# HERIT: Democratizing Global Access to Korean Historical Archives via RAG-based Data Augmentation

**Anonymous ACL submission**

## Abstract

Historical archives are invaluable resources for multidisciplinary research but remain inaccessible to the global community due to language barriers. While manual translation is prohibitively expensive and time-consuming, the direct application of existing machine translation models is often inadequate due to the unique linguistic and historical nuances of these documents. To address these challenges, we propose a novel framework that leverages Retrieval-Augmented Generation (RAG) to generate high-quality pseudo-labeled data from abundant Hanja-Korean corpora. This approach expands the training dataset, effectively mitigating data scarcity and temporal overfitting observed in human-labeled corpora. Extensive evaluations demonstrate that HERIT significantly outperforms baseline models. Finally, we employ our model to translate previously untranslated portions of the archives, aiming to democratize access to these resources for researchers worldwide.

## 1 Introduction

Historical documents have long been a vital source for discovering knowledge and deriving insights. In the humanities, these archives allow researchers to analyze political dynamics and cultural evolution. Furthermore, they serve as crucial data sources in the natural sciences for tracking historical events, including astronomical observations, such as solar eclipses and supernovae, as well as seismic activities and meteorological patterns.

The Annals of the Joseon Dynasty (**AJD**) and the Journal of the Royal Secretariat (**JRS**) are regarded as preeminent historical archives, recognized not only in Korea but globally. These archives provide comprehensive record spanning over 500 years, comprising a vast array of subjects. Extending beyond state affairs and administration, they meticulously chronicle natural phenomena, including astronomical observations, seismic activity, famines, and epidemics. Given this extensive scope, researchers worldwide actively utilize these archives across disciplines, ranging from the social sciences to natural sciences.

Given their significance, the translation of the AJD from Hanja into Korean has been completed, while the translation of the JRS is currently underway. Furthermore, to promote the global accessibility of these historical archives, the National Institute of Korean History (NIKH) has launched a project to translate the AJD into English. However, relying on manual translation by experts proficient in both Hanja and English presents significant challenges; the project is projected to span 22 years with a budget of approximately 20 million USD[1]. Furthermore, applying this manual approach to other archives such as the JRS, which is four to five times more voluminous than the AJD, would require a prohibitive time and costs. Consequently, existing human-translated English corpora are not only scarce but also heavily skewed towards specific reigns (e.g., King Sejong). Improving translation on unseen historical periods requires a method that generalizes beyond this temporal bias.

Given that historical archives in Korea, such as AJD and JRS, are written in Hanja, leveraging LLMs specialized in Chinese-English machine translation appears to be an intuitive solution for processing these corpora. However, such an approach encounters significant challenges. First, there is a fundamental script mismatch: Chinese models typically utilize Simplified Chinese, whereas Korean Hanja retains Traditional Chinese forms, resulting in character incompatibility. Second, substantial semantic and syntactic discrepancies exist. Even when characters are identical, semantic divergence is common as homographs have

[1] https://sillok.history.go.kr/eslk/about/translationProjectInfo.do

often evolved distinct meanings in the respective languages. Furthermore, the texts contain Korea-specific Hanja usage that reflects native grammatical structures, and the word order frequently deviates from standard Chinese syntax. Most critically, Chinese-English models transliterate proper nouns, such as personal names and geographical locations, based on Chinese pronunciation rather than Korean pronunciation. This misalignment not only degrades translation quality but also introduces factual errors, rendering the models unsuitable for rigorous historical analysis.

To address these challenges, we propose a novel framework for training a Hanja-to-English model specialized for historical archives, utilizing human-translated corpora as illustrated in Fig. 1. To overcome the scarcity and temporal overfitting of available expert data, we introduce a RAG-based data augmentation strategy that generates high-quality pseudo-labeled data. This method effectively alleviates data sparsity while minimizing overfitting to specific eras. Subsequently, we present **HERIT**: a **H**anja-**E**nglish **R**AG-based **I**ntelligent **T**ranslation model for historical archives. Extensive evaluations, including quantitative metrics and human expert assessments, demonstrate that HERIT significantly outperforms existing LLMs in historical text translation. In addition, we translate the remaining untranslated documents in the JRS and AJD using our model. We will make the results including the translated corpus publicly available upon publication to facilitate future research.

## 2 Related Work

Historical archives have extensively contributed to diverse research domains. In the field of social sciences, they provide foundational data for analyzing international affairs (Thies, 2002), demography (Alter, 2019), economic condition (Clark, 2007), and linguistic transitions (Wei et al., 2025). In parallel, the natural sciences leverage historical records to investigate astronomical phenomena (Wang et al., 2021; Wei and Yan, 2024), climatic change (Chen et al., 2020; Zhang et al., 2021). Thus, historical documents serve as invaluable assets across diverse fields, offering a lens through which to understand past events and derive critical insights for both the present and the future.

Korean archives, such as AJD and JRS, are recognized as invaluable resources by the global academic community for their comprehensive scope, spanning multiple centuries and covering a wide array of subjects. Notably, the continuity of these records makes them particularly advantageous for analyzing macroscopic changes. For instance, they have been extensively utilized in studies concerning astronomical observations and geological phenomena, as well as disasters including epidemics and economic crises (Stephenson, 2011; Yang et al., 2012; Kim et al., 2017; Wang et al., 2021; Wei and Yan, 2024; Rhee, 2014; Hwang, 2022).

Research is currently underway to translate historical archives into modern languages to utilize them effectively. Afli and Way (2016) proposed a method to digitize historical documents written in 17th-century French using OCR and translate them into Modern English by applying stochastic machine translation techniques. To translate Ancient Chinese into Modern Chinese, Zhang et al. (2019); Liu et al. (2019) developed passage-level alignment methods and deep neural network-based machine translation models.

Among Korean historical documents, the AJD has been fully translated into Korean by the NIKH, whereas the translation of the JRS remains ongoing due to its vast volume. Addressing this, Kang et al. (2021) developed a Hanja-to-Korean translation model leveraging the parallel corpus derived from both the AJD and JRS. Concurrently, a project to translate the original Hanja text of the AJD into English is underway, with the records from the King Sejong era (approximately 2.5% of the total) currently completed. Using this partial Hanja-English corpus alongside the Hanja-Korean data, Son et al. (2022) proposed a Hanja-Korean-English multilingual model. However, due to the scarcity of human-translated Hanja-English data, the performance of models trained from scratch remains limited.

Recent 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), have demonstrated enhanced in-context learning capabilities. This advancement enables them to utilize information during the inference phase via few-shot learning, even without having encountered it during the pre-training or fine-tuning stages (Brown et al., 2020). Specifically, RAG further improves performance by incorporating retrieved examples relevant to the input into the prompt (Lewis et al., 2020). Consequently, there is active research into data augmentation approaches for data-scarce domains that leverage these few-shot learning and RAG strate-

| Hanja | Korean | English |
|---|---|---|
| 疫疾 | 역질 | Epidemic disease |
| 世宗 | 세종 | King Sejong |
| 江原道 | 강원도 | Kangwon Province |

Table 1: Examples of glossary information: Hanja source, Korean pronunciation, and English translation.

gies (Seo et al., 2024; Song et al., 2024). Moreover, methods such as Best-of-N (BoN) (Liu et al., 2024) and Fusion-of-N (FusioN) (Khairi et al., 2025) have been proposed to further enhance performance by generating multiple candidate responses and aggregating them to derive a final answer.

We address the scarcity of historical parallel corpora by leveraging RAG-based augmentation. Specifically, we construct a high-quality pseudo-labeled dataset using Hanja-Korean corpora via RAG and FusioN. Finally, we fine-tune a LLM, leveraging its inherent English proficiency and partial Hanja understanding.

## 3 Datasets

As noted in Section 2, the translation of the AJD is ongoing; currently, only the Annals of King Sejong has been fully translated. For our experiments, we obtained the AJD documents from NIKH[2].

Subsequently, we curated a dataset from existing aforementioned human-translated documents, comprising 350K Hanja-Korean pairs and 17.7K Hanja-Korean-English triples[3]. From the Hanja-English corpus, we allocated 1,000 pairs each to the validation set $D^{valid}_{\text{human}}$ and the test set $D^{test}_{\text{human}}$, while the remaining 15.7K pairs constitute the training set $D^{train}_{\text{human}}$. Furthermore, to mitigate period-specific bias in the evaluation, we compiled an additional test dataset $D^{test}_{NT}$ of 2,080 entries by sampling 80 records for each king from the untranslated corpus. Crucially, $D^{test}_{NT}$ is strictly excluded from both the pseudo-labeling process and the training data. For the remaining pairs, we applied pseudo-labeling by utilizing the Hanja-English training set as an external knowledge base for RAG. Further details on the pseudo-labeling approach are provided in Section 4. We constructed an augmented dataset, $D^{train}_{\text{aug}}$, comprising 315K pairs from all reigns other than King Sejong's, after excluding documents longer than 800 characters. Finally, both $D^{train}_{\text{human}}$ and $D^{train}_{\text{aug}}$ were utilized for model training.

In addition, NIKH provides lexical information, including meanings and pronunciations, for proper nouns such as personal and geographical names. Employing this data enables the LLM to better comprehend Hanja documents and accurately translate proper nouns into English based on their correct pronunciation. Consequently, we collected these lexicons and compiled a glossary consisting of 25K entries and utilized it during pseudo-labeling.

## 4 Proposed Methods

In this section, we first evaluate the Hanja-to-English translation performance of existing LLMs. Next, we describe a method to improve translation performance through pseudo-labeling, which leverages both partially translated and untranslated data to address data sparsity. Finally, we introduce HERIT, a model specialized for Hanja-to-English translation, which is fine-tuned on both the pseudo-labeled dataset and human-expert translations.

### 4.1 Baseline Performance of LLMs in Hanja-to-English Translation

We evaluated the translation performance of LLMs on $D^{test}_{\text{human}}$ using base prompts with simple instructions. We employed standard metrics for translation tasks, including SacreBLEU (Post, 2018), METEOR (Banerjee and Lavie, 2005), ROUGE-L (LIN, 2004), and chrF++ (Popović, 2017).

As shown in Table 2, Hunyuan-MT-7B, despite being one of the state-of-the-art models for Chinese translation (Zheng et al., 2025), exhibits suboptimal performance in Hanja translation. Similarly, Exaone-4.0 (LG AI Research, 2025), an LLM specialized in both English and Korean, exhibits suboptimal performance. This suggests that LLMs specialized in either Chinese or Korean alone are not effective substitutes for dedicated Hanja translation. Results from the Qwen3 families demonstrate that translation quality generally improves as underlying model capabilities increase. Regarding reasoning capabilities, while Gemini-2.5-Flash shows performance gains, Sonnet-4.5 experiences a slight decline. This aligns with prior survey suggesting that reasoning tokens do not significantly contribute to machine translation performance (Zebaze et al., 2025). A similar trend is observed with Gemini-2.5-Pro as shown in Table 4.

Although Gemini-2.5-Pro achieves the highest

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

[3] Note that as all documents in the AJD were originally translated from Hanja to Korean, the Hanja-English data naturally forms (Hanja, Korean, English) triples.

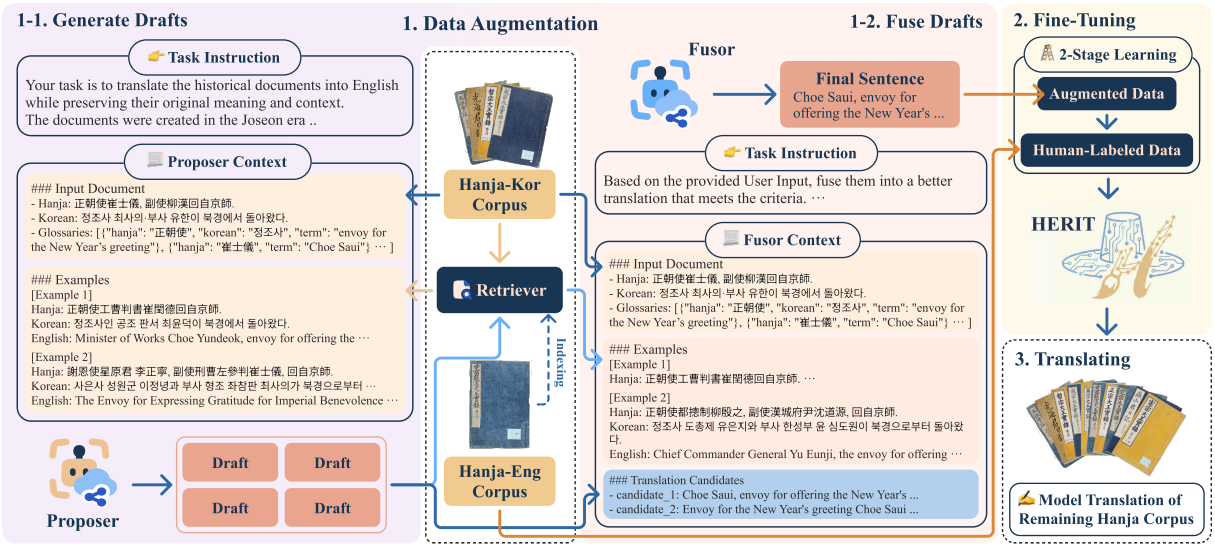


Figure 1: Overview of the proposed framework for translating historical documents into English. The framework comprises three main stages: 1) pseudo-labeling for data augmentation, 2) fine-tuning the LLM by integrating human-labeled and pseudo-labeled data, and 3) translation of the untranslated corpus using the trained model.

| Model | BLEU | chrF++ | Cost $ |
|---|---|---|---|
| Hunyuan-MT-7B | 8.082 | 34.237 | - |
| Exaone-4.0-32B | 8.961 | 35.538 | - |
| Qwen3 8B | 10.438 | 38.166 | - |
| Qwen3 32B | 11.857 | 40.677 | - |
| GPT-5.1 | 13.384 | 43.359 | 2.09 |
| Kimi-K2 (R) | 15.512 | 43.373 | 6.14 |
| Gemini-2.5-Flash | 15.777 | 44.342 | 0.46 |
| Gemini-2.5-Flash (R) | 17.245 | 46.213 | 3.34 |
| Sonnet-4.5 | 17.934 | 48.378 | 5.88 |
| Sonnet-4.5 (R) | 17.249 | 47.767 | 12.25 |
| Gemini-2.5-Pro (R) | 20.088 | 49.660 | 19.82 |

Table 2: Translation performance on $D_{\text{human}}^{test}$ using basic prompts and the total API cost. (R) denotes the integration of reasoning capabilities.

accuracy, it incurs higher inference costs due to its size and reasoning overhead. In contrast, Gemini-2.5-Flash offers the most efficient trade-off between cost and performance. Therefore, we employ Gemini-2.5-Flash for pseudo-labeling.

### 4.2 Building a Pseudo-label Dataset for Hanja-English Translation

To improve translation performance, we construct a high-quality pseudo-labeled dataset by adopting RAG and FusioN (Khairi et al., 2025) approaches, following the process shown in Step 1 of Fig. 1. First, in Step 1-1, the Proposer takes untranslated documents as input and generates translation drafts. In Step 1-2, the Fusor generates the final translation by receiving both the generated drafts and the target document used in the Proposer as inputs. Since this architecture generates the final sentence by ensembling the Proposer's drafts in the Fusor, it not only resolves mistranslation but also enhances translation quality by leveraging expanded information through a holistic examination of the drafts.

As depicted in Fig. 1, we aim to improve the pseudo-label quality of the Proposer and Fusor by enriching the input prompt with auxiliary information, instead of relying solely on the Hanja document. To address the potential limited proficiency of LLMs in Hanja, we augment the input with the corresponding Korean text to enhance contextual understanding. To ensure the accurate translation of proper nouns, we utilize a curated glossary by appending terms to the prompt when exact matches are found in the source text. We also utilize few-shot prompting to capture the nuances, including contextual and lexical information, of Hanja-to-English translation. Specifically, we adopt RAG that retrieves the top-$K$ most similar documents $\mathcal{S}(h)$ from $D_{\text{human}}^{train}$ for a given input Hanja document $h$ via an embedding-based retriever. These retrieved entries are fed into both the Proposer and the Fusor as few-shot demonstrations in the form of (Hanja, Korean, English) triples.

We optimized the hyperparameters for pseudo-labeling based on the experiments described in Section 5.1. To balance performance and inference

cost, we selected Gemini-2.5-Flash as the backbone for both the Proposer and Fusor, and employed Gemini-Embedding (Lee et al., 2025) as the retriever. The retriever used the Korean component of the input document as the query for the Proposer, while utilizing both the input Hanja and the English translation drafts as the query for the Fusor. We employed 64 and 12 few-shot examples for the Fusor and the Proposer, respectively. During pseudo-labeling, the Fusor synthesized the final translation from five drafts generated by the Proposer with a temperature of 0.3.

Finally, to construct the pseudo-labeled dataset $D_{\text{aug}}^{train}$, we applied the proposed augmentation strategy to 317.4K Hanja-Korean pairs from corpora not yet translated into English (excluding data from the King Sejong era), incurring a cost of about $3,000. This augmented dataset is utilized to mitigate data sparsity and prevent potential overfitting toward a specific historical period.

### 4.3 Model Training and Translation of the Untranslated Corpus

We present HERIT, a Hanja-to-English historical document translation model fine-tuned on Qwen3-32B. HERIT performs end-to-end translation from Hanja to English. Our training utilizes two datasets: a pseudo-labeled dataset $D_{\text{aug}}^{train}$ and a human-translated dataset $D_{\text{human}}^{train}$. We employ a two-stage training strategy: the first stage uses $D_{\text{aug}}^{train}$ to learn Hanja semantics and phonetic patterns of proper nouns, while the second stage leverages $D_{\text{human}}^{train}$ to align the model with human-level stylistic nuances. The entire training process was executed on eight H200 GPUs, taking 15.1 hours for the first stage and 1.2 hours for the second stage.

Finally, we translated the entire set of untranslated AJD and JRS documents using our model. Using eight H200 GPUs, the translation of these 2.09M documents was completed in 94.6 hours, which is significantly faster than manual translation. To facilitate global accessibility to these historical records, we will make both the resulting dataset and our model publicly available upon publication.

## 5 Experimental Results

This section investigates optimal pseudo-labeling settings and demonstrates our model's superior performance against baselines using reference-based, LLM-based, and human expert evaluations.

### 5.1 Effectiveness of Pseudo-Labeling

To comprehensively validate the effectiveness of glossary integration, input augmentation, and few-shot prompting strategies, we evaluated translation performance on $D_{\text{human}}^{test}$ using Gemini-2.5-Flash. All experimental results with costs on Batch API are summarized in Table 3.

First, comparing the baseline with the initial configurations, we observe a significant improvement in translation performance when a glossary is incorporated. This suggests that the glossary information facilitates a better understanding of Hanja terms, aiding in their translation into appropriate English pronunciations. Similarly, providing pre-translated Korean inputs alongside the Hanja source text enhances the understanding of the input documents, indicating that Korean translations effectively support the pseudo-labeling process.

Furthermore, the latter part of the table presents ablation studies examining the impact of the RAG and FusioN modules, as well as the retriever input language. Including few-shot examples in the prompt contributes substantially to performance gains. Notably, while fixed examples (static shots) unrelated to the input text offer some benefit, dynamic shots relevant to the input document lead to superior improvements. This demonstrates that RAG-based approaches are more effective for performance enhancement than naive few-shot learning. When employing dynamic shots, using the Korean translation of the corresponding Hanja sentence as the retriever input yielded better performance than using the Hanja sentence itself. This is likely because the embedding model used for retrieval was trained on a larger corpus of Korean text compared to Hanja, resulting in superior comprehension of Korean. Using both Hanja and Korean simultaneously as input did not lead to significant additional improvements.

The proposed method utilizing FusioN outperforms standard single-stage translation pipelines. Experimental results with the number of few-shots set to 12 and drafts set to 4 indicate that concatenating the Hanja input with the English draft sentences as query inputs for the retriever is effective when finding relevant documents for the Fusor. This suggests that in the Fusor, which refines drafts to generate the final translation, the contexts of both the target Hanja sentence and the English candidate sentences are crucial. Additionally, we conducted experiments on the number of few-shots provided

| | | BLEU | METEOR | chrF++ | Cost $ |
|---|---|---|---|---|---|
| Baseline | | 15.777 | 0.430 | 44.342 | 0.232 |
| + Glossary | | 24.732 | 0.518 | 52.489 | 0.257 |
| + Korean | | 27.225 | 0.539 | 54.172 | 0.278 |
| + RAG ($K$: 12) | | | | | |
| • Static shots | | 34.264 | 0.593 | 58.846 | 1.070 |
| • Dynamic shots (Hanja) | | 39.883 | 0.631 | 62.852 | 1.216 |
| • Dynamic shots (Kor) | | **40.310** | **0.633** | 63.146 | 1.367 |
| • Dynamic shots (Hanja + Kor) | | 40.008 | **0.633** | **63.279** | 1.387 |
| RAG (Dynamic Shots, Kor, $K$: 64) | | 41.763 | 0.644 | 64.138 | 6.157 |
| + FusioN ($K$: 12, $D$: 4) | | | | | |
| Proposer (Kor) | Fusor (Kor) | 40.609 | 0.641 | 63.965 | 2.959 |
| Proposer (Kor) | Fusor (Eng) | 40.722 | 0.641 | 63.956 | 2.699 |
| Proposer (Kor) | Fusor (Hanja) | 40.825 | 0.641 | 64.060 | 2.767 |
| Proposer (Kor) | Fusor (Hanja + Eng) | **40.925** | **0.642** | **64.047** | 2.881 |
| Proposer (Kor) | Fusor (Hanja + Kor + Eng) | 40.786 | 0.641 | 64.040 | 3.060 |
| $K_{\text{Proposer}}$: 64, $K_{\text{Fusor}}$: 12, $D$: 5 | | | | | |
| Proposer (Kor) | Fusor (Hanja + Eng) | **42.293** | **0.652** | **65.025** | 7.698 |

Table 3: Ablation study of pseudo-labeling. The language in parentheses denotes the query language used for retrieval, while $K$ and $D$ represent the number of few-shot examples and translation drafts, respectively.

to the Proposer and Fusor, the results of which are shown in Fig. 5. Based on these results, we set the number of drafts to 5 and few-shot examples for the Proposer and Fusor to 64 and 12, respectively.

As indicated in the bottom row of Table 3, the final pseudo-labeling performance achieved a BLEU score of 42.293, outperforms other baselines including the single translation approach that employs 64 dynamic few-shot examples, consistent with the proposer. Note that this performance does not reflect the direct translation of Hanja-only documents into English. Instead, it represents the pseudo-labeling performance facilitated by auxiliary information, such as Korean translations, glossary data, and retrieved references from $D^{train}_{\text{human}}$.

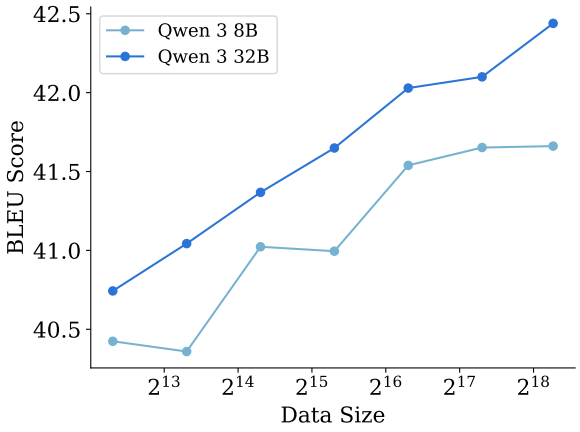


Figure 2: Model translation performance with respect to pseudo-labeling data size.

## 5.2 Evaluation on Ground-Truth Data

To verify the effectiveness of training with our augmented data, we evaluated translation performance using an expert-translated test dataset, $D^{test}_{\text{human}}$.

First, we evaluate the model performance with respect to the size of $D^{train}_{\text{aug}}$ used in the first training stage, as shown in Figure 2. We observe that translation performance improves as the volume of pseudo-labeled data increases, demonstrating the effectiveness of our proposed approach.

Furthermore, we compared models fine-tuned exclusively on $D^{train}_{\text{human}}$ against those trained on both $D^{train}_{\text{human}}$ and $D^{train}_{\text{aug}}$. To isolate the effects of fine-tuning, we employed Gemini-2.5 Flash and Pro models utilizing glossary information and RAG as baselines. As shown in Table 4, given sufficient glossary context and relevant documents, Gemini-2.5-Pro demonstrates no significant performance advantage over the Gemini-2.5-Flash.

A comparison between Qwen3 8B and 32B confirms that translation quality scales with model size. Furthermore, the Qwen3 8B model trained on $D^{train}_{\text{human}}$ outperforms Gemini-2.5-Pro equipped

| Model | BLEU | METEOR | chrF++ |
|---|---|---|---|
| Baseline (+ Glossary & RAG) | | | |
| Gemini-2.5-Flash | 39.134 | 0.626 | 62.552 |
| Gemini-2.5-Pro | 38.602 | 0.625 | 62.643 |
| Fined-tuned on $D_{\text{aug}}^{train}$ | | | |
| Qwen3 8B | 38.690 | 0.624 | 62.326 |
| Qwen3 32B | 38.827 | 0.624 | 62.411 |
| Fined-tuned on $D_{\text{human}}^{train}$ | | | |
| Qwen3 8B | 40.317 | 0.634 | 63.589 |
| Qwen3 32B | 40.885 | 0.640 | 64.099 |
| Mixed training: $D_{\text{aug}}^{train} \cup D_{\text{human}}^{train}$ | | | |
| Qwen3 8B | 40.813 | 0.640 | 63.840 |
| Qwen3 32B | 40.880 | 0.641 | 63.887 |
| 2-Stage learning: $D_{\text{aug}}^{train} \rightarrow D_{\text{human}}^{train}$ | | | |
| Qwen3 8B | **41.661** | **0.649** | **64.647** |
| Qwen3 32B | **42.439** | **0.653** | **65.060** |

Table 4: Translation performance of baseline models and fine-tuned LLMs on the test dataset $D_{\text{human}}^{test}$.

| $D_{\text{human}}^{train}$ | Score | $D_{\text{aug}}^{train} \rightarrow D_{\text{human}}^{train}$ | Score |
|---|---|---|---|
| Qwen3 8B | 6.531 | Qwen3 8B | **7.938** |
| Qwen3 32B | 7.071 | Qwen3 32B | **8.043** |

Table 5: Evaluation results using LLM-as-a-judge on the held-out test dataset $D_{\text{NT}}^{test}$.

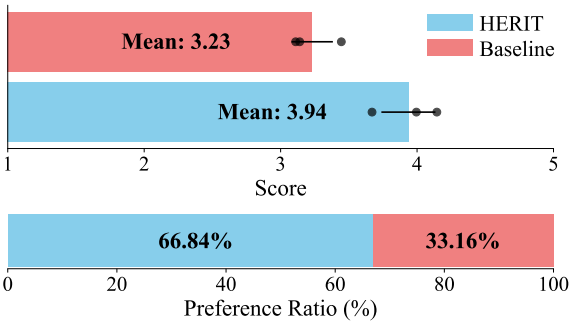


Figure 3: Human expert evaluation scores and preference ratios for our model versus the baseline.

with glossary information and RAG, demonstrating the effectiveness of fine-tuning for historical documents. While the model trained solely on $D_{\text{human}}^{train}$ achieves higher performance than the one trained only on $D_{\text{aug}}^{train}$, it is noteworthy that the Qwen3 32B model trained exclusively on $D_{\text{aug}}^{train}$ still surpasses the performance of Gemini-2.5. Most importantly, incorporating our constructed $D_{\text{aug}}^{train}$ yields better results than training on $D_{\text{human}}^{train}$ alone, validating our proposed pseudo-labeling method for low-resource historical documents. Crucially, our two-stage learning method outperforms the strategy of simply mixing $D_{\text{aug}}^{train}$ and $D_{\text{human}}^{train}$. We attribute this to the significant size imbalance between the smaller $D_{\text{human}}^{train}$ and the larger $D_{\text{aug}}^{train}$, which hinders the model from leveraging the strengths of both datasets when simply combined. In contrast, our approach demonstrates that decoupling the training process allows the model to effectively capture the advantages of both datasets.

### 5.3 LLM and Human Evaluation Results

Although $D_{\text{human}}^{test}$ consists of human-translated data, it is restricted to specific reign periods, which limits generalization across the entire chronological scope of the AJD corpus. To address this, we performed an LLM-based evaluation on $D_{\text{NT}}^{test}$ by sampling 80 documents from each of the remaining reign periods excluded from pseudo-labeling. Adopting the method proposed by Zheng et al. (2023), models first generate translations for $D_{\text{NT}}^{test}$. Subsequently, the LLM Evaluator assesses the translation quality on a scale of 1 to 10, taking the source Hanja text and the model-generated English translation as inputs. To further enhance evaluation reliability, we provided the LLM Evaluator with the corresponding human-translated Korean text and glossary information as reference materials. We selected Sonnet-4.5 as the evaluator to prevent bias from the pseudo-labeling model.

As shown in Table 5, the LLM-based evaluation demonstrates that for both the 8B and 32B models, the 2-stage learning approach combining pseudo-labeled and human-translated data outperforms models trained exclusively on human-translated data. Notably, given that $D_{\text{NT}}^{test}$ covers the entire AJD period disjoint from the training dataset, these results confirm the effectiveness of pseudo-labeling in enhancing generalization.

Additionally, using 200 examples sampled uniformly across different kings from $D_{\text{NT}}^{test}$, we recruited domain experts in Sinographic literature and Korean history who are proficient in English to evaluate our model against the baseline (Qwen3 32B trained solely on $D_{\text{human}}^{train}$). We asked the experts to rate each translation on a scale of 1-5. Further, to differentiate between models in cases where scores were tied (excluding identical translations), we performed a comparative evaluation where experts were instructed to select the better response.

| | |
|---|---|
| Input Hanja | 兩司啓趙應奎事, 不允. |
| Hunyuan-MT-7B | a petition was submitted by the two departments regarding Zhao Yingkui's case, but it was denied. |
| Qwen3 32B trained on $D_{\text{human}}^{train}$ | The two investigative bodies reported on the matter of Jo Eunggyu, but the king did not approve their request. |
| HERIT | The two offices reported on the matter of Jo Eunggyu, but the king did not approve. |
| Input Hanja | 宜給宣飯及房子, 炊飯, 汲水人等 |
| Hunyuan-MT-7B | It is appropriate to provide food, housing, cooking facilities, and water for the people. |
| Qwen3 32B trained on $D_{\text{human}}^{train}$ | They should be provided with food, a house, a female palace servant, and a water carrier. |
| HERIT | We should provide them with meals, female palace servants, cooks, and water carriers. |
| Input Hanja | 宥魚思漢, 京外從便 |
| Hunyuan-MT-7B | Yu Yu thinks of Han; outside the capital, do as you please. |
| Qwen3 32B trained on $D_{\text{human}}^{train}$ | The king pardoned Eo Sahan and allowed him to reside in a place of his choice in the capital or in the provinces. |
| HERIT | The king pardoned Eo Sahan and allowed him to reside in a place of his choice outside the capital. |

Table 6: Qualitative examples of translations by different models using samples from test dataset. Red denotes incorrect translations, while blue indicates correct ones.

As shown in Fig. 3, our model achieved a score of 3.94, outperforming the baseline score. In addition, HERIT achieved a selection rate of 66.84% in the preference evaluation, more than doubling the baseline figure and confirming its superior translation quality in expert assessments. Moreover, Fig. 4 in Appendix B.1 demonstrates that our pseudo-labeling method mitigates temporal overfitting.

### 5.4 Qualitative Analysis of Translations

We also conducted a qualitative comparison of HERIT against two baselines using samples from the test set: Qwen3 32B, fine-tuned exclusively on human-translated data, and Hunyuan-MT-7B, a leading Chinese translation model.

As shown in the first example of Table 6, both Qwen3 32B and HERIT correctly transliterated proper nouns adhering to Korean pronunciation rules, whereas Hunyuan-MT-7B incorrectly applied Chinese phonetics. In the second example, the term "房子" carries distinct semantics depending on the context: it denotes a 'house' in standard Chinese but refers to a "a female palace servant" in the context of the Joseon Dynasty. While Hunyuan-MT-7B defaulted to the standard meaning and Qwen3 32B conflated the interpretations, HERIT accurately identified the historical context and translated the term correctly. Finally, in the third example, Hunyuan-MT-7B failed to recognize the proper name "魚思漢", and neither baseline recognized the legal term "京外從便" (a form of punishment). In contrast, HERIT accurately translated these domain-specific terms.

These results highlight HERIT's capability to generate accurate English translations by capturing both Sino-Korean nuances and the intricacies of historical documents.

## 6 Conclusion

In this study, we propose HERIT, a framework designed to overcome the scarcity of parallel corpora in historical document translation. By leveraging abundant intermediate resources through a RAG-based strategy, we demonstrated a robust methodology for generating high-quality pseudo-labels without relying solely on expensive human annotation. Our extensive evaluations confirm that HERIT not only outperforms baselines but also effectively mitigates temporal overfitting, a common challenge in chronological archives. Crucially, this work suggests that data augmentation can be a viable solution for other low-resource classical languages where direct translation pairs are rare but modern interpretations exist. Finally, we democratize access to these archives by releasing the translated corpus of AJD and JRS. Future work will focus on extending the context window to handle longer documents and applying this framework to diverse historical domains, thereby broadening the scope of digital humanities research.

## Limitations

Due to budgetary constraints, we primarily employed Gemini-2.5-Flash for the Proposer and Fusor components, focusing exclusively on AJD documents for pseudo-labeling. We anticipate that employing a more diverse set of LLMs as Proposers to enhance translation diversity, as well as incorporating JRS documents as targets, would yield higher-quality pseudo-labels.

Furthermore, hardware limitations necessitated restricting the maximum sequence length to 2,048 tokens, with a limit of 800 Hanja characters. Consequently, approximately 6.7% of the training data was filtered out. For inference, documents exceeding this limit were processed using a chunking strategy. We expect that extending the model's maximum input length will further improve translation performance.

## References


Haithem Afli and Andy Way. 2016. Integrating optical character recognition and machine translation of historical documents. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 109–116.

George C Alter. 2019. The evolution of models in historical demography. *Journal of Interdisciplinary History*, 50(3):325–362.

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

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Siying Chen, Yun Su, Xiuqi Fang, and Jia He. 2020. Climate records in ancient chinese diaries and their application in historical climate reconstruction–a case study of yunshan diary. *Climate of the Past Discussions*, 2020:1–28.

Gregory Clark. 2007. The long march of history: Farm wages, population, and economic growth, england 1209–1869 1. *The Economic History Review*, 60(1):97–135.

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*.

Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 2022. 8-bit optimizers via block-wise quantization. *International Conference on Learning Representations*.

Kun Hwang. 2022. A disastrous situation of plague and collapsed community depicted in a joseon dynasty tale. *Infection & Chemotherapy*, 54(2):388.

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 *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4031–4042.

Ammar Khairi, Daniel D'souza, Marzieh Fadaee, and Julia Kreutzer. 2025. Making, not taking, the best of n. *arXiv:2510.00931*.

Kwang-Hee Kim, Jung-Ho Park, Yongcheol Park, Tian-Yao Hao, and Han-Joon Kim. 2017. Crustal structure beneath the southern korean peninsula from local earthquakes. *Geophysical Journal International*, 209(2):969–978.

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. *arXiv:2503.07891*.

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.

LG AI Research. 2025. Exaone 4.0: Unified large language models integrating non-reasoning and reasoning modes. *arXiv:2507.11407*.

CY LIN. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop, Barcelona, Spain*, pages 74–81.

Dayiheng Liu, Kexin Yang, Qian Qu, and Jiancheng Lv. 2019. Ancient–modern chinese translation with a new large training dataset. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(1):1–13.

Tianqi Liu, Yao Zhao, Rishabh Joshi, Misha Khalman, Mohammad Saleh, Peter J Liu, and Jialu Liu. 2024. Statistical rejection sampling improves preference optimization. In *International Conference on Learning Representations*.

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

Maja Popović. 2017. chrf++: words helping character n-grams. In *Proceedings of the conference on machine translation*, pages 612–618.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Conference on Machine Translation: Research Papers*, pages 186–191.

Young Hoon Rhee. 2014. Economic stagnation and crisis in korea during the eighteenth and nineteenth centuries. *Australian Economic History Review*, 54(1):1–13.

Minju Seo, Jinheon Baek, James Thorne, and Sung Ju Hwang. 2024. Retrieval-augmented data augmentation for low-resource domain tasks. *arXiv:2402.13482*.

J Son, J Jin, H Yoo, J Bak, K Cho, and A Oh. 2022. Translating hanja historical documents to contemporary korean and english. *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1260–1272.

Fangzhou Song, Bin Zhu, Yanbin Hao, and Shuo Wang. 2024. Enhancing recipe retrieval with foundation models: A data augmentation perspective. In *European Conference on Computer Vision*, pages 111–127.

F Richard Stephenson. 2011. Historical eclipses and earth's rotation: 700 bc–ad 1600. In *Highlighting the History of Astronomy in the Asia-Pacific Region: Proceedings of the ICOA Conference*, pages 3–20.

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*.

Cameron G Thies. 2002. A pragmatic guide to qualitative historical analysis in the study of international relations. *International studies perspectives*, 3(4):351–372.

Yuqi Wang, Si Chen, Kaihua Xu, Limei Yan, Xinan Yue, Fei He, and Yong Wei. 2021. Ancient auroral records compiled from korean historical books. *Journal of Geophysical Research: Space Physics*, 126(1):e2020JA028763.

Yong Wei and LiMei Yan. 2024. Solar cycles during the seventeenth century revealed by equatorial aurora records. *Earth and Planetary Physics*, 9(1):182–187.

Yuting Wei, Meiling Li, Yangfu Zhu, Yuanxing Xu, Yuqing Li, and Bin Wu. 2025. A diachronic language model for long-time span classical chinese. *Information Processing & Management*, 62(1):103925.

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*.

Hong-Jin Yang, Changbom Park, and Myeong-Gu Park. 2012. Analysis of historical meteor and meteor shower records: Korea, china and japan. *Proceedings of the International Astronomical Union*, 10(H16):150–151.

Armel Zebaze, Rachel Bawden, and Benoît Sagot. 2025. Llm reasoning for machine translation: Synthetic data generation over thinking tokens. *arXiv:2510.11919*.

Can Zhang, Cheng Zhao, Aifeng Zhou, Haixia Zhang, Weiguo Liu, Xiaoping Feng, Xiaoshuang Sun, Tianlong Yan, Chengcheng Leng, Ji Shen, and 1 others. 2021. Quantification of temperature and precipitation changes in northern china during the "5000-year" chinese history. *Quaternary Science Reviews*, 255:106819.

Zhiyuan Zhang, Wei Li, and Qi Su. 2019. Automatic translating between ancient chinese and contemporary chinese with limited aligned corpora. In *CCF international conference on natural language processing and chinese computing*, pages 157–167.

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.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Mao Zheng, Zheng Li, Bingxin Qu, Mingyang Song, Yang Du, Mingrui Sun, and Di Wang. 2025. Hunyuan-mt technical report. *arXiv:2509.05209*.

## A Training Details

In this survey, we trained the models for one epoch at each stage. To mitigate overfitting and prevent bias toward the final stage, we adjusted the learning rates. Specifically, for Qwen3 32B, the learning rate was set to $1 \times 10^{-5}$ for the first stage and reduced to $5 \times 10^{-6}$ for the second. For Qwen3 8B, we used $2 \times 10^{-5}$ and $1 \times 10^{-5}$, respectively. A cosine decay scheduler was employed. We set the batch size to 64 for the first stage and 16 for the second. All models, including baselines, were trained using the PagedAdamW 8-bit optimizer (Dettmers et al., 2022) combined with FSDP (Zhao et al., 2023) on eight NVIDIA H200 GPUs.

## B Experimental Results

### B.1 Temporal Overfitting

We further analyzed the human evaluation results to investigate whether our proposed pseudo-labeling method effectively mitigates temporal overfitting, which typically arises when models are trained on data concentrated on specific kings. Given that the evaluation examples were stratified by king (as described in Section 5.3), we calculated the average scores for each king for both HERIT and the baseline, Qwen3 32B (trained exclusively on human labels). We then computed the performance gap between the two models for each king and visualized these differences chronologically.

As shown in Fig. 4, human experts primarily translated documents from the reign of King Sejong. Consequently, Qwen3, trained on this biased data, exhibits performance comparable to or marginally better than HERIT during this specific period. However, HERIT consistently demonstrates superior performance even in periods outside this reign. These results suggest that training on data confined to a specific era can lead to temporal overfitting, whereas the proposed pseudo-labeling approach effectively alleviates this issue and enhances overall translation robustness.

### B.2 Number of Few-shot Examples

Fig. 5 demonstrates that the number of few-shot examples influences the performance of both the Proposer and Fusor. Notably, the Proposer shows consistent performance improvements up to 128 shots. However, to strike a balance between performance and the increased LLM inference costs associated with more shots, we configured the Proposer with 64 shots for pseudo-labeling. Additionally, we determined that a temperature of 0.3 and a draft count of 5 yielded the best performance; thus, these settings were applied during pseudo-labeling. In contrast, while the Fusor exhibits performance improvements as the number of examples increases, the marginal gains are less pronounced compared to the Proposer. Therefore, considering the trade-off between performance and cost, we set the number of few-shot examples for the Fusor to 12.

| | BLEU | METEOR | chrF++ |
|---|---|---|---|
| No aggregation | 41.763 | 0.644 | 64.138 |
| Best-of-N | 42.148 | 0.649 | 64.678 |
| Fuse-of-N | **42.293** | **0.652** | **65.025** |

Table 7: Comparison of pseudo-labeling performance among FusioN, BoN, and Baseline.

### B.3 Fuse-of-N vs Best-of-N

We also evaluated the pseudo-labeling performance of multi-candidate aggregation methods, specifically FusioN and BoN. As shown in Table 7, BoN outperforms the baseline, which generates translations in a single stage without candidate aggregation. However, FusioN demonstrates superior performance compared to BoN, suggesting that synthesizing the final output from all candidates is more effective than selecting a single best candidate.

## C Human Evaluation

To compare our model against the baseline, we recruited domain experts in Sinographic literature and Korean history who are proficient in English to evaluate 200 items. Prior to the task, the experts were provided with guidelines on the interface usage and assessment criteria.

We developed a custom web-based interface to facilitate this evaluation, as shown in Fig. 6. The interface displays the source Hanja text and associated glossary information at the top, while the anonymized outputs from both models are presented in the center. Evaluators rated each translation using a Likert scale and were encouraged to provide qualitative feedback on errors such as semantic distortion, omission, terminology misuse, or unnatural phrasing. To enable a strict comparison, a forced-choice mechanism was activated whenever identical scores were assigned, requiring the evaluator to select the superior translation.

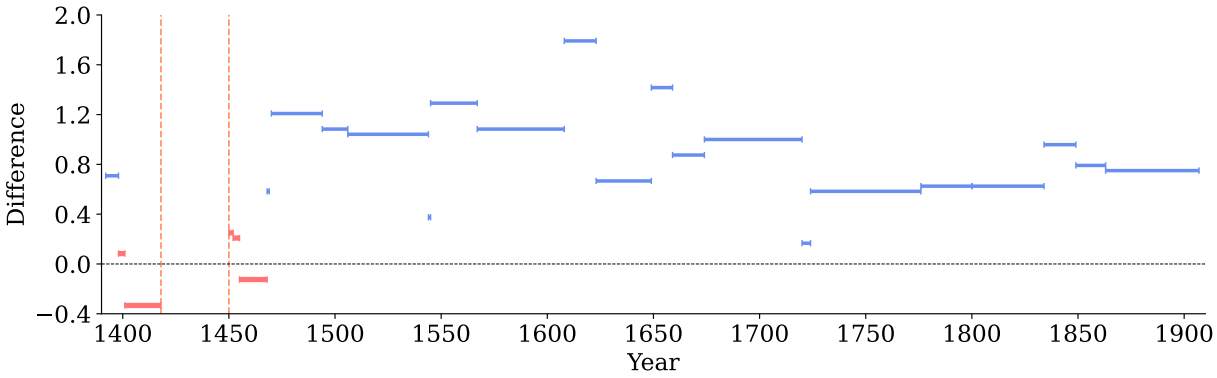


Figure 4: Differences in human expert scores between HERIT (ours) and Qwen3 32B (trained on soly human labels) across each king, arranged by their reign periods. Positive values indicate that HERIT achieved higher expert evaluation scores compared to the baseline, highlighting its robustness across different time periods.

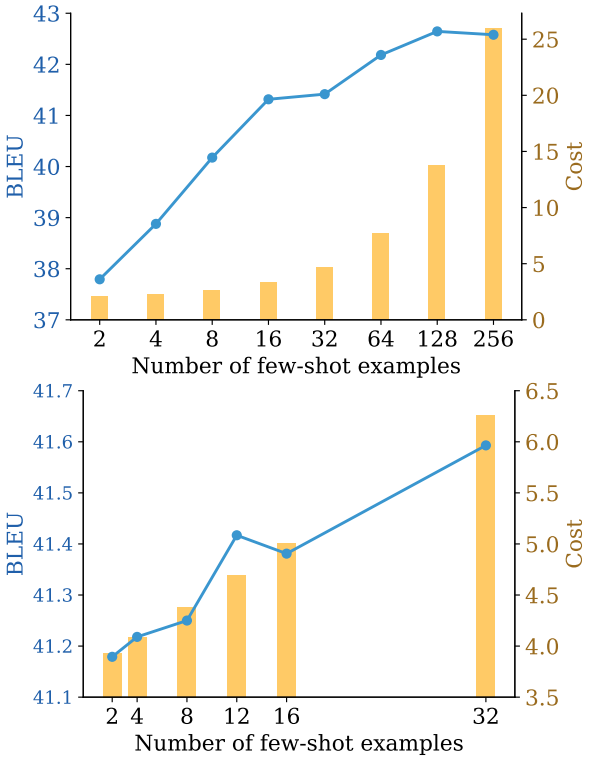


Figure 5: Performance and cost of the Proposer and Fusor across different numbers of few-shot examples.


Figure 6: Web-based evaluation tool for human experts.

## D Prompt Format

Fig. 7 illustrates the base prompt template used for direct Hanja-to-English translation in the baseline models, as discussed in Sections 4.1 and 5. The system prompt defines the instructions for the translation task and provides a description of Hanja within the *Task* section. It also outlines *Guidelines for translation* and briefly explains the user prompt structure and expected output format. In the user prompt, the source Hanja text and relevant glossary information are provided in the *Input Document* section. For RAG-based baselines, relevant *Examples* are included as (Hanja, English) pairs.

Fig. 8 and 9 depict the prompt structures utilized for the pseudo-labeling process described in Section 4.2. In contrast to the base prompt, the Proposer and Fusor prompts integrate Korean translations provided by human experts. The Fusor prompt follows a structure similar to the Proposer but differs in the following aspects: drawing upon the implementation by Khairi et al. (2025), it presents a *Task* section describing the draft fusion method and *Fusion Steps* specifying the procedure. In the user prompt, the task instructions and guidelines from the Proposer stage are provided under *Translation Instruction*, while the drafts generated by the Proposer are listed under *Translation Candidates*.

**System Prompt**

**### Task**
Your task is to translate the historical documents into English while preserving their original meaning and context. The documents were created in the Joseon era in Korea (from the late 14th century to the early 20th century) and were written in Hanja, a traditional writing system used at the time.

**### Guidelines**
Focus on accurately conveying the meaning, nuances, and context of the original texts.

**### Input**
The document contains the original Hanja text, and may optionally include glossaries or examples of expert translations.

**### Output JSON Format**
`{"english_translation": str}`

**User Prompt**

**### Examples**
[Example 1]
Hanja: 上王幸壤, 觀離宮之役而還. 離宮乃壤縣古基, 去都城四十里.
English: The abdicated king visited Pungyang, observed the construction of a detached palace, and returned. The detached ...
[Example 2]
Hanja: 上王幸離宮南郊, 觀放鷹.
English: The abdicated king went to the southern outskirts of the temporary palace and viewed falconry.
...

**### Input Document**
- hanja: 上王幸毋岳, 觀離宮造成之役.
- glossaries: [{"hanja": "上王", "korean": "상왕", "term": "Abdicated King"}, {"hanja": "毋岳", "term": "Muak"}, ...]

Figure 7: An example of the base prompt.

**System Prompt**

**### Task**
Your task is to translate the historical documents into English while preserving their original meaning and context. The documents were created in the Joseon era in Korea (from the late 14th century to the early 20th century) and were written in Hanja, a traditional writing system used at the time.

**### Guidelines**
Focus on accurately conveying the meaning, nuances, and context of the original texts.

**### Input**
The document contains the original Hanja text, its Korean translation, and may optionally include glossaries or examples of expert translations.

**### Output JSON Format**
`{"english_translation": str}`

**User Prompt**

**### Examples**
[Example 1]
Hanja: 正朝使工曹判書崔閏德回自京師.
Korean: 정조사인 공조 판서 최윤덕이 북경에서 돌아왔다.
English: Minister of Works Choe Yundeok, envoy for offering the New Year's greetings to the Emperor of Ming China ...
[Example 2]
Hanja: 謝恩使星原君 李正寧, 副使刑曹左參判崔士儀, 回自京師.
Korean: 사은사 성원군 이정녕과 부사 형조 좌참판 최사의가 북경으로부터 돌아왔다.
English: The Envoy for Expressing Gratitude for Imperial Benevolence, Lord of Seongwon Yi Jeongnyeong, and ...
...

**### Input Document**
- Hanja: 正朝使崔士儀, 副使柳漢回自京師.
- Korean: 정조사 최사의 · 부사 유한이 북경에서 돌아왔다.
- Glossaries: [{"hanja": "正朝使", "korean": "정조사", "term": "envoy for the New Year's greeting"},
  {"hanja": "崔士儀", "term": "Choe Saui"} ... ]

Figure 8: The prompt structure designed for the Proposer.

**System Prompt**

**### Task**
Based on the provided User Input, fuse them into a better translation that meets the criteria. The source texts are Joseon-era documents originally written in Hanja between the late 14th and early 20th centuries. You must create a fused translation that faithfully preserves the original meaning, nuances, and historical context.

**### User Input**
- Examples: (reference material) Certified example translations.
- Input Document
  - Hanja: The original Hanja text.
  - korean: Its Korean translation.
  - Glossaries: (reference material) Glossaries to clarify specific words and characters.
- Translation Instruction: Additional instructions for translation.
- Translation Candidates: Multiple English translation candidates.

**### Fusion Steps**
1. Read all source materials and translation candidates.
2. Compare strengths and weaknesses.
3. Generate a fused final translation.

**### Output JSON Format**
```
{"final_translation": str}
```

**User Prompt**

**### Examples**
[Example 1]
Hanja: 正朝使工曹判書崔閏德回自京師.
Korean: 정조사인 공조 판서 최윤덕이 북경에서 돌아왔다.
English: Minister of Works Choe Yundeok, envoy for offering the New Year's greetings to the Emperor of Ming China ...
[Example 2]
Hanja: 正朝使都摠制柳殷之, 副使漢城府尹沈道源, 回自京師.
Korean: 정조사 도총제 유은지와 부사 한성부 윤 심도원이 북경으로부터 돌아왔다.
English: Chief Commander General Yu Eunji, the envoy for offering the New Year's greetings to the Emperor of Ming ...
...

**### Input Document**
- Hanja: 正朝使崔士儀, 副使柳漢回自京師.
- Korean: 정조사 최사의 · 부사 유한이 북경에서 돌아왔다.
- Glossaries: [{"hanja": "正朝使", "korean": "정조사", "term": "envoy for the New Year's greeting"},
  {"hanja": "崔士儀", "term": "Choe Saui"} ... ]

**### Translation Instruction**
Task: Translate the historical documents into English while preserving their original meaning and context.
Guidelines: Focus on accurately conveying the meaning, nuances, and context of the original texts.

**### Translation Candidates**
- Candidate_1: Envoy for the New Year's greeting Choe Saui and Vice-Envoy Yu Han returned from Beijing.
- Candidate_2: Choe Saui, envoy for the New Year's greeting, and Yu Han, vice-envoy, returned from Beijing.
- Candidate_3: Choe Saui, envoy for offering the New Year's greetings to the Emperor of Ming China, and Yu Han ...
...

Figure 9: An example of the prompt format used in the Fusor.