scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Reweaving the Threads of Korean History: AI-Driven Restoration of the Daegu-bu Household Registers (1681–1876)

Seongmin Mun[1,5], Donggue Lee[2,5], Jane Yoo[2,3]✉ & Sangkuk Lee[2,4]✉

In this study, we have applied advanced masked language models (MLMs)—BERT, DistilBERT, ELECTRA, and RoBERTa—to infer missing and misinterpreted values in comprehensive family register data. Our data compiles Daegu-bu household register books, triennially published from 1392 to 1910 in Joseon (premodern Korea) for listing the demographic characteristics of taxpayers. MLM models primarily detect and infer transcription errors and blanks caused by the historical deterioration of original books. The results show that RoBERTa outperforms the other models using byte-pair encoding tokenization. The accuracy rates of non-RoBERTa models have been found to deteriorate with an increase in the number of variables. The performance of RoBERTa in inferring morphologically rich and sparse data extends the frontiers of historical research by reconstructing primary sources for understanding the past.

Background & Summary

Historical document restoration is a meticulous process involving the rehabilitation of structures, subjects, and contents of historical and cultural significance, with the primary goal of returning these elements to their original or near-original state, thereby reviving their authenticity and historical integrity. Recently, more historians are employing artificial intelligence (AI)-based technologies to expedite the restoration of old materials and ensure that the restored details are contextually congruent with historical records. This study examines the efficacy of an AI-based approach for a specific historical restoration task: inferring missing or omitted values of multiple variables from the partial information within historically deteriorated documents. Such tasks are similar to predicting missing words or tokens in a sentence based on surrounding context, a principal function of masked language models (MLMs). Consequently, this study investigates whether the contextual understanding capabilities of advanced MLMs are applicable to the restoration of historical data.

We restore the household registers (Hojeokdaejang) of the Joseon Dynasty (1392–1910). Hojeokdaejang served as a demographic record and a critical administrative tool for taxation, conscription, and social control, formalized through the implemented triennial household survey system[1–3]. These registers are particularly insightful sources as they document individuals across the social spectrum. Each register detailed demographic information for families based on the household unit (Ho, 戶), where (Ho, 戶) is typically organized around a marital relationship comprising a male householder (Juho, 主戶), his wife, and other family members. For example, the social status and occupation of individuals were recorded to assess their capacity as potential taxpayers and military personnel. The documents were compiled every three years across 334 administrative districts (Gun–Hyeon, 郡縣) in the eight provinces of Joseon, totaling 187 volumes, approximately 44,300 pages, and 1.7 million data cells in spreadsheets. The historical significance of these registers is widely acknowledged by historians[4–8]. For example, diverse occupational roles (Jikyeok, 職役), reflecting one's social and familial background, represent a projection of the social structure. The longitudinal data pertaining to each family member are also highly valued by historical demographic researchers whose primary research often involves tracing

[1]Department of English Language and Literature, Kyungpook National University, Daegu, Republic of Korea. [2]Humanities Research Institute, Ajou Universty, Suwon-si, Republic of Korea. [3]Department of Financial Engineering, Ajou University, Suwon, Republic of Korea. [4]Department of History, Ajou University, Suwon, Republic of Korea. [5]These authors contributed equally: Seongmin Mun, Donggue Lee. ✉e-mail: janeyoo@ajou.ac.kr; okllsskh@ajou.ac.kr

genealogical links. The comprehensive reconstruction and linkage of such cross-sectional data have long posed a significant challenge in studying social dynamics and the transition from premodern to modern society[9].

Among extant registers, the Daegu-bu register is notable for its data quality. Daegu, as the capital of Gyeongsang-do—one of the Joseon Dynasty's eight provinces—administered 30 myeons, which encompassed both urban and rural areas. It documented households and individuals from diverse classes and occupations. Unlike a smaller rural Gun-Hyeon (郡縣), Daegu-bu functioned as a significant political and economic hub in Gyeongsang-do, containing government offices, military outposts, and a vibrant merchant class. The Hojeokdaejang in this region reflected a highly stratified society comprising urban dwellers, landowning elites, merchants, and military personnel, thus constituting a particularly valuable dataset for studying social mobility and economic activity within the provincial capital[10].

Despite being a critical source in Korean historical studies, systematic analysis of the Daegu-bu register is constrained by the substantial volume and complexity of the records. Moreover, its utility for studying social mobility is impeded by several additional factors. First, only a limited number of household registers remain after significant reforms and alterations to the population policy in 1895, compounded by losses during the Japanese colonial period and the Korean War[11]. Second, because the source materials are handwritten by various scribes across different periods and records, calligraphy lacks standardization, thereby presenting analytical challenges. Third, the level of recorded details varies among individuals, with limited information available for household members other than the primary householder and his wife. Although primary householders and their spouses have been documented more comprehensively, knowledge concerning other family members remains inherently limited. However, these non-householders often became new primary householders over time, inheriting the economic and intangible assets of their predecessors. Therefore, the considerable data volume, fragmented information, and irregular transcription errors cannot be easily resolved using traditional methods.

To overcome the difficulties of using large-scale data, researchers at the Daedong Institute for Korean Studies at Sungkyunkwan University began digitizing Hojeokdaejang in the early 2000s[12]. It began with Danseong-hyeon in the early 2000s, followed by the large-scale digitization of the Daegu-bu Hojeokdaejang, a provincial capital, in the 2010s. The first phase of digitization involved categorizing the original records into 72 variables and converting them into structured Excel datasets, with the records of an individual assigned to a single cell[13]. The initial phase of digitization prioritized preserving the original text. Despite efforts to preserve the original texts as accurately as possible, digital data included inconsistencies across books, missing/omitted values, and transcription errors. Furthermore, erratic transcription errors and physical eradication add complexity to the understanding of this dataset.

To the best of our knowledge, this is the first attempt to apply MLMs to restore individual-level longitudinal demographic data by leveraging fragmented and scattered information. Based on the accuracy rates of targeting the original text, RoBERTa outperforms BERT, DistilBERT, and ELECTRA. Given the same hyperparameters and simulation environment, RoBERTa accurately infers more missing values based on byte-pair encoding (BPE) tokenization. Specifically, the overall performance of the non-RoBERTa models deteriorates exponentially with an increase in the number of variables. The key strength of MLMs for historical restoration lies in their ability to efficiently process morphologically rich and sparse data.

Methods

Data cleaning. To address the aforementioned issues and enhance the usability of the dataset, this study employs a two-step methodology: (1) comprehensive data cleaning to ensure internal consistency and (2) AI-driven analysis to automate missing data reconstruction. The initial phase involved categorizing the household register data into 72 variables and converting them into an Excel database. Additional refinements were performed to ensure usability of the dataset for research applications. The refinement process was significant as the groundwork for MLM designs to process data from historical perspectives.

The raw household register data were originally digitized and curated by the Daedong Institute for Korean Studies at Sungkyunkwan University. Our research team obtained official authorization from the Institute to refine and publish the dataset for research purposes. The formal document is deposited together with the dataset in the Open Science Framework (OSF) repository (OSF, https://doi.org/10.17605/OSF.IO/A8MH3).

We refined the dataset to ensure consistency of data across books and usability in both qualitative and quantitative analyses. Three rules for data cleaning were established: (1) standardizing the transcription of proper nouns (place names, personal names, and official titles); (2) correcting input errors; and (3) supplementing the data by cross-referencing with other historical sources.

The first step in data cleaning involved addressing inconsistencies in the Chinese character transcriptions. For example, the surname "金" was primarily transcribed as "Kim," but in certain cases, it appeared as "Geum." As the same Chinese character can have multiple pronunciations in Korean, standardization was necessary. To resolve this problem, transcriptions of the same individuals and their family members were sorted and examined based on contextual relevance. Additionally, discrepancies in the names, official titles, and ages of the same individual across different record years were manually corrected.

After the first step, we rectified the errors caused by digitalization. Specifically, variant Chinese characters (Icheja, 異體字) and different calligraphic styles were detected and sorted. We also corrected the transcription errors made by the scribes. Recovering missing entries from faded, damaged, and contaminated documents required meticulous effort, as detecting blanks were intentionally left for certain reasons. For example, place names were expressed in different versions. Different Chinese characters were used for naming the same place, like "羅州 (Naju)" written as "那州" and "忠州 (Chungju)" as "衷州." These discrepancies arose because of the handwritten nature of these records. To resolve these issues, we applied a consistent framework to unify place names and correct erroneous Chinese-character transcriptions.

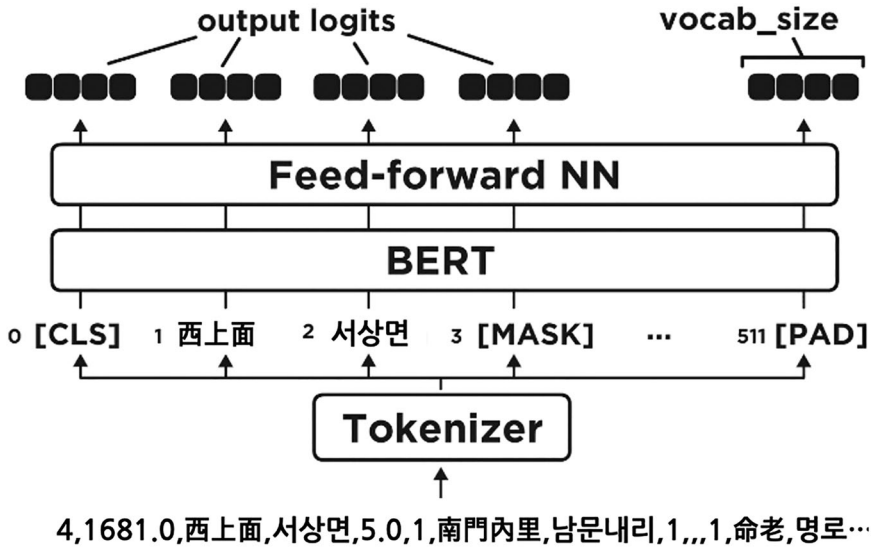

Fig. 1 Masked language modeling process using BERT for historical family register data.

To further enhance data accuracy, supplementary historical sources were consulted to validate and supplement missing entries. Cross-referencing with the Jokbo (Korean genealogies) and household registers from nearby Daegu, including Danseong and Ulsan, was conducted to verify the records of individuals and families[14]. The Dictionary of Ancient Script Usage, provided by the Korean Historical Manuscript Center, was consulted to confirm the correct forms and pronunciations of specific Chinese characters. Through this preprocessing, researchers and assistants refined incomplete records and ensured the continuity of family records. The digitized dataset of the Daegu-bu Hojeokdaejang became more structured and consistent, reducing transcription errors and improving cross-record compatibility. This process provided a reliable foundation for subsequent AI-driven analyses and automated data refinement, ultimately expanding the potential of large-scale computational studies in historical research.

The household register dataset (CleanedData.csv) was constructed primarily from the Excel files prepared by the Daedong Institute for Korean Studies, which provided the foundation for the digitized records employed in this study. Genealogical materials (Jokbo) were not incorporated into the deposited dataset; rather, they were referred to only in limited instances during the data cleaning process, chiefly to verify or supplement missing and corrupted information such as surnames, given names, and places of origin. For this purpose, we drew upon the Korean Historical Figures Information System (KHFIS), including Bangmok (examination registers), maintained by the Academy of Korean Studies[15], and, where necessary, also referred to genealogical collections that survive predominantly in image form, such as those preserved in the Genealogy Library at Inje University.

Model Creation. Despite the comprehensive preprocessing, numerous entries in the Daegu-bu family register remained missing or incomplete. This issue is characteristic of historical records, particularly those spanning extensive periods and multiple generations. Researchers have attempted to correct these errors and ensure data consistency. However, because the dataset is large and complex, it is nearly impossible to manually fill in every missing piece.

We used a modern computer-based method to automatically fill in the missing information. We used a type of AI called the large language model (LLM). These models are trained on large amounts of text and can guess missing words based on context. A special type of LLM, called the MLM, is particularly suitable for this task. MLMs work by hiding one word in a sentence and training the model to predict the hidden words by looking at the remainder of the sentence. This makes them useful in estimating missing names, jobs, places, or family relationships in historical records.

Several well-known MLMs were tested in this study. A popular one is the BERT[16]. BERT understands the meaning of words in a sentence by examining both the left and right sides of the masked words. We have also used a distilled version of BERT[17], which is a smaller and faster version that still performs well. Furthermore, we have employed ELECTRA[18], which adopts a distinct training method by replacing words and checking if the model can determine real versus fake ones. Finally, we have used RoBERTa[19], a version of BERT that has been trained for longer periods and on more data, making it more accurate.

All these models were trained in the same setting based on the number of layers, hidden sizes, and sequence length. We trained each model to learn from the structure and patterns in the Daegu-bu register and predict missing entries. Figure 1 shows how BERT-based MLMs learn information. In the Daegu-bu family register, each individual is recorded with 72 different pieces of information. The MLM randomly hides one of these pieces by replacing it with a special token, called [MASK]. As shown in the figure, the model hides the person's order number in the register and learns to predict it.

Through this process, the model better understands the overall structure and content of the family register. Table 1 summarizes the training settings for each model, as suggested by previous studies[20].

	BERT	DistilBERT	ELECTRA	RoBERTa
Tokenization	**WordPiece**	**WordPiece**	**WordPiece**	**Byte-pair encoding (BPE)**
Model parameters	Hidden activation: GeLU Vocab size: 52,000 Batch size: 32 Sequence length: 512 Dropout rate: 0.1 Attention dropout: 0.1 Learning rate: 3e-5 Epochs: 5 Hidden size: 768 Attention heads: 12 Hidden layers: 6 Train Steps: 5,000			

Table 1. Pretraining hyperparameters of each MLM model.

All four models were based on BERT and shared several training parameters such as batch size, learning rate, and number of layers. However, unlike other models, RoBERTa uses BPE instead of WordPiece in the token generation stage. We used a server to generate each MLM model, and the server specifications were as follows: NVIDIA Quadro RTX A5000 - Blower, 8,192 CUDA Cores, 256 Tensor Cores, and 24GB GDDR6 memory. The final models were saved in their respective folders and subsequently used for model evaluation.

Data Records

The clean version of the Korean Family Register Data and four MLMs created by our research team are available under a CC-BY at https://doi.org/10.17605/OSF.IO/A8MH3[21]. The dataset includes information on 1,758,301 individuals with 72 attributes (see Section 2.1). The data folder contains both the integrated original data from Sungkyunkwan University (RawData.csv) and the cleaned data processed by our research team (CleanedData.csv). Both datasets were provided in the CSV format using commas as delimiters and were compatible with a UTF-8 encoding environment. Detailed descriptions of the data variables are provided in the description.pdf file. Additionally, the TransformedData.txt file is a text format of CleanedData.csv and was used for training the MLMs. Four adjusted MLMs are available in the Model folder, and instructions for using each model can be followed using the ModelUsage.ipynb file, which runs on a Python Jupyter Notebook. For more detailed explanations of the dataset, including clarifications on Korean and Chinese characters appearing in the Daegu-bu household registers, please also refer to the description.pdf file available in the OSF repository.

Technical Validation

To restore historical information, we developed an MLM based on four distinct language models—BERT, DistilBERT, ELECTRA, and RoBERTa—trained on Korean family register data. To evaluate the effectiveness of these models, we analyzed their performance in inferring information across different family hierarchical levels, a structure inherent to the dataset. The family register data included variables that showed each one's position or role in a family, such as "*juho* (head of family, 328,693 individuals)," "*cheo* (wife, 260,859 individuals)," "*nobi* (servant, 448,989 individuals)," and "*other* (719,760 individuals)." This hierarchical structure provided a meaningful framework for assessing the inference capabilities of the models. For the model evaluation, we employed a method in which the AI models inferred information for variables masked with the [MASK] token. The evaluation process involved randomly selecting 1,000 variables across all data, masking their values with [MASK], and measuring the accuracy of the models in correctly predicting the masked values. If a model accurately inferred the correct value 500 times out of 1,000 trials, its accuracy was measured as 0.5. This procedure was repeated 10 times to ensure robustness, and the results were averaged for analysis.

RoBERTa consistently outperformed the other models across all hierarchical levels. Notably, while the other three models showed a marked decline in inference accuracy for individuals with a substantial amount of recorded information, such as *juho* and *cheo*, RoBERTa maintained a superior performance. In categories with more missing values per variable such as "*nobi*" and "other," the performance gap between RoBERTa and the other models narrowed significantly. This performance disparity can be attributed to several factors. First, individuals with fewer missing values have more variables to be inferred. Subsequently, we identified a higher degree of freedom for specifying the characteristics of these individuals. Models such as BERT, DistilBERT, and ELECTRA are more likely to struggle with managing the complexity of such a high degree of freedom per cell, leading to a lower accuracy compared with RoBERTa. Second, the difference in the structure of each model explains the performance variation. Unlike other models that use WordPiece tokenization, RoBERTa employs BPE for tokenization. Our results suggests that BPE is more likely to predict the value accurately than the others by virtue of its morphological richness and sparse contexts often found in historical datasets.

Figure 2 shows the performances of the four models across the *juho*, *cheo*, *nobi*, and *other* categories. The figure summarizes the results of the 10 evaluation trials, highlighting the effectiveness of RoBERTa and illustrating its robust handling of complex and diverse historical information.

The initial evaluation confirmed that RoBERTa consistently demonstrated superior performance in inferring information about individuals recorded in the Korean family register compared with the other three models. Additionally, all models exhibited a decline in inference accuracy for hierarchical levels with substantial amounts of information, such as *juho* (head of the family) and *cheo* (wife). This trend underscores how the uncertainty in predictions increases when additional information is processed.

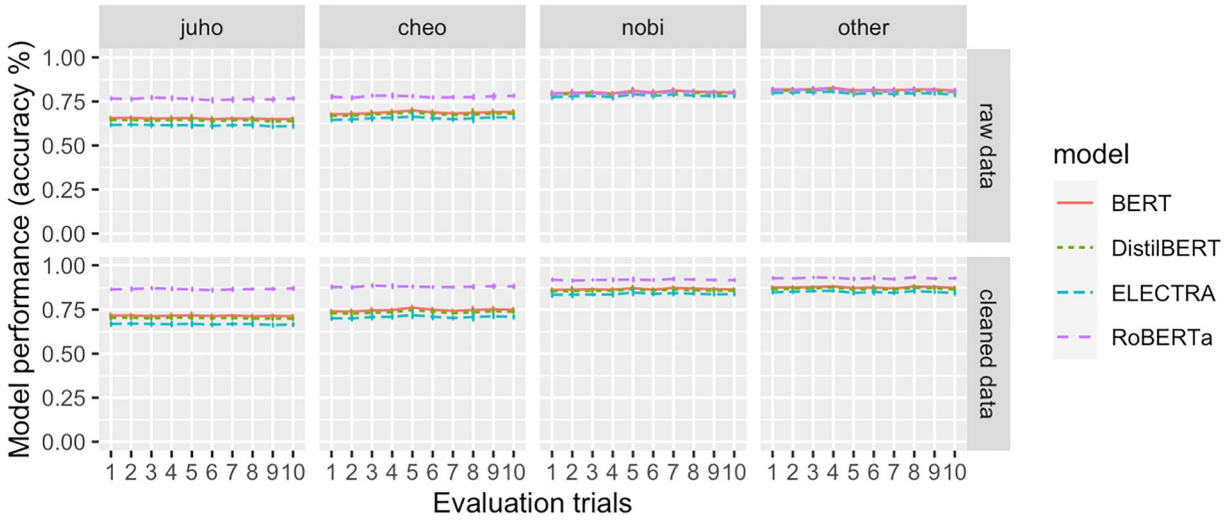

Fig. 2 Comparison of language model performance by family hierarchy. Note: x-axis: evaluation trials (1–10); y-axis: model performance (accuracy %). Panels show four family hierarchy categories (juho = household head, cheo = wife, nobi = slave, other = additional members) for both raw data (top row) and cleaned data (bottom row). Colored lines indicate models: red = BERT, green = DistilBERT, blue = ELECTRA, purple = RoBERTa. Results demonstrate that RoBERTa consistently achieves higher accuracy than the other models across all categories, while accuracy declines for roles with greater recorded information (juho, cheo), reflecting increased uncertainty with complex data.

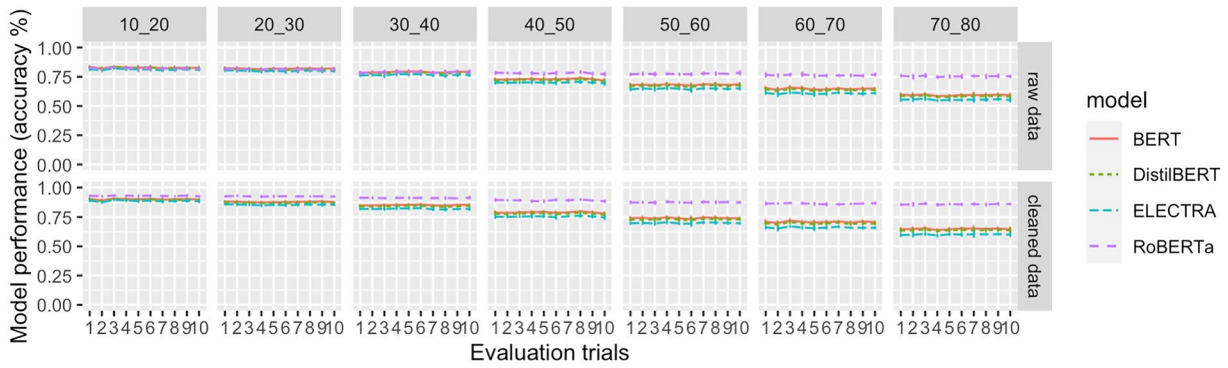

Fig. 3 Comparison of language model performance by variable information size. Note: x-axis: evaluation trials (1–10); y-axis: model performance (accuracy %). Panels show model accuracy across different levels of variable information (10–20% through 70–80%) for raw data (top row) and cleaned data (bottom row). Colored lines indicate models: red = BERT, green = DistilBERT, blue = ELECTRA, purple = RoBERTa. Results highlight the consistent superiority of RoBERTa across all information levels, while other models show substantial declines in accuracy as information density increases, reflecting the greater uncertainty introduced by more complex data.

To further explore this phenomenon, we conducted an additional test to evaluate how the models responded to varying amounts of information rather than hierarchical categories. Figure 3 shows the results. The results are given by the performance of the models for nine categories of data based on the amount of variable information associated with individuals. Specifically, the information availability per variable ranged from 10% to 100% per group. We excluded groups with barely available information (0–10%) or almost full information (ranged between 80%–90% and 90%–100%) For each group, the same evaluation procedure as the previous analysis was applied, which involved masking variable values and assessing the accuracy of the model in predicting the correct information.

The results of this evaluation revealed a clear pattern: as the amount of information associated with individuals increased, the inference accuracy of all models decreased. Moreover, for the three models other than RoBERTa, the sharpest decline in accuracy was observed with variable information filled at approximately 40%–50%. This finding highlights the challenge posed by larger datasets with more complex choices, as increased data introduce greater uncertainty in the inference process. The results also emphasize the significance of the model architecture and algorithmic features in influencing the inference performance. The relative robustness of RoBERTa across varying information levels suggests that its use of BPE tokenization, rather than the WordPiece tokenization used by the other models, facilitates its handling of diverse and sparse historical contexts.

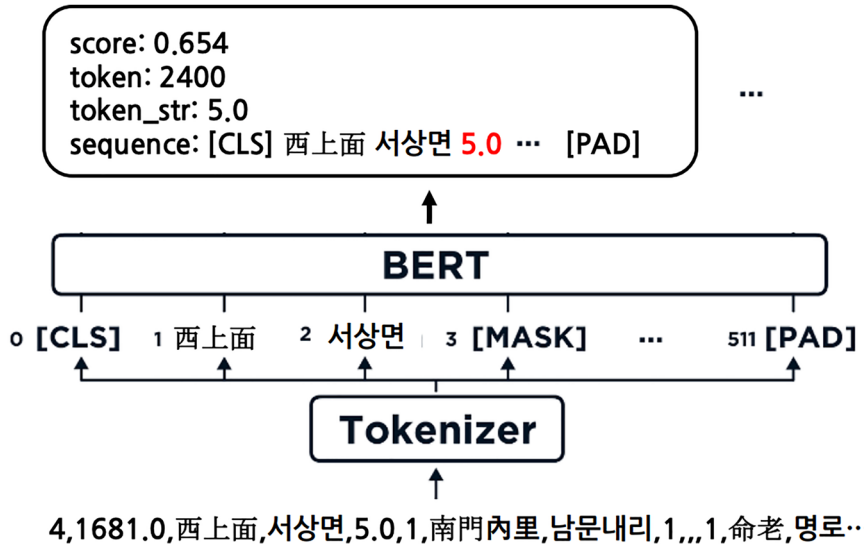

Fig. 4 Example of AI model predictions for a masked token.

Figure 3 shows these findings and presents the performances of all four models across different information levels. The graph highlights the consistent superiority of RoBERTa, along with the substantial decrease in performance observed in the other models as the information density increased. This analysis reaffirms the importance of tailored approaches when leveraging AI for historical data restoration. Understanding the interaction of model structures and data complexities is crucial for enhancing the inference accuracy and unlocking the full potential of historical datasets.

In addition, our results indicate that the AI models tend to achieve higher prediction accuracy when less information is available for a given individual. This can be interpreted as a result of increased uncertainty: as more information is provided, the number of possible values that could fill a masked [Mask] position also increases. Furthermore, as illustrated in Fig. 4, the models do not generate a single prediction but instead provide multiple possible outputs with associated information, including a score (confidence level, summing to 1 across all candidates), a token (the index of the predicted token), a token_str (the actual predicted string), and a sequence (the reconstructed output combining the predicted token with contextual information).

In this study, we evaluated accuracy by using only the predictions with the highest probability. However, depending on their research goals, scholars may apply alternative strategies. For example, they might (1) adopt a threshold rule, using predictions only if their probability exceeds a given cutoff, or (2) randomly select among the top three predictions ranked by probability. Such methodological choices could lead to different outcomes, underscoring the importance of transparent reporting and careful interpretation when working with AI-inferred historical data.

Usage Notes

The clean data created in this study integrate information that can be recorded differently by each researcher and can be used without issues if the encoding is set to UTF-8. Additionally, the MLMs created in this study can be used by following the procedures recorded in the ModelUsage.ipynb file in the Model folder. These models are used to predict missing values owing to time or missing information. Finally, if one wishes to replicate the methods of this study, the IPYNB files can be used for the four models included in the Code folder. To ensure smooth model training, it is recommended that the code be run in an environment that meets or exceeds the specifications of the server used in this study (see Section 2.2).

In addition to the technical guidance provided above, the restored Daegu-bu household register dataset allows systematic investigation of intergenerational mobility by tracing the transmission of status and occupation across household units. At the same time, its detailed occupational classifications enable analyses of the composition of artisans, merchants, military personnel, and landowning elites, thereby illuminating the economic activities of a provincial capital. Moreover, the strengthened documentation of maternal lineages and non-householder members makes it possible to examine family structures and kinship ties that are often obscured in demographic studies. Thus, the dataset demonstrates its potential for broad reuse and expands the scope of comparative and interdisciplinary research on premodern Korean society.

Against this backdrop, prior work using Daegu-bu and other local registers—many of which are synthesized in *Joseon wangjo hojeok, saeroun yeongu bangbeomnon-eul wihayeo* [Joseon Dynasty Household Registers: Toward a New Research Methodology][22]—has advanced our understanding of political, economic, social, cultural, military, and demographic patterns in specific locales and periods. Yet, because most datasets were manually compiled, analyses often remained limited to a single occupational group, a small administrative unit (myeon/dong), or a short temporal span.

For example, prior case studies examined changes in status-group composition within one or several subdistricts (myeon), shifts in surname–bon'gwan distributions during clan-village formation, and late-nineteenth-century

household entries and exits. In contrast, our restored Daegu-bu dataset and versioned code pipeline provide city-wide coverage across subdistricts and social groups, a long temporal span (1681–1876), and harmonized variables (e.g., standardized occupations and origin/bon'gwan), together with explicit correction flags and audit logs. These features enable researchers to track recorded groups over multiple decades, link physical and social mobility to identifiable events (e.g., institutional reforms, crises), and reconstruct the dynamics of socioeconomic change in a provincial capital with a degree of scale, consistency, and reproducibility that was previously difficult to achieve.

Data availability

The dataset is available online at the OSF repository [https://doi.org/10.17605/OSF.IO/A8MH3]. It contains information on 1,758,301 individuals with 72 attributes, including both the original data from Sungkyunkwan University (RawData.csv) and the cleaned version processed by our team (CleanedData.csv). All files are in CSV format (UTF-8 encoded), and detailed descriptions of the variables are provided in description.pdf.

Code availability

The code we used to generate the four masked language models is publicly available through the following OSF repository: https://doi.org/10.17605/OSF.IO/A8MH3.

Received: 9 July 2025; Accepted: 10 November 2025;
Published online: 18 December 2025

References

1. Palais, J. B. *Confucian Statecraft and Korean Institutions*. University of Washington Press (1996).
2. Deuchler, M. *The Confucian Transformation of Korea: A Study of Society and Ideology*. Harvard University Asia Center (1992).
3. Peterson, M. *Korean Adoption and Inheritance: Case Studies in the Creation of a Classic Confucian Society*. Cornell University East Asia Program (1996).
4. Hojŏk taejang yŏn'gu t'im. *Tansŏng hojŏktaejang yŏn'gu*. Sungkyunkwan University Daedong Institute for Korean Studies (2003).
5. Park, H. & Lee, S. A survey of data sources for studies of family and population in Korean history. *The History of the Family* **13**, 258–267 (2008).
6. Son, B. & Lee, S. The effect of social status on women's age at first childbirth in late seventeenth- to early eighteenth-century Korea. *The History of the Family* **15**, 430–442 (2010).
7. Lee, S. & Son, B. Long-term patterns of seasonality of mortality in Korea from the seventeenth to the twentieth century. *Journal of Family History* **37**, 270–283 (2012).
8. Kim, K., Park, H. & Jo, H. Tracking individuals and households: longitudinal features of Danseong household register data. *The History of the Family* **18**, 378–397 (2013).
9. Kwon, N. Digitalization of and new research trends in Hojeokdaejang from the late Joseon period. *The Review of Korean Studies* **7**, 229–246 (2004).
10. Lee, D. & Han, S. Families in the household registers of seventeenth-century Korea: capital, urban and rural areas. European. *Journal of Korean Studies* **20**, 1–34 (2020).
11. Hwang, K. M. Citizenship, social equality and government reform: changes in the household registration system in Korea, 1894–1910. *Modern Asian Studies* **38**, 355–387 (2004).
12. Paek, S., Park, J. H. & Lee, S. Building an archival database for visualizing historical networks: a case for pre-modern Korea. *Historical Life Course Studies* **12**, 42–57 (2022).
13. Daedong Institute for Korean Studies. *DataBase*. Daedong Institute for Korean Studies, https://ddmh.skku.edu/ddmh/db/intro.do (2018).
14. Son, B. In'gu-sa jŏk ch'ŭngmyŏn esŏ pon hojŏk kwa chokpo ŭi charyo chŏk sŏnggyŏk—17–19 segi Kyŏngsang-do Tansŏng-hyŏn ŭi hojŏk taejang kwa Hapchŏn Yi-ssi ga ŭi chokpo. *Taedong munhwa yŏn'gu* **46**, 79–109 (2004).
15. Academy of Korean Studies. *Korean Historical Figures Information System (KHFIS)*. Academy of Korean Studies, https://people.aks.ac.kr/front/main/main.do (2025).
16. Devlin, J., Chang, M., Lee, K. & Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint* arXiv:1810.04805 (2018).
17. Sanh, V., Debut, L., Chaumond, J. & Wolf, T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint* arXiv:1910.01108 (2019).
18. Clark, K., Luong, M., Le, Q. & Manning, C. D. ELECTRA: Pre-training text encoders as discriminators rather than generators. *Int. Conf. Learn. Represent*. (ICLR) (2020).
19. Liu, Y. *et al*. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint* arXiv:1907.11692 (2019).
20. Pires, T., Schlinger, E. & Garrette, D. How multilingual is multilingual BERT? *Proc. 57th Annual Meeting of the Association for Computational Linguistics*, 4996–5001 (2019).
21. Mun, S., Lee, D., Yoo, J. & Lee, S. *Daegu-bu Household Register Dataset (1681–1876)*. Open Science Framework, https://doi.org/10.17605/OSF.IO/A8MH3 (2025).
22. Son, B. Joseon wangjo hojeok, saeroun yeongu bangbeomnon-eul wihayeo [Joseon Dynasty Household Registers: Toward a New Research Methodology]. Sungkyunkwan University Press, Seoul (2020).

Acknowledgements

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2022S1A5C2A02090368).

Author contributions

S.K.L. and J.N.Y. conceived and supervised the overall design and development of the manuscript. D.G.L. was responsible for describing and refining the dataset. S.M.M. developed the AI model for historical data inference using the refined dataset and evaluated the model. All authors contributed to writing and revising the manuscript and approved the final version.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to J.Y. or S.L.