Bridging Modern AI and Historical Language: A Study on Large Language Model Adaptation to 1920s–30s Korean

Anonymous ACL submission

Abstract

Large Language Models (LLMs) primarily train on modern texts, limiting their ability to process historical language effectively. This study investigates methods for enhancing LLM adaptability to historical Korean, specifically focusing on the 1920s–30s literary domain. We construct a novel dataset of Korean literature from this period and introduce a sentence-final ending prediction task to evaluate historical linguistic adaptation. Our results demonstrate that adapting LLMs with targeted historical text exposure improves their ability to generate era-specific linguistic patterns while maintaining stability in longer contexts. The findings provide insights into the broader challenge of modeling diachronic language variations and highlight the potential of historical text adaptation techniques for computational humanities research.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities in processing modern texts, yet their adaptability to historical languages remains a challenge. Historical texts often exhibit distinct grammatical structures, vocabulary, and stylistic conventions that diverge significantly from contemporary usage, posing difficulties for models trained predominantly on modern corpora (Manjavacas Arevalo and Fonteyn, 2021; Palmero Aprosio et al., 2022; Rastas et al., 2022; Gabay et al., 2022; Yamshchikov et al., 2022; Al-Laith et al., 2024). While pretraining on historical corpora has been shown to improve performance in diachronic text processing, retraining models from scratch is often impractical due to the computational demands and scarcity of historical text data. Thus, effective methods for adapting existing models to historical language remain an open research question.

Korean from the 1920s–30s presents an ideal testbed for studying historical language adaptation due to its substantial syntactic and morphological differences from modern Korean. However, historical Korean has received relatively little attention in NLP research, particularly in the context of large-scale language models. This study addresses this gap by investigating how exposure to historical Korean texts can improve LLMs' ability to process diachronic linguistic patterns.

Since Western languages dominate the training data of most LLMs (Lai et al., 2024), low-resource historical languages such as early 20th-century Korean provide an opportunity to explore the extent to which LLMs can generalize beyond their primary training distribution. To evaluate historical language adaptation, we introduce a sentence-final ending prediction task—a linguistically significant feature in Korean that encodes grammatical, stylistic, and pragmatic information (Han, 2020). Given that sentence-final endings play a crucial role in meaning construction, their accurate prediction serves as a proxy for assessing a model's ability to internalize historical linguistic structures.

To this end, we constructed a dataset of 1,000 books from the 1920s–30s, capturing a diverse range of sentence-final endings and stylistic patterns. Using this dataset, we adapted various contemporary LLMs and evaluated their performance on the sentence-final ending prediction task. Our experiments show that exposing models to historical text significantly enhances their ability to generate era-specific linguistic patterns, with an average accuracy improvement of 3.78 percentage points. Moreover, models trained on historical data demonstrated greater stability in longer contexts, suggesting that historical text adaptation can lead to improved coherence in diachronic text processing.

These findings provide insights into the broader challenge of modeling historical languages and highlight the potential of targeted adaptation strategies for computational humanities research. By examining how LLMs respond to historical linguis-

tic variations, this study contributes to the development of more adaptable AI systems capable of processing diverse textual domains across time periods.

2 Related Work

Adapting language models to historical texts presents unique challenges, as linguistic features evolve over time, leading to significant shifts in syntax, vocabulary, and orthographic conventions. Various studies have explored how different training strategies influence language model performance on diachronic text processing.

One approach involves pretraining models on historical corpora to improve their ability to process texts from specific time periods. For example, MacBERTh, trained on a historical English corpus spanning 1450 to 1900, demonstrated that era-specific linguistic patterns strongly influence model performance (Manjavacas Arevalo and Fonteyn, 2021). Similarly, BERToldo, which integrates both pretraining and fine-tuning, has shown significant improvements in processing Dante Alighieri's classical Italian texts (Palmero Aprosio et al., 2022). In French, D'AlemBERT, trained on the early modern French corpus FREEMmax, has exhibited cross-temporal generalization, adapting effectively to language variations across different periods (Gabay et al., 2022). Greek and Latin NLP models have leveraged transfer learning techniques to improve performance on historical texts, particularly for author identification tasks (Yamshchikov et al., 2022). In Scandinavian languages, models trained on 19th-century Danish and Norwegian corpora have outperformed modern-text-trained models in tasks such as word sense disambiguation and sentiment analysis (Al-Laith et al., 2024).

While historical text adaptation has been explored in several Indo-European languages, low-resource historical languages remain underrepresented in this field. Since most LLMs are trained predominantly on modern and Western-language corpora (Lai et al., 2024), investigating how models adapt to historical Korean—a low-resource language with substantial diachronic linguistic variation—offers new insights into the generalizability of language models. Unlike Indo-European languages, Korean is an agglutinative language where grammatical meaning is heavily encoded in sentence-final endings (Han, 2020). These endings convey stylistic, pragmatic, and syntactic information, making them a linguistically rich evaluation metric for historical adaptation.

This study builds on prior research by examining how targeted exposure to historical Korean texts influences LLMs' ability to internalize diachronic linguistic shifts. Unlike previous work that primarily focuses on pretraining, our study investigates whether existing models can be effectively adapted to historical language through structured domain adaptation techniques, without requiring computationally expensive full-scale pretraining. By focusing on sentence-final ending prediction as a key evaluation metric, we aim to provide a fine-grained analysis of how LLMs process linguistic variation across time.

3 Corpus and Task Design

3.1 Corpus Construction

Korean has undergone substantial linguistic transformations over the past century due to major socio-historical events, including the Japanese colonial period and the Korean War (Hong, 2009). By the 1960s, Korean had shifted toward its modern form, exhibiting significant changes in syntax, morphology, and vocabulary (Han, 2015). This study focuses on Korean from the 1920s–30s, a period that retains distinct linguistic structures, making it an ideal testbed for evaluating historical language adaptation.

To facilitate this research, we constructed a corpus of 1,145 novels written by 56 authors, sourced from *Gongumadang*[1], a South Korean government platform for public domain works. Only texts with expired copyrights were included to ensure legal compliance. The dataset reflects authentic language usage from the early 20th century, capturing key grammatical and stylistic patterns of the era.

3.2 Sentence-Final Ending Prediction Task

To systematically evaluate how models adapt to historical Korean, we introduce sentence-final ending prediction as a core evaluation task. Sentence-final endings in Korean encode grammatical, stylistic, and pragmatic information, making them a linguistically rich feature for assessing historical text understanding (Han, 2020). Predicting the correct ending requires a model to grasp historical morphosyntactic structures and stylistic conventions, beyond simple lexical memorization.

[1] https://gongu.copyright.or.kr/gongu/main/main.do

For this task, sentence-final endings were removed from literary texts, and models were prompted to generate the most contextually appropriate completion. This design allows us to measure a model's ability to internalize historical linguistic patterns and maintain coherence over different context lengths.

To ensure that models were not inadvertently exposed to the dataset during pretraining, we applied the Name Cloze task(Chang et al., 2023), which tests whether a model can recover masked character names from context. With a measured accuracy of 0%, we confirmed that the corpus had not been included in pretraining data, ensuring unbiased evaluation. Additional details are provided in AppendixB.

4 Experimental setup

4.1 Model Selection

To assess how exposure to historical texts influences model adaptation, we selected seven base models spanning different architectures and training paradigms: Qwen 2.5 7B, Mistral 7B, Llama 3 8B, Llama 3.2 3B, Llama-3-Korean-Bllossom-8B, EEVE-Korean 2.8B, and EEVE-Korean 10.8B (see Appendix C for details).

Each model was further adapted using the LoRA-based parameter-efficient tuning approach (Hu et al., 2021), resulting in 14 models (7 base and 7 adapted). LoRA enables efficient fine-tuning while maintaining the generalization capabilities of the base models. Additional training details are provided in Appendix D.

4.2 Evaluation Procedure

We designed two evaluation settings to assess historical language adaptation across different context lengths. In both cases, the sentence-final ending prediction task was used as the primary metric:

3-Sentence Test: Each test instance consisted of three sentences, ensuring sufficient contextual information while filtering out overly short examples. This resulted in 939 test instances (average 59.1 tokens per example).

10-Sentence Test: To analyze the impact of longer contexts, each instance contained ten sentences, producing 304 test instances (average 232.5 tokens per example).

4.3 Performance Metrics

Model performance was measured as the proportion of correctly predicted sentence-final endings, averaged over three trials. This metric provides insights into how well models internalize historical morphosyntactic patterns and whether longer contexts contribute to improved prediction consistency.

5 Results and Discussion

Rank	Llama 3 8B	Llama 3 8B (Fine-tuned)
1	**-오(-o)** (3.98%)	-지(-ji) (5.95%)
2	-지(-ji) (2.99%)	-다(-da) (2.97%)
3	-습니다(-seumnida) (2.10%)	**-읍니다(-umnida)** (2.42%)
4	-어(-eo) (1.99%)	-니(-ni) (2.20%)
5	-다(-da) (1.88%)	-나(-na) (2.20%)
6	-요(-yo) (1.88%)	**-소(-so)** (2.09%)
7	-고(-go) (1.77%)	**-느냐(-neunya)** (1.98%)
8	-니(-ni) (1.77%)	**-오(-o)** (1.87%)
9	-는가(-neunga) (1.66%)	-요(-yo) (1.87%)
10	라(-ra) (1.55%)	-구나(-guna) (1.65%)
	Others (78.43%)	Others (74.80%)

Table 1: Top 10 most frequent sentence-final endings generated by Llama 3 8B and Llama 3 8B (Fine-tuned) in the 10-sentence prediction task. **Bolded** endings indicate archaic sentence-final forms.

5.1 Adaptation to Historical Linguistic Norms

Table 1 presents the impact of historical text adaptation on sentence-final endings. In the base model (Llama 3 8B), only *-오(-o)* appears among the Top 10 archaic endings. However, after adaptation, the model generates multiple archaic forms, including *-읍니다(-umnida)*, *-소(-so)*, and *-느냐(-neunya)*, demonstrating an improved grasp of period-specific linguistic structures.

A notable example is *-읍니다(-umnida)*, a sentence-final ending officially replaced by *-습니다(-seumnida)* in a 1988 language reform. The adapted model correctly generates this form, suggesting that historical text exposure enhances the model's ability to reflect linguistic norms from earlier eras.

5.2 Performance Across Models

Table 2 summarizes sentence-final ending prediction performance across different models. On average, historical text adaptation improved accuracy by 3.78 percentage points. The largest relative gain was observed in Llama 3 8B, where accuracy in the

Model	Test			
	Base		Fine-tuned	
	3 sentence	10 sentence	3 sentence	10 sentence
Qwen 2.5 7B	10.87	8.36	12.12	12.43
Mistral 7B	5.08	4.73	9.63	10.34
Llama 3 8B	7.92	5.61	12.19	12.54
Llama 3.2 3B	6.79	4.18	8.17	8.18
Bollossom 8B	8.39	6.38	11.19	11.55
EEVE 2.8B	5.86	4.18	6.82	8.03
EEVE 10.8B	13.96	13.75	16.95	**18.70**
GPT 4o	14.39	15.84	-	-

Table 2: The table compares accuracy between base and fine-tuned models across different context lengths. The **bolded** values represent the highest accuracy recorded across all test conditions, including different context lengths and whether the model was fine-tuned.

10-sentence test increased from 5.61% to 12.54% (123.5%).

Interestingly, GPT-4o achieved the highest overall accuracy among baseline models (14.39% in the 3-sentence test, 15.84% in the 10-sentence test). However, it still underperformed compared to the adapted EEVE 10.8B model (16.95% and 18.70%, respectively). This suggests that even state-of-the-art models struggle with historical Korean, reinforcing the importance of targeted historical adaptation for diachronic language modeling.

5.3 Effect of Context Length

Context length significantly influenced model performance. While most base models showed a decline in accuracy as context length increased (with some dropping by 38.4%), adapted models maintained or improved performance.

Adapted models showed an average improvement of 2.60 percentage points in the 3-sentence test, which increased to 4.94 percentage points in the 10-sentence test, suggesting that historical adaptation not only enhances single-sentence completion but also benefits longer-context coherence.

Interestingly, GPT-4o deviated from this trend by improving with longer context lengths, similar to the adapted models. This is likely due to its advanced long-context comprehension ability (Li et al., 2024), but its lower overall accuracy compared to specialized models indicates that pretraining alone does not fully compensate for historical language shifts.

5.4 Generalization Across Multilingual Models As shown in Table 2, the best-performing model overall was EEVE 10.8B, a Korean-specialized model. However, adaptation also significantly improved multilingual models, particularly Llama 3 8B, which showed the highest relative accuracy gain.

This finding suggests that historical adaptation benefits are not limited to language-specific models. Even models without explicit Korean pretraining exhibited improved historical text understanding, indicating that adaptation techniques could be extended to low-resource historical languages beyond Korean.

6 Conclusion

This study investigated how adapting LLMs to historical texts enhances their ability to process diachronic linguistic structures. We constructed a corpus of 1920s–30s Korean literature and introduced sentence-final ending prediction as an evaluation method tailored to the linguistic characteristics of historical Korean.

Our findings demonstrate that historical text adaptation improves model accuracy in generating period-specific linguistic forms, with an observed 3.78 percentage point increase in prediction performance. Adapted models exhibited greater stability over longer contexts and a stronger ability to reflect historical grammatical norms, as seen in their use of archaic sentence-final endings. Notably, multilingual models such as Llama 3 8B also benefited significantly, indicating that historical adaptation techniques can enhance performance even in models without explicit pretraining on Korean.

These results suggest that targeted adaptation strategies can improve LLMs' handling of diachronic language variations, offering valuable insights for both NLP research and computational humanities. Future work could explore more efficient adaptation techniques, such as domain-adaptive pretraining or contrastive learning, to further refine historical language modeling. Additionally, expanding evaluation beyond sentence-final ending prediction—through style transfer, historical text translation, or broader syntactic analysis—could provide a more comprehensive assessment of model adaptation to historical languages.

Limitations

Single-Task Evaluation This study primarily employs sentence-final ending prediction as an evaluation metric for historical text adaptation. While this task captures key morphosyntactic and stylistic patterns, it does not comprehensively assess semantic

coherence, discourse-level understanding, or pragmatic adaptation to historical texts. Future research should complement this approach with style transfer, historical text translation, or broader syntactic evaluation to provide a more holistic measure of diachronic language adaptation.

Additionally, sentence-final ending prediction is a newly introduced task, and its reliability in measuring historical adaptation warrants further validation. While our findings indicate its potential as an effective proxy for evaluating historical linguistic adaptation, future work should explore inter-annotator agreement studies, benchmark comparisons, and human evaluation frameworks to strengthen its credibility.

Single-Language Evaluation This study focuses exclusively on historical Korean, and the generalizability of our approach to other languages remains untested. Given that languages such as Japanese, Mongolian, and Turkish also rely on sentence-final structures to encode grammatical and stylistic variation, extending this approach to other agglutinative languages would provide stronger cross-linguistic validation.

Furthermore, while our results suggest that historical adaptation improves diachronic text modeling even in multilingual models, the extent to which similar gains can be observed in non-Korean datasets remains an open question. Future research should investigate how historical adaptation impacts low-resource historical languages beyond Korean, particularly in underrepresented linguistic families.

Data Availability and Bias Although our dataset is derived from public domain Korean literary texts from the 1920s–30s, it may not fully capture the linguistic diversity of the time period. The dataset consists primarily of formal literary works, which could introduce stylistic bias, making it less representative of spoken language or informal writing styles from that era.

Additionally, authorial and regional biases may exist within the dataset, as certain dialects or sociolects could be underrepresented. Future work should aim to expand the dataset by incorporating historical newspapers, personal letters, and legal documents to enhance linguistic diversity. Furthermore, evaluating models on diachronic corpora spanning multiple historical periods (e.g., 18th–20th centuries) could provide deeper insights into how models adapt to gradual linguistic shifts.

Model Size and Computational Constraints Our study primarily examines mid-sized models (2.8B–10.8B parameters), leaving open the question of whether larger-scale LLMs (e.g., GPT-4, PaLM-2, Llama 70B) exhibit similar historical adaptation trends. While preliminary results suggest that even state-of-the-art models struggle with diachronic variation, further investigation is needed to determine whether scaling up model size mitigates the need for explicit historical adaptation.

Moreover, our adaptation approach relies on parameter-efficient tuning (LoRA) rather than full pretraining on historical data. While this provides a computationally feasible strategy, it remains unclear whether pretraining from scratch on historical corpora would yield fundamentally different adaptation patterns. Future research could compare the effectiveness of pretraining vs. fine-tuning vs. domain-adaptive pretraining in historical language modeling.

Practical Applications and Real-World Deployment While our results demonstrate improvements in historical text modeling, the practical applicability of these adaptations remains underexplored. Real-world applications, such as historical document analysis, machine translation for archaic texts, and linguistic preservation efforts, require additional validation to determine whether adapted models generalize beyond controlled evaluation settings.

References

Ali Al-Laith, Alexander Conroy, Jens Bjerring-Hansen, and Daniel Hershcovich. 2024. Development and evaluation of pre-trained language models for historical Danish and Norwegian literary texts. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 4811–4819, Torino, Italia. ELRA and ICCL.

Kent Chang, Mackenzie Cramer, Sandeep Soni, and David Bamman. 2023. Speak, memory: An archaeology of books known to ChatGPT/GPT-4. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7312–7327, Singapore. Association for Computational Linguistics.

ChangSu Choi, Yongbin Jeong, Seoyoon Park, Inho Won, HyeonSeok Lim, SangMin Kim, Yejee Kang, Chanhyuk Yoon, Jaewan Park, Yiseul Lee, HyeJin Lee, Younggyun Hahm, Hansaem Kim, and Kyung-Tae Lim. 2024. Optimizing language augmentation for multilingual large language models: A case study on Korean. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics,*

Language Resources and Evaluation, pages 12514–12526, Torino, Italia. ELRA and ICCL.

Simon Gabay, Pedro Ortiz Suarez, Alexandre Bartz, Alix Chagué, Rachel Bawden, Philippe Gambette, and Benoît Sagot. 2022. From FreEM to d'AlemBERT: a large corpus and a language model for early Modern French. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3367–3374, Marseille, France. European Language Resources Association.

Seungkyu Han. 2020. A study on the use of final endings in korean language conversation. *The Journal of Humanities and Social science 21*, 11(4):2315–2328.

Young Gyun Han. 2015. On setting-up 'establishing stage of modern korean' and subdivision of the era. *Hanminjok Emunhakhoe*, (70):63–108.

Damian Hodel and Jevin West. 2024. Response: Emergent analogical reasoning in large language models. *Preprint*, arXiv:2308.16118.

Jongseon Hong. 2009. The diachronic change in korean grammar of the 20th century. *The Korean Language and Literature*, (152):35–61.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Zhiqiang Hu, Lei Wang, Yihuai Lan, Wanyu Xu, Ee-Peng Lim, Lidong Bing, Xing Xu, Soujanya Poria, and Roy Lee. 2023. LLM-adapters: An adapter family for parameter-efficient fine-tuning of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5254–5276, Singapore. Association for Computational Linguistics.

Seungduk Kim, Seungtaek Choi, and Myeongho Jeong. 2024. Efficient and effective vocabulary expansion towards multilingual large language models. *Preprint*, arXiv:2402.14714.

Wen Lai, Mohsen Mesgar, and Alexander Fraser. 2024. LLMs beyond English: Scaling the multilingual capability of LLMs with cross-lingual feedback. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8186–8213, Bangkok, Thailand. Association for Computational Linguistics.

Zhuowan Li, Cheng Li, Mingyang Zhang, Qiaozhu Mei, and Michael Bendersky. 2024. Retrieval augmented generation or long-context LLMs? a comprehensive study and hybrid approach. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 881–893, Miami, Florida, US. Association for Computational Linguistics.

Inbal Magar and Roy Schwartz. 2022. Data contamination: From memorization to exploitation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 157–165, Dublin, Ireland. Association for Computational Linguistics.

Enrique Manjavacas Arevalo and Lauren Fonteyn. 2021. MacBERTh: Development and evaluation of a historically pre-trained language model for English (1450-1950). In *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*, pages 23–36, NIT Silchar, India. NLP Association of India.

National Institute of Korean Language. 2005. *Korean Grammar for Foreigners: Systematic Approach*. Korean Language Education Series. Communication Books.

Alessio Palmero Aprosio, Stefano Menini, and Sara Tonelli. 2022. BERToldo, the historical BERT for Italian. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 68–72, Marseille, France. European Language Resources Association.

Iiro Rastas, Yann Ciarán Ryan, Iiro Lassi Ilmari Tiihonen, Mohammadreza Qaraei, Liina Repo, Rohit Babbar, Eetu Mäkelä, Mikko Tolonen, and Filip Ginter. 2022. Explainable publication year prediction of eighteenth century texts with the bert model. In *Proceedings of the 3rd International Workshop on Computational Approaches to Historical Language Change 2022*, pages 68–77, United States. The Association for Computational Linguistics. Workshop on Computational Approaches to Historical Language Change.

Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. 2024. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1819–1862, Mexico City, Mexico. Association for Computational Linguistics.

Ivan P. Yamshchikov, Alexey Tikhonov, Yorgos Pantis, Charlotte Schubert, and Jürgen Jost. 2022. BERT in plutarch's shadows. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6071–6080, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

7 Appendix

A Korean Sentence-Final Endings

Table 3 compares sentences in English and Korean. While English sentences exhibit variation through changes in word order and the addition or omission of constituents, Korean sentences are primarily determined by sentence-final endings such as *-다(-da)*, *-라(-ra)*, *-니(-ni)*, and *-자(-ja)*.

	English	Korean
Declarative	I went to school.	나는 학교에 **갔다**. *naneun hakgyoe gatda.*
Imperative	Go to school.	너는 학교에 **가라**. *neoneun hakgyoe gara.*
Interrogative	Did you go to school?	너는 학교에 **갔니**? *neoneun hakgyoe gatni?*
Propositive	Let's go to school.	우리, 학교에 **가자**. *uri, hakgyoe gaja.*

Table 3: Comparison of English and Korean sentence types.

Unlike English, Korean features a highly developed honorific system. The appropriate use of honorific expressions requires consideration of factors such as age, rapport, position, and family hierarchy, all of which are reflected in the forms of sentence-final endings. For example, the propositive sentence-final ending *-자(-ja)* shown in Figure 2 can be replaced by forms such as *-시지요(-sijiyo)*, *-ㅂ시다(-psida)*, *-세(-se)*, *-세요(-seyo)*, *-셔요(-syeoyo)*, and *-지(-ji)* depending on the degree and manner of honorificity (National Institute of Korean Language, 2005). Considering the additional variations introduced by the speaker's attitude, intent, dialect, and historical context, selecting the appropriate sentence-final ending poses a significant challenge for beginners learning Korean.

B Name Cloze

또 결혼두 그렇지, 법률에 성년이란 게 있는데 스물하나가 돼야 비로소 결혼을 할 수 있는 걸세. 자넨 물론 아들이 늦을 걸 염려하지만 [MASK]를 말하면 이제 겨우 열여섯이 아닌가. 그렇지만 아까 빙장의 말씀이 올 갈에는 열일을 제치고라도 성례를 시켜 주겠다니 좀 고마울 겐가.
→ **점순이** (Kim Yu-jeong, *Spring-Spring*, 1935)

고개가 앞에 놓인 까닭에 세 사람을 나귀를 내렸다. 둔덕은 험하고 입을 벌리기도 대근하여 이야기는 한동안 끊겼다. 나귀는 건듯하면 미끄러졌다. [MASK]는 술이 차 몇 번이고 다리를 쉬지 않으면 안 되었다. 고개를 넘을 때마다 나이가 알찼다. → **허생원** (Lee Hyo-seok, *When Buckwheat Flowers Bloom*, 1936)

Figure 1: An example of the *Name Cloze* task applied to a Korean literary dataset from the 1920s–1930s.The bold text is the name of the character that goes in [MASK]

In this study, the sentence-final endings in the literary works were removed, and LLMs were tasked with predicting them. If an LLM had previously encountered this data during pretraining, its inference results might be biased, thereby compromising experimental objectivity. To mitigate this, the *Name Cloze* task was employed across the set of LLMs used in our experiments to assess potential contamination within the 1920s–30s Korean literature dataset as illustrated in Figure 1.

Modern LLMs demonstrate remarkable reasoning and generation capabilities through training on vast datasets; however, debate continues as to whether these models truly understand text or merely memorize it (Wu et al., 2024; Hodel and West, 2024). Additionally, the issue of data contamination has become a significant concern—referring to the unintentional inclusion of downstream test sets in a model's pretraining data, which can affect evaluation and performance (Magar and Schwartz, 2022).

Chang et al. (2023) proposed the *Name Cloze* task to assess data contamination in copyright-expired English book datasets. In this task, a single name in a given sentence is masked as `[MASK]`, with the sentence consisting of 40 to 60 tokens and containing no other names. The model must predict the correct name for the `[MASK]` solely based on the provided context; a prediction is considered correct only if it exactly matches the actual name. Human accuracy on the *Name Cloze* task approaches 0%, making it nearly impossible to infer the correct answer based solely on context. Even a strategy that always predicts the most frequent name achieves only 0.6% accuracy. Consequently, this test is well suited for evaluating the extent to which a model has memorized the text rather than its general language understanding capabilities.

According to Chang et al. (2023), data contamination was evident in copyright-expired English book datasets. For example, GPT-4 achieved an accuracy of 98% on Lewis Carroll's *Alice's Adventures in Wonderland*, and similarly high accuracies were observed for *Harry Potter and the Sorcerer's Stone*, *1984*, and *The Lord of the Rings*.

C Korean-Specific Models

Bllossom Llama-3-Korean-Bllossom-8B is a variant of the Llama 2 enhanced for Korean. This model achieves vocabulary expansion by combining the vocabularies of KoBERT and Llama 2 and employs bilingual pretraining to align semantic knowledge between high-resource and low-resource languages. It has demonstrated performance improvements ranging from an average of 1.8% to 8% across eight benchmark tasks, outper-

forming existing Korean models such as Kullum and Polyglot by 93% (Choi et al., 2024).

EEVE EEVE-Korean-2.8B/10.8B is an efficient model designed to extend an English-centric large language model to Korean. Through parameter freezing and subword initialization, it achieves excellent performance with relatively few training tokens. The model sequentially trains from input embeddings through to all parameters, effectively transferring the advanced capabilities of the original English model to Korean. As a result, EEVE exhibits outstanding performance in Korean and is recognized as a leading Korean model within the open-source community. The 2.8B version was fine-tuned as Phi-2, while the 10.8B version was meticulously fine-tuned as SOLAR-10.7B (Kim et al., 2024)

D Fine-Tuning

This study aimed to construct a language model that reflects the characteristics of historical Korean by applying a uniform fine-tuning approach to various pretrained models. To achieve consistency, data preprocessing and training processes were standardized while taking into account the architectural differences among models.

For fine-tuning, the dataset constructed from Korean literary works of the 1920s–1930s was employed. This dataset comprises 1,145 novels by 56 authors, from which approximately 5 million tokens were extracted through preprocessing. The dataset was subsequently split into training (80%), validation (10%), and test (10%) sets. Tokenization was performed using AutoTokenizer, and an appropriate maximum sequence length was set for each model.

Fine-tuning was conducted in an unsupervised manner, ensuring that the models learned patterns from the historical text without explicit annotations. To optimize memory usage, LoRA was applied during training, which consistently enhances performance across various LLMs and datasets, particularly in complex reasoning tasks. LoRA scales effectively regardless of model size and achieves high efficiency with minimal computational resources (Hu et al., 2023). The training process was carried out on four GeForce RTX 3090 GPUs, enabling efficient parallelization. Despite varying model architectures, fine-tuning for each model was completed within 5 hours.

Common hyperparameters were set across all models: the learning rate was fixed at 2e-4, the optimizer used was AdamW, and a linear decay schedule was employed. Weight decay was set to 0.01, and the maximum training steps were capped at 1,000—adjusted as needed based on each model's characteristics. Batch sizes varied by model, with gradient accumulation steps set to 2 and warm-up steps to 5 to ensure training stability.

Model performance was monitored by saving checkpoints based on evaluation loss, and the model with the lowest evaluation loss was selected for the experiments.