors. Such compromised sections are difficult for both OCR systems and human experts to analyze. As a result, they are often treated as blanks or damaged tokens, which significantly hinders downstream analysis. For instance, by matching those damaged tokens, we identified that approximately

*Corresponding author
[1]https://sillok.history.go.kr/intro/english.do
[2]https://sjw.history.go.kr/intro/engInfo.do

Findings of the Association for Computational Linguistics: ACL 2026, pages 43290–43304
July 2-7, 2026



○上在昌慶宮。□常參·經□□邊司□□□九月二十九日□講時□□□□□李景奭所啓，戰亡將士等父母妻子，給□□□□□護恤事，上曰，啓辭甚當，令備局施行事，傳敎矣。

Diaries of the Royal Secretariat, vol. 61, 1/10/1637 (15th year of King Injo)

Figure 2: Real-world example of a damaged historical document with masked characters (□).

41.9K characters across 11.1K documents in the JRS remain damaged or unrecognizable.

To restore illegible parts of historical archives, several approaches have been proposed. Kang et al. (2021) addressed the restoration of damaged documents by training a model via masked language modeling (MLM). However, a key limitation of these methods is their exclusive reliance on the damaged document itself, neglecting external information. Integrating external knowledge is crucial for correctly inferring named entities, particularly proper nouns. For example, in the sentence "In 1492, Columbus first landed in **[M]**.", predicting the masked token **[M]** is challenging using only the local context. In such cases, the model must either possess internalized historical knowledge or utilize external resources to ensure accurate restoration.

To overcome this limitation, we introduce a novel framework for historical document restoration utilizing LLM-based retrieval-augmented generation. Our method utilizes a pre-trained LLM that has indirectly learned from extensive datasets including historical knowledge and documents available on the web. This implicitly learned context acts as vital external knowledge, significantly improving the restoration of damaged characters. Additionally, we integrate RAG to allow the model to explicitly access external knowledge by incorporating relevant documents into the prompt. Finally, we developed our model, **ARI** (**A**rchive **R**estoration **I**ntelligence), specifically trained on ancient corpora. Our model demonstrates performance that surpasses that of baselines. We will make our model and data publicly available upon publication to facilitate future research.

2 Related Work

Ancient documents serve as crucial data sources with high research value across diverse domains, ranging from history to the social and natural sciences. However, as previously discussed, many historical records have been damaged from various factors, inevitably limiting their research utility.

Accordingly, methods have been proposed to restore these damaged parts. Initial efforts focused on deep learning approaches to correct typographical mistakes in historical texts (Tang et al., 2018; Domingo and Nolla, 2018). Pythia, designed to predict missing characters in Greek epigraphy, achieved a top-20 accuracy of 73.5% (Assael et al., 2019). Kang et al. (2021) achieved both effective restoration and translation by jointly training MLM and machine translation. While Ithaca enhanced interpretability by jointly predicting textual, geographical, and chronological attributes (Assael et al., 2022), our target corpora (AJD and JRS) inherently lack geographical metadata. Leveraging the improved general performance and linguistic capabilities of pretrained LLMs, Liu et al. (2025) demonstrated that restoration is feasible in a zero-shot fashion.

However, prior works have restricted the model input to a single corrupted document, limiting predictions to the document's internal context. As a result, these methods exhibit significant shortcomings in restoring information that necessitates external knowledge, such as named entities. This highlights the necessity of leveraging external data sources to effectively restore damaged historical documents. Another limitation stems from the fixed masking probability used during training (Kang et al., 2021; Assael et al., 2022). This creates a discrepancy, as the masking rate does not match the corruption distribution in actual documents. This mismatch between training and inference conditions can degrade the real-world performance.

Language models trained on extensive web data demonstrate superior general performance. BERT (Devlin et al., 2019), for example, utilizes a bi-directional Transformer Encoder trained with MLM that predicts original tokens by masking random positions in the training data. Given its resemblance to recovering damaged text, MLM has been extensively used for historical document restoration. However, BERT-based approaches have a key drawback: They predict masked tokens in a non-autoregressive manner, often missing the dependencies among them. Recently, causal language modeling has become the standard pre-training method for the latest LLMs, such as Gemini-2.5 (Comanici et al., 2025), Sonnet 4.5 (Anthropic, 2025), Kimi-K2 (Team et al., 2025), GPT-5.1 (OpenAI, 2025),

and Qwen3 (Yang et al., 2025), due to the simplicity of its training objective. Effectively ingesting extensive linguistic data and general knowledge, these models internalize implicit knowledge from vast archives (e.g., Chinese documents, Korean Hanja texts, and historical records). This model-intrinsic knowledge can then be leveraged as a powerful source of external knowledge for the restoration of damaged historical documents.

Additionally, RAG has been developed to improve LLM performance by directly integrating external knowledge. RAG operates by retrieving documents relevant to a user's query via lexical-based (e.g., BM25) or embedding-based methods from external repositories like the web or databases (Lewis et al., 2020). Providing these documents as context enables the model to utilize information not learned during pre-training and effectively alleviates hallucinations. RAG is particularly effective for domain-specific and knowledge-intensive tasks, such as those involving low-resource languages (Nie et al., 2023; Chang et al., 2025) and named entity recognition (Xie et al., 2025). In addition, research suggests that applying fine-tuning to RAG systems enables them to better leverage relevant documents, resulting in significant performance gains on downstream tasks (Zhang et al., 2024).

In this study, we propose a method for effectively restoring historical documents by leveraging external knowledge. We utilize LLM that employs a causal language modeling approach to model the contextual dependencies of damaged characters. Additionally, we employ RAG to provide relevant documents to the model, thereby allowing it to draw on external knowledge. This approach is further refined through fine-tuning, which optimizes the model's capacity to leverage external information and enhances final restoration accuracy.

3 Dataset

In this study, we utilized AJD and JRS to train and evaluate our restoration model. Both corpora consist of documents written in Hanja and contain rich metadata, including temporal information (e.g., year, month, and date) and named entities (e.g., personal names, organizations, book titles) provided by the National Institute of Korean History.[3] Table 1 presents the statistics for the raw data. We observed a significantly higher prevalence of damaged characters in the JRS compared to the AJD. Furthermore, within the documents containing damage, the average number of damaged characters was 3.77 per document. To investigate the characteristics of these damaged characters in detail, we employed Gemini-2.5-Pro to predict the named entity type for each character.[4] As illustrated in Fig. 3, approximately 44.8% of the 42K damaged characters were identified as named entities. A breakdown of the categories reveals that PER (Person) was the most frequent, followed by LOC (Location), ETC (Organizations, Official Titles, etc.), DAT (Date), and POH (Publication of History). While Gemini-2.5-Pro may not be perfectly accurate, these results function as a proxy, indirectly reflecting the distribution trends regarding the presence and types of named entities. Therefore, these findings underscore the importance of character restoration not merely for arbitrary positions, but specifically for named entity segments.

We strictly held out documents containing real-world corrupted parts for the human-evaluation dataset D_{RD}. The remaining data comprised 10K

	AJD	JRS
Number of Docs.	0.37M	1.75M
Number of Chars.	71.9M	292.6M
Avg. NEs per Doc.	1.79	4.73
Number of Docs. w/ [D]	56	11.1K
Number of [D]	0.11K	41.9K
Avg. [D] per Doc.	1.98	3.77

Table 1: Statistics of the AJD and JRS raw data corpora. NEs and [D] represent named entities and damaged tokens, respectively.

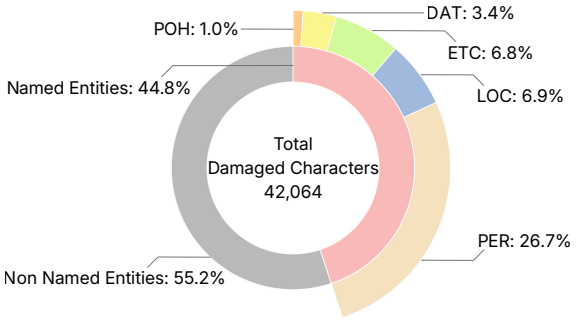


Figure 3: Pie chart depicting the distribution of damaged characters, including the categorical ratios of named entities: PER, LOC, DAT, POH, and ETC.

[3] https://www.history.go.kr/en/main/main.do

[4] To evaluate the reliability of the entity classification labels generated by Gemini-2.5-Pro, we manually inspected the entities in 200 randomly sampled documents. The results showed that only 4 instances (0.7%) were misclassified.

documents each for testing and validation, with the other 2.02M used for training. To prevent data leakage, we strictly removed from the training corpus any samples duplicated in the validation or test sets. This training corpus acted as the retrieval source for the RAG system and was used to train our model and baselines. For building the test dataset using the test corpus, we introduced synthetic corruptions at random positions, taking into account the distribution of corruption span lengths observed in real damaged documents. We set the corruption rate stochastically to best emulate the original data's corruption distribution. Specifically, the rate was centered around a mean of 2.96% (the average rate from the original corrupted documents), and we guaranteed a minimum of one corrupted character per document. In addition, assessing the reconstruction accuracy of the damaged parts is crucial, especially in cases requiring external data. Accordingly, we designed our valid and test datasets to comprise the following two versions:

- D_{Rand}: Dataset where synthetic corruptions are applied at random positions. This dataset serves to assess the overall restoration performance of the model.
- D_{NE}: Dataset containing synthetic corruptions originating from named entities. This dataset is used to evaluate the model's ability to restore segments requiring external knowledge.

These evaluation datasets consist of pairs of partially corrupted input documents and their corresponding ground-truth originals. The data construction pipeline is detailed in Section 4.

4 Proposed Methods

As shown in Fig. 4, this section presents our restoration framework. To validate the framework, we first evaluate the performance of baseline models. We then demonstrate how restoration accuracy can be improved by leveraging external knowledge via metadata and RAG. Finally, we present our proposed model, ARI, alongside the baseline, BERT-Res trained from scratch on our dataset.

4.1 LLM Baseline Restoration Performance

Kang et al. (2021) utilized a Transformer Encoder model to directly restore original characters at each masked position. However, given the autoregressive nature of most pre-trained language models, it is crucial to explicitly identify the index (position) of each corrupted character and its corresponding restoration within the input and output. Therefore, as shown in Fig. 5, we designed the base prompt format to incorporate positional information for each corrupted character and extract the restored characters accordingly.

Model	Thinking	Acc. (NE)	Acc. (Rand)
Qwen3 8B	No	2.86	7.43
Qwen3 32B	No	5.57	13.37
Kimi K2	No	5.83	22.45
Kimi K2	Yes	6.53	20.66
Gemini-2.5-Flash	No	6.82	24.38
Gemini-2.5-Flash	Yes	7.10	22.35
Gemini-2.5-Pro	Yes	9.57	**32.62**
GPT-5.1	Yes	10.03	28.69
Sonnet 4.5	Yes	**13.30**	32.35

Table 2: Reconstruction performance (Top-1 accuracy) of LLMs on D_{NE} and D_{Rand} with the base prompt.

We then evaluated the basic restoration performance by measuring character prediction accuracy on the valid dataset across multiple LLMs. The accuracy for each model is presented in Table 2. Within the Qwen3 and Gemini-2.5 families, we observed a positive correlation between model scale and restoration accuracy. Notably, the results from Kimi K2 and Gemini-2.5-Flash demonstrate that enabling thinking mode fosters deeper reasoning, thereby enhancing named entity restoration capabilities. However, a slight decline was observed in the restoration performance for random characters. Gemini-2.5-Pro and Sonnet 4.5 surpassed other models in restoring characters at random positions, reflecting their linguistic and contextual understanding of Hanja characters. For named entities, Sonnet 4.5 achieved the highest performance. This suggests that Sonnet 4.5 likely engaged in more extensive learning of historical background knowledge during the pre-training phase compared to other models.

4.2 Enhancing Restoration Performance via External Knowledge

External knowledge plays a crucial role in the precise restoration of damaged characters. To leverage this, we propose an enhanced prompting strategy that integrates chronological metadata directly into the LLM context. Specifically, we augment the input with temporal identifiers such as the reigning

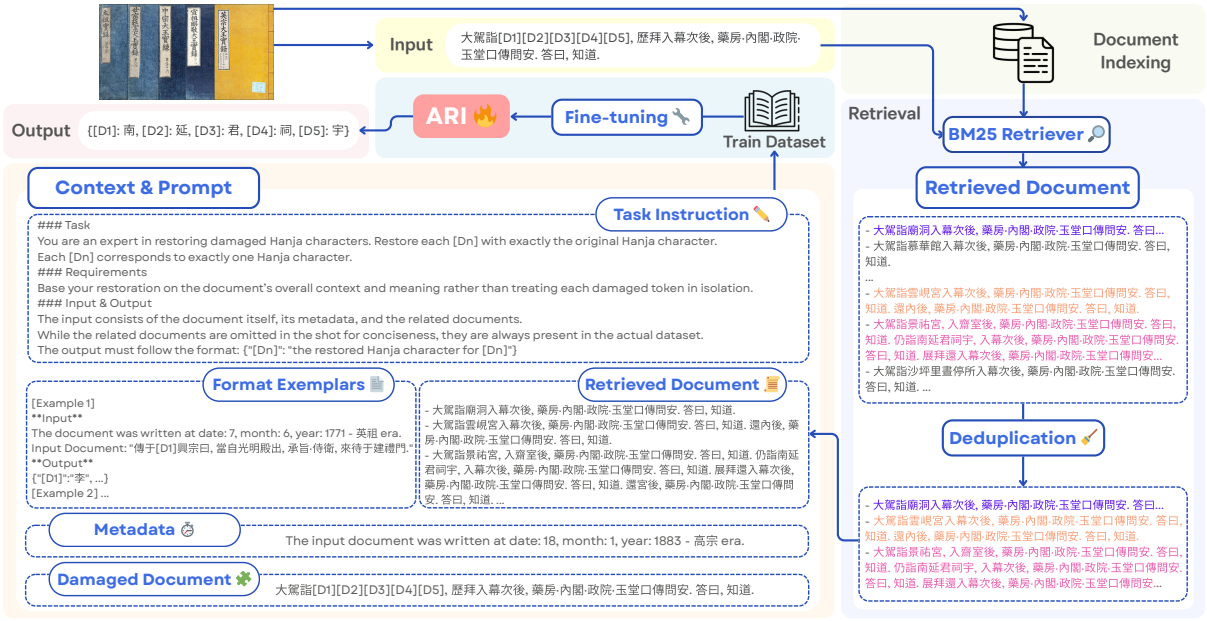


Figure 4: Overview of the proposed framework for restoring damaged documents. The inputs to the LLM include a task description, examples of formats, documents related to the damaged text, and metadata such as temporal information. Based on this structure, ARI is fine-tuned specifically for the historical document restoration task.

```
System prompt: Restore [Dn] with the
original Hanja character. Each [Dn]
corresponds to exactly one Hanja
character. Output must be a valid json
object with the following format:
{“[Dn]”: “Restored character for [Dn]”}
Input: "A[D1]CDE[D2]G"
Output: {"[D1]": "B", "[D2]": "F"}
```

Figure 5: An example of the base prompt for LLM-based restoration.

	Acc. of NE	Acc. of Rand
Baseline	5.57	13.37
+ Metadata	**5.63**	**14.83**

Table 3: Restoration performance results based on the inclusion of metadata.

king, year, month, and day. As detailed in Table 3, this integration improves the model's reconstruction accuracy, validating that temporal grounding is essential for effective restoration.

Beyond simply adding metadata, a key contribution of our approach is the strategic integration of external knowledge via few-shot prompting and Retrieval-Augmented Generation. We specifically designed our experiments to demonstrate how injecting this external context augments the restoration capabilities of the Qwen3 32B. The significant performance improvements yielded by these techniques are detailed in Table 4.

	Acc. of NE	Acc. of Rand
Baseline	5.63	14.83
+ Static shots	6.27	18.98
+ Dynamic shots		
• Random	8.55	26.18
• Embedding	19.88	59.91
• BM25	24.28	60.36
+ Deduplication	**27.85**	**61.45**

Table 4: Impact of adding few-shot, RAG, and deduplication on restoration performance.

To ensure the model's strict adherence to the required output schema, we initially incorporated five few-shot format exemplars into the system prompt. These examples comprise distinct input-output pairs following a format similar to that shown in Fig. 5. Empirical results indicate that providing these format guides significantly reduces formatting errors and aligns the model's generation with the intended restoration protocol.

Furthermore, to incorporate external knowledge via RAG, the top 20 most relevant documents for

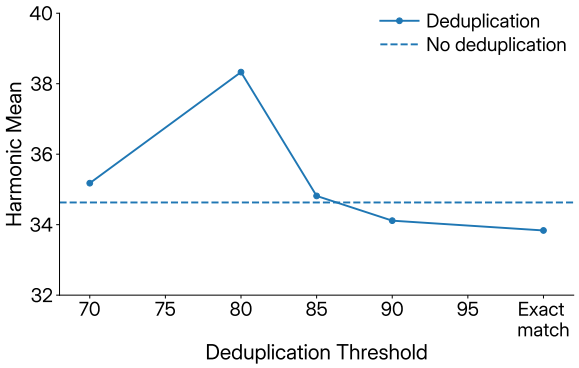


Figure 6: Restoration performance across deduplication thresholds. The plot represents the harmonic mean of D_{NE} and D_{Rand} restoration performances.

each input were retrieved from the training corpus. We evaluated three retrieval methods: random selection, embedding-based retrieval, and BM25. Specifically, we used Gemini Embedding (Lee et al., 2025) for dense retrieval and BM25S (Lù, 2024) for sparse retrieval. The results indicate that even randomly selecting reference documents improves performance compared to using no references. These random examples function as few-shot demonstrations, providing the model with general knowledge regarding Hanja and its usage. However, both embedding-based retrieval and BM25 significantly outperformed random selection, confirming the effectiveness of retrieving contextually relevant documents. Notably, BM25 achieved the highest performance among the three methods. This suggests that character matching is particularly effective for restoring damaged Hanja characters, due to the logographic nature of Hanja where each character encapsulates a specific concept. Further analysis of the retrieval strategies is provided in Appendix B.1.

Furthermore, the corpus contains duplicates and near-duplicates, which are common in short texts. Retaining such redundant content in the retrieved documents limits the model's access to diverse information and risks introducing bias. Consequently, we employ a deduplication method based on string similarity[5] among the retrieved documents. Fig. 6 illustrates the impact of various deduplication thresholds on restoration performance. The solid line depicts the harmonic mean of the performance on D_{NE} and D_{Rand}, while the dotted line indicates the baseline without deduplication.

Significant improvements over the baseline are observed at thresholds of 70% and 80%, with performance peaking at 80%. However, performance declines as the threshold exceeds 85%, reaching its lowest point at the exact-match threshold (100%), where only identical documents are removed. This suggests that preserving highly similar documents restricts the scope of external knowledge and biases the model toward duplicates, thereby degrading restoration performance. By contrast, ensuring diversity among relevant documents through effective deduplication enhances performance by providing richer contexts and cues for restoration. Therefore, we filter out documents that exceed a similarity threshold of 80%. This approach yields further performance improvements in restoring both named entities and randomly positioned content.

Overall, our experiments indicate that few-shot prompting and RAG act synergistically to enhance LLM-based restoration. This combination facilitates task comprehension and external knowledge retrieval, which are critical for accurately restoring damaged characters. The detailed prompt format is provided in Fig. 14 in Appendix E.

4.3 Fine-Tuning the ARI Model

To construct the training dataset for fine-tuning, we masked characters at random positions within a training corpus explicitly isolated from the test corpus. As shown in Table 7, prioritizing named entities masking in 25% of the training data enhanced restoration performance on D_{NE} while maintaining performance on D_{Rand}. Accordingly, we applied this named entity-prioritized masking strategy to 25% of the training dataset.

We then fine-tuned an open-source LLM with the prompt format described above to build a restoration model specialized for the Hanja domain. To mitigate overfitting and improve pattern diversity, we adopted a dynamic masking strategy in which mask positions are varied across epochs, following RoBERTa (Liu et al., 2019). Finally, we developed our model, **ARI-32B**, by fine-tuning Qwen3 32B, incurring a computational cost of approximately 1,500 H200 GPU hours. In addition, we trained a computationally efficient variant, **ARI-8B**, based on Qwen3 8B, while maintaining competitive performance against other baseline models.

Additionally, we implemented a baseline following Kang et al. (2021) that relies solely on the damaged input text, without access to external context. We denote the model that prioritizes named entities for masking at a 25% rate as **BERT-Res**. We employed ModernBERT-large (Warner et al., 2025)

[5] https://github.com/rapidfuzz/RapidFuzz

for its architectural advancements over the original BERT. Since the original ModernBERT vocabulary lacks sufficient coverage for Hanja characters, we extended the tokenizer with Hanja characters. The model was then trained from scratch on our restoration dataset for 10 epochs to ensure convergence. This baseline therefore provides an appropriate setting for evaluating the impact of external knowledge, as it performs restoration without access to external context. Training details, including hyperparameters for our model and the baseline, are provided in Appendix A.

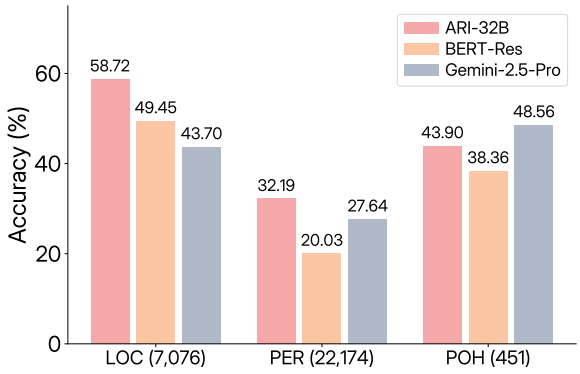


Figure 7: Top-1 accuracy comparison across different named entity categories on the D_{NE}.

5 Experimental Results

5.1 Evaluation on Test Dataset

To evaluate our model against the BERT-Res baseline and other LLMs, we measured top-K accuracy on D_{Rand} and D_{NE}. Specifically, we employed the identical input prompt for both our model and the open-source/proprietary LLMs to ensure the consistent utilization of external knowledge. As shown in Fig. 1, the restoration performance for D_{NE} is significantly lower than that for D_{Rand}. This implies that restoring named entities is more challenging than restoring general characters.

Notably, LLMs utilizing only external knowledge without fine-tuning outperformed BERT-Res in named entity restoration. This demonstrates that incorporating external knowledge effectively enhances restoration capabilities for named entities. Regarding model scale, the Qwen3 32B exhibited superior performance compared to the 8B variant within the same family, confirming that model size significantly impacts performance. With respect to reasoning strategies, we observed performance improvements across Qwen3 32B, Kimi K2, and Gemini-2.5-Flash in random position restoration tasks. Similar improvements were noted in named entity restoration, with the exception of Gemini-2.5-Flash. These findings suggest that reasoning mechanisms enable LLMs to better leverage knowledge acquired during pre-training, facilitating the resolution of complex restoration tasks.

The BERT-Res model, trained from scratch without external knowledge, outperformed Gemini-2.5-Pro in the general character restoration task. This result underscores the importance of training for restoring general characters in historical documents. Furthermore, despite having fewer parameters, the model surpassed decoder-based models, likely due to its bidirectional encoder architecture. However, its performance on named entities fell short of other LLMs, suggesting that relying solely on intra-document context is insufficient. Ultimately, the highest restoration performance was achieved by ARI-32B, which integrates external knowledge via RAG with fine-tuning. Our model demonstrates superior performance for both general text and named entities compared to other open-source and proprietary LLMs, including those with significantly larger parameter counts.

We also conducted a fine-grained performance evaluation stratified by named entity types. First, we selected samples containing explicit named entity tags from the test dataset, D_{NE}. We then constructed a specialized evaluation dataset of 7,707 documents by artificially masking only the target named entity tokens. For named entity classification, we employed Gemini-2.5-Pro. The entities were categorized into Person (PER), Location (LOC), and Publication of History (POH), comprising 22,174, 7,076, and 451, respectively. Finally, we measured the Top-1 accuracy for each category to compare the restoration performance across models; the results are presented in Fig. 7.

First, ARI-32B demonstrated superior performance in named entity restoration, exceeding both BERT-Res and Gemini-2.5-Pro in the LOC and PER categories. Conversely, Gemini-2.5-Pro achieved the highest performance in the POH category, where such instances are less prevalent. Our qualitative analysis revealed that the POH category contains a significant number of Confucian classics, which originated in China and are widely shared across East Asia. This suggests that Gemini-2.5-Pro benefits from extensive knowledge of Chinese classics acquired through multilingual pre-training, unlike models specialized solely for the Joseon

	ARI-32B	BERT-Res	Gemini
Accuracy@1	**0.383**	0.207	0.273
Accuracy@10	**0.583**	0.403	-
nDCG@10	**0.477**	0.296	-
Win Ratio (%)	**46.0**	24.0	30.0

Table 5: Restoration performance of ARI-32B, BERT-Res, and Gemini-2.5-Pro as evaluated by human experts.

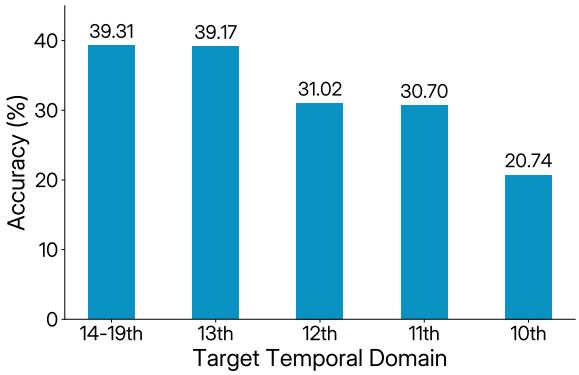


Figure 8: Restoration performance across unseen temporal domains, highlighting the impact of temporal shifts.

dynasty domain, resulting in relatively higher performance in this specific category.

5.2 Evaluation by Human Experts on Real-World Damaged Documents

To evaluate the efficacy of our restoration model as a practical collaborative tool, we engaged three experts in Sinographic literature and Korean history. We randomly sampled 100 documents from D_{RD}, which consists of real-world damaged documents. Remarkably, when classifying named entities using Gemini-2.5-Pro based on prior methods, we observed a high density of 74.6%.

For the evaluation, we presented restoration candidates generated by ARI-32B, BERT-Res, and Gemini-2.5-Pro[6] to experts in a blinded manner. Unlike the black-box nature of Gemini, ARI-32B and BERT-Res provide access to logits, enabling the generation of Top-K candidates as described in Appendix C. Experts selected valid candidates based on contextual coherence, grammatical correctness, and factual validity. Quantitatively, we designated these expert selections as the ground truth to compute Top-1 and Top-10 accuracy, as well as nDCG@10. We also measured the win rate to determine which model served as the most effective assistive tool, considering both the accuracy and diversity of the candidates.

As demonstrated in Table 5, ARI-32B outperforms both Gemini-2.5-Pro and BERT-Res in terms of Top-1 and Top-10 accuracy. Regarding nDCG@10, the model consistently ranked expert-validated candidates higher than BERT-Res. These results indicate that our model restores damaged documents more accurately than the baseline models. Additionally, it achieved a higher win ratio, reflecting a strong preference for its utility as a collaborative restoration tool. This suggests that ARI-32B effectively places reliable candidates at the top ranks while ensuring diversity, significantly reducing the search cost for experts in identifying the correct characters.

[6]As Gemini-2.5-Pro operates as a black-box model without logit access or beam search, we evaluated only its Top-1 candidate and excluded ranking-dependent metrics.

5.3 Qualitative Analysis of Restoration

To evaluate qualitative restoration performance, Table 6 presents restoration results from ARI-32B, Gemini-2.5-Pro, and BERT-Res on representative sample documents.

The first example involves the restoration of person names and general text. Although all models correctly recovered the general text in [D4], only ARI-32B successfully restored the person names in [D1], [D2], and [D3]. The second example focuses on book titles and general text. Both Gemini-2.5-Pro and ARI-32B successfully restored the book titles in [D1] and [D2]. For the general text in [D3] and [D4], only BERT-Res produced exact matches. However, our analysis confirmed that the prediction generated by ARI-32B, although differing from the ground truth "子孫" by producing "後孫", was semantically equivalent. The last example concerns the red tide phenomenon (Lee, 2017). In this case, ARI-32B demonstrated robustness in handling proper nouns; specifically, it was the only model to correctly restore the region names in [D1] and [D2]. Overall, these qualitative results suggest that ARI-32B outperforms the other models, particularly in restoring proper nouns such as personal and geographic names, by effectively leveraging external knowledge.

5.4 Impact of Temporal Domain Shift

To investigate the impact of the temporal gap between the retrieved and target documents on ARI's performance, we additionally performed a restora-

Damaged Text	吏曹判書**[D1][D2][D3]**初度呈辭. 入啓. **[D4]**由. (from JRS, 12/6/1654, King Hyojong)
BERT-Res	吏曹判書李壽恒初度呈辭. 入啓. 給由.
Gemini-2.5-Pro	吏曹判書沈之源初度呈辭. 入啓. 給由.
ARI-32B	吏曹判書李厚源初度呈辭. 入啓. 給由.
Damaged Text	上詣敬奉閣行禮, 仍御承文院, 誦<**[D1][D2]**><下泉>詩, 命三學士及諸忠**[D3][D4]**孫錄用. (from AJD, 10/6/1770, King Yeongjo)
BERT-Res	上詣敬奉閣行禮, 仍御承文院, 誦<詩鑑><下泉>詩, 命三學士及諸忠臣子孫錄用.
Gemini-2.5-Pro	上詣敬奉閣行禮, 仍御承文院, 誦<匪風><下泉>詩, 命三學士及諸忠臣之孫錄用.
ARI-32B	上詣敬奉閣行禮, 仍御承文院, 誦<匪風><下泉>詩, 命三學士及諸忠臣後孫錄用.
Damaged Text	全羅道監司書狀, 去九月間, **[D1][D2]**池水變赤色十餘日, 魚蝦浮出盡死事. (from AJD, 17/11/1588, King Seonjo)
BERT-Res	全羅道監司書狀, 去九月間, 大山池水變赤色十餘日, 魚蝦浮出盡死事.
Gemini-2.5-Pro	全羅道監司書狀, 去九月間, 海南池水變赤色十餘日, 魚蝦浮出盡死事.
ARI-32B	全羅道監司書狀, 去九月間, 光州池水變赤色十餘日, 魚蝦浮出盡死事.

Table 6: Comparison of restoration examples between our model and baseline models.

tion experiment on the *Goryeosa*[7], a historical chronicle of the Goryeo Dynasty, which predates our training dataset. Following the methodology of our survey, we constructed the test dataset from this corpus, employing the AJD and the JRS spanning the 14th to 19th centuries as reference corpora. As shown in Fig. 8, while ARI-32B remains robust under moderate temporal shifts, its performance declines as the temporal gap widens. This degradation is likely due to both diachronic linguistic changes and discrepancies in the period-specific external knowledge encapsulated within the reference corpora. Mitigating the impact of such domain mismatches remains an area for future research.

6 Conclusion

In this work, we introduced a RAG framework that enables LLMs to effectively utilize external knowledge for restoring damaged historical documents. Experimental results demonstrated consistent improvements in overall restoration performance, with especially notable gains in recovering named entities that depend on external knowledge. Furthermore, quantitative evaluation on a ground-truth test set and expert assessment on real-world damaged data showed that our model outperforms existing baselines and provides practical value for domain experts. Although developed for Korean historical archives, the proposed framework has broader potential as a general approach to low-resource ancient language restoration. By reducing the time and cost of restoring damaged historical documents, we expect this framework to help reveal previously obscured information and encourage the wider use of historical archives.

Limitations

In this study, we constructed training, valid, and test datasets by aligning them with statistical characteristics, such as the frequency of damaged characters, observed in real-world documents. We additionally validated our model's performance through human-expert evaluation employing actual damaged documents. Despite these efforts, a potential discrepancy may persist between our constructed datasets and the intrinsic characteristics of real-world damaged documents.

Due to computational constraints, we limited the maximum input length of the model to 4,096 tokens. Consequently, approximately 1.7% of data exceeding this length were filtered. In future work, we aim to extend the context window to accommodate longer documents and further improve performance. Beyond this, while our current approach leverages external knowledge and textual context, it relies solely on the text modality. Since the National Institute of Korean History provides original images for each document, incorporating these visual features into a multimodal restoration framework could significantly enhance performance.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2025-00555320) and the National Supercomputing Center with supercomputing resources including technical support (KSC-2024-CRE-0388).

[7] https://db.history.go.kr/goryeo

References

Anthropic. 2025. Introducing claude sonnet 4.5. https://www.anthropic.com/news/claude-sonnet-4-5. Accessed: 2025-01-03.

Yannis Assael, Thea Sommerschield, and Jonathan Prag. 2019. Restoring ancient text using deep learning: a case study on greek epigraphy. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6368–6375.

Yannis Assael, Thea Sommerschield, Brendan Shillingford, Mahyar Bordbar, John Pavlopoulos, Marita Chatzipanagiotou, Ion Androutsopoulos, Jonathan Prag, and Nando De Freitas. 2022. Restoring and attributing ancient texts using deep neural networks. *Nature*, 603(7900):280–283.

Chen-Chi Chang, Chong-Fu Li, Chu-Hsuan Lee, and Hung-Shin Lee. 2025. Enhancing low-resource minority language translation with llms and retrieval-augmented generation for cultural nuances. In *Intelligent Systems Conference*, pages 190–204. Springer.

Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, Yifeng Lu, and 1 others. 2023. Symbolic discovery of optimization algorithms. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 49205–49233.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv:2507.06261*.

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.

Miguel Domingo and Francisco Casacuberta Nolla. 2018. Spelling normalization of historical documents by using a machine translation approach. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation: 28-30 May 2018, Universitat d'Alacant, Alacant, Spain*, pages 129–137. European Association for Machine Translation.

Kyeongpil Kang, Kyohoon Jin, Soyoung Yang, Soojin Jang, Jaegul Choo, and Youngbin Kim. 2021. Restoring and mining the records of the joseon dynasty via neural language modeling and machine translation. In *2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021*, pages 4031–4042. Association for Computational Linguistics (ACL).

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Jinhyuk Lee, Feiyang Chen, Sahil Dua, Daniel Cer, Madhuri Shanbhogue, Iftekhar Naim, Gustavo Hernández Ábrego, Zhe Li, Kaifeng Chen, Henrique Schechter Vera, and 1 others. 2025. Gemini embedding: Generalizable embeddings from gemini. *CoRR*.

Sankyun Lee. 2017. Occurrence of and countermeasure against red tide in joseon dynasty. *The Korean Historical Association*, pages 75–108.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Kiara MH Liu, Martin Mueller, and Matthew Wilkens. 2025. Zero-shot methods for historical text restoration. *Anthology of Computers and the Humanities*, 3:984–995.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv:1907.11692*.

Xing Han Lù. 2024. Bm25s: Orders of magnitude faster lexical search via eager sparse scoring. *arXiv:2407.03618*.

Ercong Nie, Sheng Liang, Helmut Schmid, and Hinrich Schütze. 2023. Cross-lingual retrieval augmented prompt for low-resource languages. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8320–8340.

OpenAI. 2025. Gpt-5.1: A smarter, more conversational chatgpt. https://openai.com/index/gpt-5-1/. Accessed: 2026-01-03.

Gongbo Tang, Fabienne Cap, Eva Pettersson, and Joakim Nivre. 2018. An evaluation of neural machine translation models on historical spelling normalization. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1320–1331.

Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, and 1 others. 2025. Kimi k2: Open agentic intelligence. *arXiv:2507.20534*.

Changhan Wang, Kyunghyun Cho, and Jiatao Gu. 2020. Neural machine translation with byte-level subwords. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9154–9160.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, and 1 others. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547.

Tingyu Xie, Jian Zhang, Yan Zhang, Yuanyuan Liang, Qi Li, and Hongwei Wang. 2025. Retrieval augmented instruction tuning for open ner with large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2904–2918.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv:2505.09388*.

Tianjun Zhang, Shishir G Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E Gonzalez. 2024. Raft: Adapting language model to domain specific rag. In *First Conference on Language Modeling*.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, and 1 others. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv:2506.05176*.

Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, and 1 others. 2023. Pytorch fsdp: Experiences on scaling fully sharded data parallel. *Proceedings of the VLDB Endowment*, 16(12):3848–3860.

$D_{\text{Rand}} : D_{\text{NE}}$	Acc. on NE	Acc. on Rand
100% : 0%	42.75	75.67
75% : 25%	45.41	76.13
50% : 50%	45.06	75.12
25% : 75%	45.89	74.58
0% : 100%	45.25	69.68

Table 7: Performance comparison across different mixing ratios for constructing the training dataset.

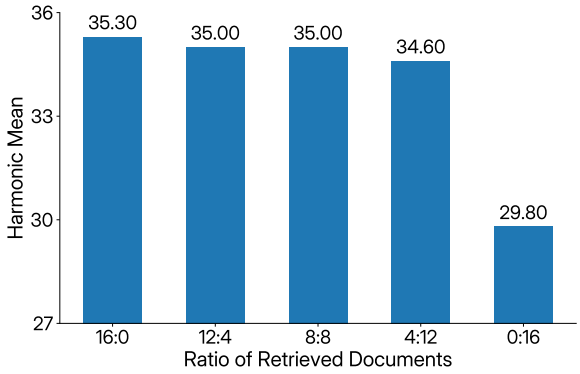


Figure 9: Restoration performance across different BM25-to-embedding retrieval ratios. The y-axis represents the harmonic mean of the scores on D_{NE} and D_{Rand} for each configuration.

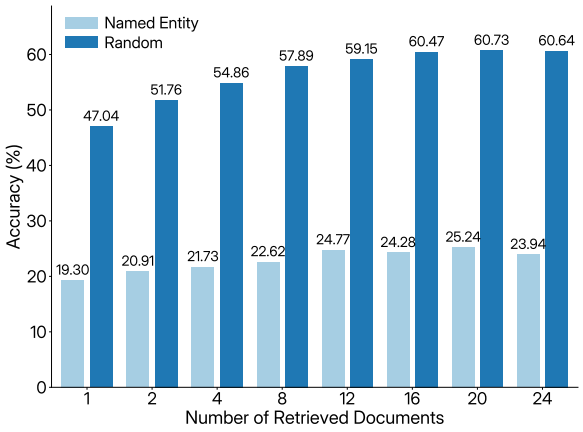


Figure 10: Performance comparison across different numbers of retrieved documents.

A Training Dataset Construction and Training Settings

In this study, we conducted a comparative experiment on training data mixing ratios to examine how each dataset affects model performance. The total size of the training set was fixed at 100K samples. Each configuration comprised a mixture of D_{NE} and D_{Rand}, with the proportion of D_{NE} varied across settings. As shown in Table 7, increasing the proportion of D_{NE} consistently improved restoration performance on named entities, while random character restoration performance showed a downward trend. Considering this trade-off, the setting with 25% D_{NE} achieved the most balanced performance across the two metrics. Based on this result, we constructed the final training dataset using this ratio.

We trained ARI-32B on eight HGX H200 GPUs utilizing FSDP (Zhao et al., 2023) and the Lion-8Bit optimizer (Chen et al., 2023). We employed a cosine learning rate schedule with a peak learning rate of 6×10^{-6} and a warmup ratio of 0.05. Training was performed for two epochs with a global batch size of 64 and a maximum sequence length of 4,096 tokens, totaling 16.06 billion training tokens. ARI-8B was trained under the same settings, except that it used four HGX H200 GPUs and a global batch size of 48. For the BERT-Res baseline, we used ModernBERT-large and trained it on two RTX A6000 GPUs. We again used a cosine learning rate schedule, with a learning rate of 1×10^{-4} and a warmup ratio of 0.05. The model was trained for 10 epochs with a global batch size of 1,024 and a maximum sequence length of 2,048 tokens, requiring approximately 12 hours.

B Experiments

B.1 Analysis of Retrieval Strategies

We further evaluated a hybrid RAG-based approach to determine the optimal RAG strategy for Hanja document restoration. We varied the composition of the 16 retrieved documents between BM25 and Gemini-Embedding. As shown in Fig. 9, the BM25-only configuration (i.e., a 16:0 ratio) achieved the best performance, yielding a harmonic mean of 35.3%. This suggests that, given the character-level nature of the restoration task, lexical similarity plays a more important role than semantic similarity. Furthermore, we investigated a hybrid RAG strategy that additionally incorporated re-ranking using Qwen3-Reranker-8B (Zhang et al., 2025). This configuration achieved a D_{NE} score of 37.70%, which was lower than the 38.58% obtained by the original BM25-based RAG pipeline. This result indicates that re-rankers pretrained on modern corpora have limitations in reordering retrieved Hanja documents for restoration tasks. Accordingly, we adopted a BM25-only strategy, prioritizing both restoration quality and architectural simplicity.

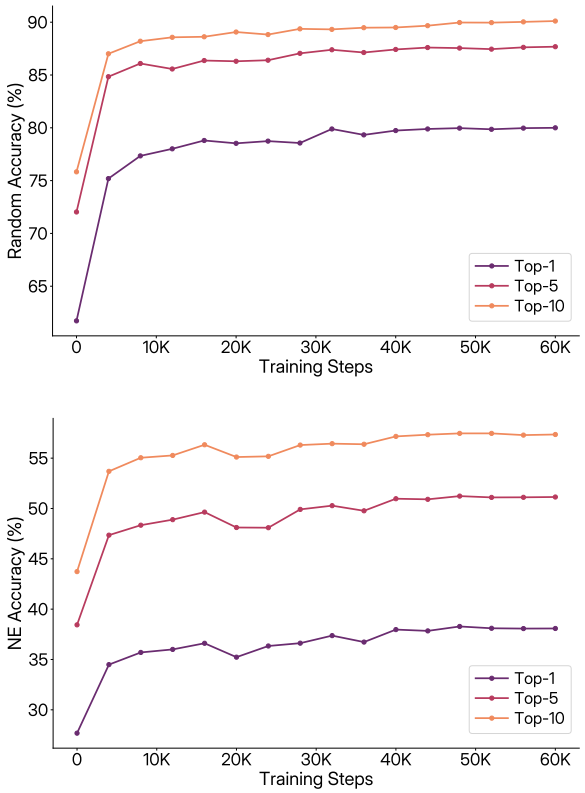


Figure 11: Fine-tuning accuracy of ARI-32B on the D_{Rand} (top) and D_{NE} (bottom) datasets.

After determining the retrieval strategy, we further investigated the effect of varying the number of retrieved documents. Fig. 10 illustrates the performance trends with respect to the number of retrieved documents. We find that, although performance generally improves on both D_{NE} and D_{Rand} as more documents are retrieved, it shows a slight decline when the number reaches 24. To balance computational efficiency and performance stability, we configure the ARI models to use 20 retrieved documents as few-shot examples.

B.2 Top-k Accuracy during Training

Fig. 11 illustrates the Top-1, Top-5, and Top-10 accuracies of ARI-32B measured on D_{Rand} and D_{NE}, respectively, at each training step. These results demonstrate that the model's restoration performance improves as training progresses.

C Byte-level Constrained Decoding for Top-K Candidate Selection

With a Byte-level BPE tokenizer (Wang et al., 2020), frequent Chinese characters are typically mapped to single tokens. In contrast, Hanja characters, often absent from the tokenizer's vocabulary,

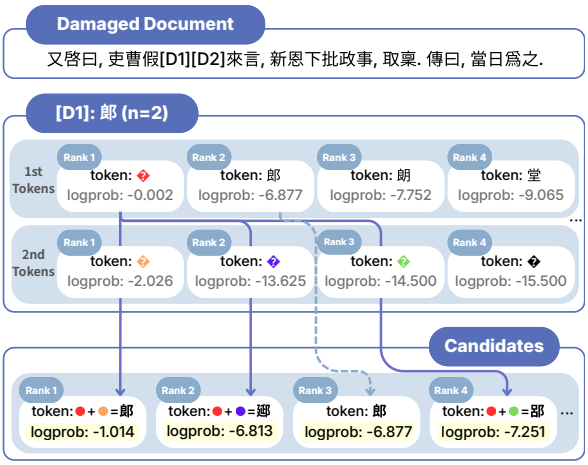


Figure 12: Illustration of Top-K candidate selection via byte-level constrained decoding.

are decomposed into sequences of multiple byte tokens (typically 2–3).

To address this fragmentation during the restoration of the i-th damaged character (illustrated in Fig. 12), we employ a constrained decoding approach. Once an initial byte token is generated, the model continues generating tokens until a valid character sequence is formed. Candidate scores are then calculated by averaging the log probabilities of their constituent byte tokens, and the top-K options are presented to the user. By default, the generation of the subsequent $(i+1)$-th character is conditioned on the top-1 prediction. However, if the user manually selects an alternative candidate (e.g., the c-th option), the context is updated to reflect this selection. We leverage the vLLM (Kwon et al., 2023) framework to efficiently extract these top-K candidates.

D Expert Evaluation

To facilitate efficient expert evaluation, we implemented a web interface, as shown in Fig. 13. The evaluation process consists of two questions for each real-world damaged document. First, evaluators are asked to select all valid answers from the candidates generated by different models (allowing for multiple selections), based on contextual coherence and factual validity. Subsequently, the second question investigates model preference; evaluators select the single most suitable model by comprehensively considering overall restoration quality and its utility as a collaboration tool.

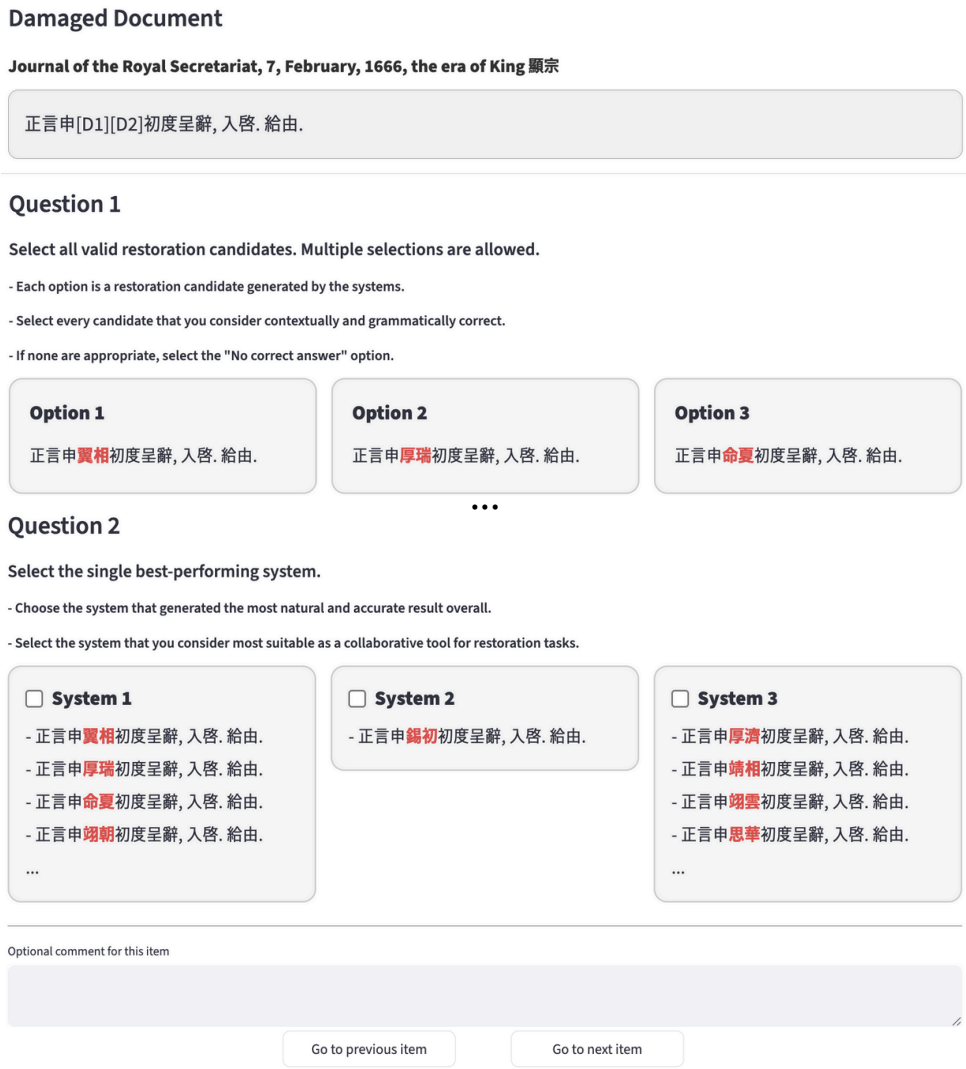


Figure 13: Web interface for expert evaluation.

E Prompt Format

Fig. 14 presents the final prompt format, based on the experimental results in Section 4.2. The system prompt comprises four main components: Task, which defines the model's objective; Requirements, which outlines restoration guidelines tailored to the damaged document's context; Input & Output, which dictates the interaction format; and Example Input & Output, which provides few-shot examples. The user prompt reiterates basic instructions and supplies the target damaged document for restoration. Additionally, to enhance accuracy, we incorporate external knowledge comprising the document's corresponding date and 20 related documents retrieved from the training dataset.

System Prompt

```
### Task
You are an expert in restoring damaged Hanja characters. Restore each [Dn] with exactly the
original Hanja character. Each [Dn] corresponds to exactly one Hanja character.

### Requirements
Base your restoration on the document's overall context and meaning rather than treating each
damaged token in isolation.

### Input & Output
The input consists of the document itself, its metadata, and the related documents.
While the related documents are omitted in the shot for conciseness, they are always present in
the actual dataset.
The output must follow the format: {"[Dn]": "the restored Hanja character for [Dn]"}

### Example Input & Output
[Example 1]
**Input**
The document was written at date: 7, month: 6, year: 1771 - 英祖 era.
Input Document: "傳于[D1]興宗日, 當自光明殿出, 承旨·侍衛, 來待于建禮門."
**Output**
{"[D1]":"李"}

[Example 2]
**Input**
The document was written at date: 26, month: 8, year: 1824 - 純祖 era.
Input Document: "[D1][D2]口傳政事, 副護軍單李紀淵."
**Output**
{"[D1]":"兵","[D2]":"曹"}

[Example 3]
**Input**
The document was written at date: 30, month: 6, year: 1686 - 肅宗 era.
Input Document: "府[D1]啓, 請[D2]禮·壽進·於[D3]·龍洞·明安公主房折受處[D4]查正事. 入啓."
**Output**
{"[D1]": "前", "[D2]": "明", "[D3]": "義", "[D4]": "一"}

...
```

User Prompt

```
Use the following documents as references to accurately restore the input document.

Related Documents:

- 傳曰, 呈告工判·同敦寧許遞, 今日政差出.
- 傳曰, 在外敦寧都正·同敦寧, 許遞, 今日政差出.
- 傳曰, 呈告禮曹判書·同敦寧·兵曹參判·同成均許遞, 今日政差出.
- 傳于金敬均曰, 同敦寧許遞, 今日政差出.
- 傳于趙秉翊曰, 同敦寧·敦寧都正·同經筵許遞, 今日政差出.
...

The input document was written at date: 28, month: 7, year: 1887 - 高宗 era.
Input Document: 傳曰, 同敦寧·[D1][D2]都正竝許遞, 今日政差出.
```

Figure 14: An example of the final prompt used for the restoration task.