https://doi.org/10.1038/s44401-025-00036-1

Protecting medical privacy with pretrained language models: named entity recognition-based de-identification in Korean unstructured electronic medical records

Check for updates

Jiahn Seo[1], Yunha Kim[1,2], Heejung Choi[1], Minkyoung Kim[1], JiYe Han[1], Gaeun Kee[1,2], Soyoung Ko[1,2], HyoJe Jung[1], Byeolhee Kim[1,2], Boeun Choi[1,2], Sanghyun Park[3], Tae Joon Jun[4,6] ✉ & Young-Hak Kim[5,6] ✉

This study introduces a lightweight de-identification software tailored to the Korean healthcare environment, where heterogeneous document formats and limited computing resources hinder clinical data integration. The proposed system protects sensitive patient information while preserving essential clinical content—such as diagnoses, surgical schedules, and prescriptions—for research purposes. We defined de-identification categories specific to the Korean context and implemented a preprocessing pipeline optimized for discharge summaries containing mixed Korean, English, and special characters. To address the lack of high-quality Korean Named Entity Recognition (NER) datasets, we applied Korean-specific data augmentation and fine-tuned the Korean Language Understanding Evaluation (KLUE) Bidirectional Encoder Representations from Transformers (BERT) model to enhance generalizability. Model performance was compared against lightweight Korean large language models (LLMs), including Llama3-Open-Ko-8B and EEVE-Korean-Instruct-10.8B. The KLUE BERT model, trained on an augmented dataset, achieved an F1 score of 91.42% on the internal validation set and maintained 94.30% on real discharge summary data. Notably, it outperformed LLMs in recognizing more than 200 categories of sensitive entities, demonstrating superior performance. This compact solution offers a scalable and privacy-preserving approach for anonymizing electronic medical records (EMRs) in clinical settings.

The digitization of Electronic Medical Records (EMRs) has dramatically improved the efficiency and accuracy of clinical practice, establishing itself as a core infrastructure of modern healthcare systems. EMRs contain not only sensitive personal information such as patient names, dates of birth, and medical histories, but also detailed clinical data, including patient health status and medical progress. As such, EMRs have become an invaluable resource for research and analysis, supporting efforts in disease pattern analysis, prognosis prediction, and novel drug development[1]. Notably, the potential of EMRs now extends beyond individual institutions, with growing international initiatives—such as those led by the European Union

[1]Department of Information Medicine, Asan Medical Center, 88, Olympicro 43gil, Songpagu, 05505 Seoul, Republic of Korea. [2]Department of Medical Science, University of Ulsan College of Medicine, 88, Olympicro 43gil, Songpagu, 05505 Seoul, Republic of Korea. [3]Department of Computer Science, Yonsei University, 50, Yonsei-ro, Seodaemungu, 03722 Seoul, Republic of Korea. [4]Department of Medical Informatics and Statistics, Asan Medical Center, University of Ulsan College of Medicine, 88, Olympicro 43gil, Songpagu, 05505 Seoul, Republic of Korea. [5]Division of Cardiology, Department of Internal Medicine, Asan Medical Center, University of Ulsan College of Medicine, 88, Olympicro 43gil, Songpagu, 05505 Seoul, Republic of Korea. [6]These authors contributed equally: Tae Joon Jun, Young-Hak Kim. ✉e-mail: taejoon@chilab.kr; mdyhkim@amc.seoul.kr

—aiming to establish integrated platforms for cross-border healthcare data interoperability. This expansion underscores EMRs as a driving force behind medical innovation[2].

While the integrated use of medical datasets enables advancements in personalized treatment and precision medicine, the protection of patient privacy remains a major challenge[3]. Despite their clinical value, EMRs are structurally vulnerable to various cybersecurity threats, including hacking, viruses, worms, data breaches, and theft[4,5]. Recently, concerns have intensified with the emergence of prompt-based adversarial attacks using Large Language Models (LLMs), which pose a novel risk of exposing sensitive personal information embedded in clinical narratives[6]. Such vulnerabilities make the presence of personally identifiable information in EMRs a significant barrier to data sharing and secondary analysis. In response, regulatory authorities such as the European Medicines Agency (EMA) have recommended the transparent disclosure of clinical trial information[7]. Consequently, the development and refinement of de-identification technologies has become a fundamental prerequisite for the secure and ethical use of EMR data in healthcare research[8].

In line with the global trend toward data utilization, South Korea enacted the Personal Information Protection Act (PIPA) in 2011 to strike a balance between individual privacy protection and the use of data for research and innovation[9]. In 2020, the revision of Korea's "Three Data Laws" further expanded the legal framework by enabling the processing and utilization of pseudonymized information, thereby facilitating Artificial Intelligence (AI)-driven medical research and large-scale data analysis. Similar to the United States' Health Insurance Portability and Accountability Act (HIPAA), which defines a "Safe Harbor" standard requiring the removal of 18 types of Protected Health Information (PHI) to consider clinical data de-identified[10], Korea has established official Guidelines for the Utilization of Health and Medical Data. These guidelines aim not only to ensure compliance with privacy protection principles but also to promote national strategies for advancing precision medicine and data-driven healthcare innovation[11].

Korean EMRs consist of a complex mixture of Korean, English, and special characters, posing significant challenges to the direct application of conventional English-based Named Entity Recognition (NER) models. In particular, the lack of high-quality training datasets for Korean NER has often necessitated manual annotation or review by domain experts and engineers, resulting in a time-consuming and resource-intensive process. In contrast, a variety of specialized NER technologies and models have been actively developed for English and Chinese[12–14], with Deep Learning (DL) approaches—especially those based on self-attention mechanisms—commonly adopted for NER tasks[15]. Several studies have proposed de-identification techniques that are effective in multi-institutional settings using 500 clinical records[16], while others have explored end-to-end frameworks that integrate multiple DL models to enhance de-identification performance[17]. For the Korean language, a Bidirectional Encoder Representations from Transformers (BERT) model applying WordPiece tokenization was trained using data from online question-and-answer platforms[18], and domain-specific Korean medical corpora have also been used to build specialized NER models for medical entity recognition[19,20]. Nevertheless, rule-based methods such as regular expressions continue to face limitations in accuracy and scalability, making them inefficient and impractical for application to real-world clinical data[21,22].

This study aims to develop de-identification software tailored to the Korean healthcare environment, enabling the protection of sensitive patient information while retaining essential clinical content—such as diagnoses, surgical schedules, and medication prescriptions—for research purposes. In practice, it is challenging to implement a universal de-identification tool across medical institutions due to heterogeneous document formats across hospitals and departments, as well as limited computing resources in clinical settings[23]. Although the theoretical use of LLMs for EMR de-identification via prompt-based input has been proposed, uploading sensitive medical data to external commercial platforms raises serious security and legal concerns under privacy protection laws. To address this, we designed a lightweight LLM-based de-identification tool that can be deployed locally on consumer-grade hospital hardware. The developed software was validated using discharge summary data and compared against open LLMs. The model trained with augmented data demonstrated the best performance. These results confirm that high-accuracy de-identification of real-world clinical documents is feasible even under resource-constrained conditions, and that the proposed model serves as a practical and effective tool for EMR-based NER tasks.

Results

Performance of models by entity category

We used two base models: the Korean Language Understanding Evaluation (KLUE) BERT[24], a BERT architecture specialized for the Korean language, and bert-base-multilingual-cased[25], a multilingual BERT model. The training datasets consisted of the Base dataset, which includes the Naver dataset and additional institution names from Seoul Asan Medical Center, and the Base+AD dataset, which adds augmented data to the Base dataset. Using these, we developed a total of four models. Table 1 presents the performance evaluation results by entity category for each model.

Among the four models, the KLUE–BERT trained on the Base + AD dataset achieved the highest precision, recall, and F1 scores across categories,

Table 1 | Performance comparison of fine-tuned BERT-based models

	BERT-base-multilingual-cased			KLUE BERT		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Base			0.86			0.88
PER	0.86	0.87	0.87	0.91	0.88	0.89
ORG	0.86	0.84	0.85	0.89	0.87	0.85
LOC	0.83	0.83	0.83	0.82	0.87	0.85
DAT	0.89	0.92	0.90	0.88	0.93	0.90
TIM	0.70	0.87	0.78	0.83	0.86	0.84
Base + AD			0.89			**0.91**
PER	0.91	0.90	0.90	0.90	0.82	**0.91**
ORG	0.88	0.89	0.88	0.92	0.91	**0.92**
LOC	0.84	0.89	0.87	0.82	0.89	**0.90**
DAT	0.91	0.93	0.92	0.93	0.93	**0.93**
TIM	0.84	0.91	0.87	0.92	0.85	**0.88**

BERT bidirectional encoder representations from transformers, *KLUE* Korean language understanding evaluation.

The bold values represent the highest F1-score within each model category. Since the F1-score evaluates the overall performance of labels without bias toward any specific label, it is an important performance metric in NER.

with an F1 score of 0.91. This represents an improvement of approximately 0.05 points compared to the model trained with bert-base-multilingual-cased on the Base dataset.

Performances of KLUE BERT

Among the four models, the KLUE BERT model fine-tuned with the augmented Base + AD dataset demonstrated the highest performance. This model was evaluated on a manually de-identified test dataset derived from real discharge summaries extracted and curated from Seoul Asan Medical Center (AMC). For evaluation, 30% of the test data was reserved, and training was conducted over five epochs. The model achieved F1 scores of 93.87%, 93.54%, 94.30%, 93.71%, and 93.31%, with an average F1 score of 93.74% (Table 2).

Table 2 | De-identification performance of KLUE BERT with augmented data

	KLUE BERT					
	Epoch1	Epoch2	Epoch3	Epoch4	Epoch5	Average
F1-score	93.87%	93.51%	**94.30%**	93.71%	93.31%	93.74%

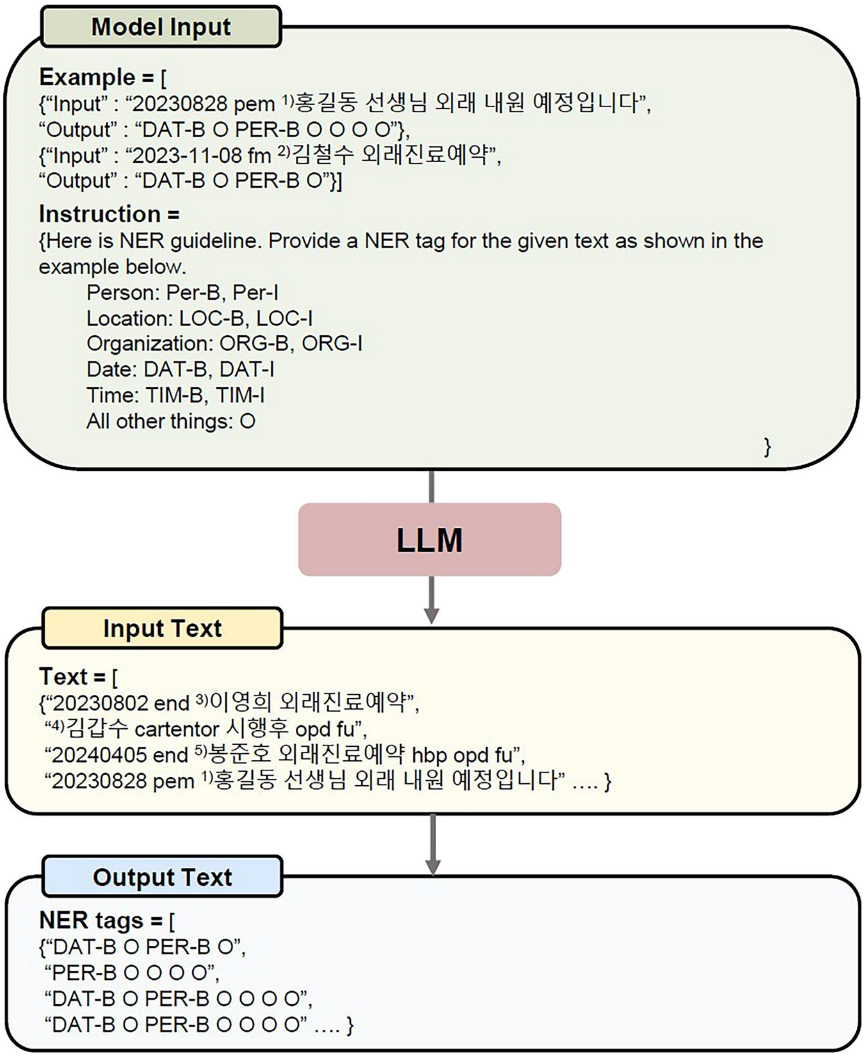
KLUE Korean language understanding evaluation.
The bold values represent the highest F1-score achieved at the epoch with the highest performance in KLUE BERT.

Comparison of LLM performance

To compare our proposed model with open-source LLMs, we applied a manually de-identified discharge summary dataset from Seoul AMC to the following models: Llama3-Open-Ko-8B[26], EEVE-Korean-Instruct-10.8B[27], and Llama3-8B-Instruct[28]. We effectively instructed open LLMs using few-shot prompting combined with Chain-of-Thought (CoT) prompting[29], along with prompt instructions that included multiple few-shot examples. As shown in Fig. 1, this approach guided each LLM to perform personal information anonymization. Notably, both Llama3-Open-Ko-8B and EEVE-Korean-Instruct-10.8B were able to accurately tag entity categories for each token as instructed. However, the multilingual Llama3-8B-Instruct model failed to properly understand the discharge summary content and produced inaccurate tagging results. Therefore, this model was excluded from the final performance comparison due to its unsuitability for EMR de-identification.

The tagging outputs of all models were assessed using this same procedure, and the proposed model's NER tagging performance was compared against that of publicly available LLM models. Specifically, KLUE-BERT, Llama3-Open-Ko-8B, and EEVE-Korean-Instruct-10.8B were evaluated against a manually annotated test set. A score of "1" was assigned when both the number and labels of the tagged tokens exactly matched the reference, and "0" was assigned otherwise.

As shown in Fig. 2, KLUE BERT achieved the highest performance among the baseline test dataset of 1000 samples, with a total of 583 matching tags, which represents the largest overlap with the manually annotated



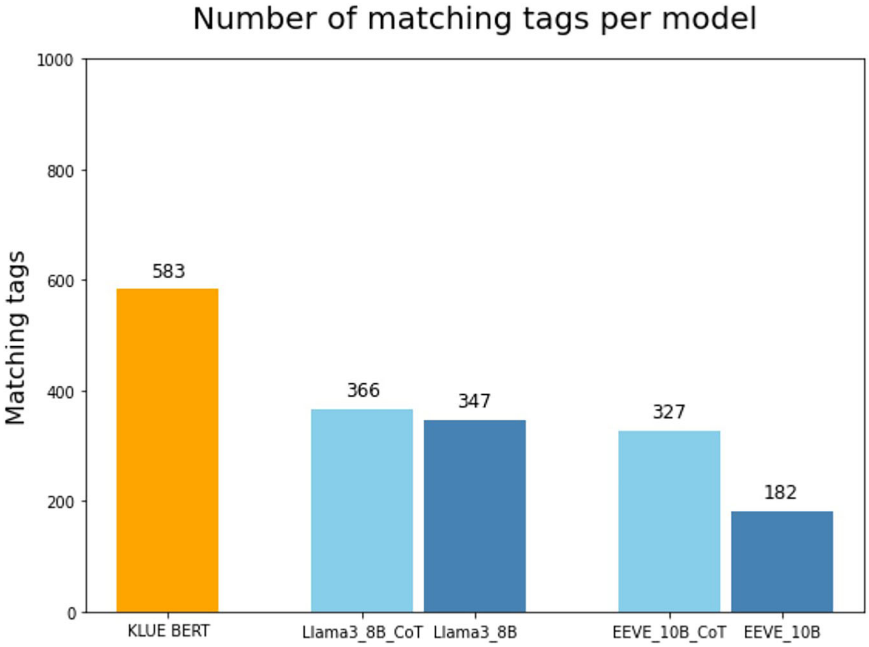
```
Example = [
{"Input" : "20230828 pem 1)홍길동 선생님 외래 내원 예정입니다",
"Output" : "DAT-B O PER-B O O O O"},
{"Input" : "2023-11-08 fm 2)김철수 외래진료예약",
"Output" : "DAT-B O PER-B O"}]
Instruction =
{Here is NER guideline. Provide a NER tag for the given text as shown in the
example below.
    Person: Per-B, Per-I
    Location: LOC-B, LOC-I
    Organization: ORG-B, ORG-I
    Date: DAT-B, DAT-I
    Time: TIM-B, TIM-I
    All other things: O
                                                                    }

Text = [
{"20230802 end 3)이영희 외래진료예약",
 "4)김갑수 cartentor 시행후 opd fu",
 "20240405 end 5)봉준호 외래진료예약 hbp opd fu",
 "20230828 pem 1)홍길동 선생님 외래 내원 예정입니다" .... }

NER tags = [
{"DAT-B O PER-B O",
 "PER-B O O O O",
 "DAT-B O PER-B O O O O",
 "DAT-B O PER-B O O O O" .... }
```

Fig. 1 | Process of NER tagging using LLM. English interpretation of Korean sentences: $^{1)}$ Dr. Hong Gil-dong is scheduled to visit the outpatient clinic, $^{2)}$ Kim Cheol-su's outpatient appointment, $^{3)}$ Lee Young-hee's outpatient appointment. $^{4)}$ Kim Gap-su cartentor after opd fu. $^{5)}$ Bong Joon-ho's outpatient appointment.

Fig. 2 | Number of matching tags per model.



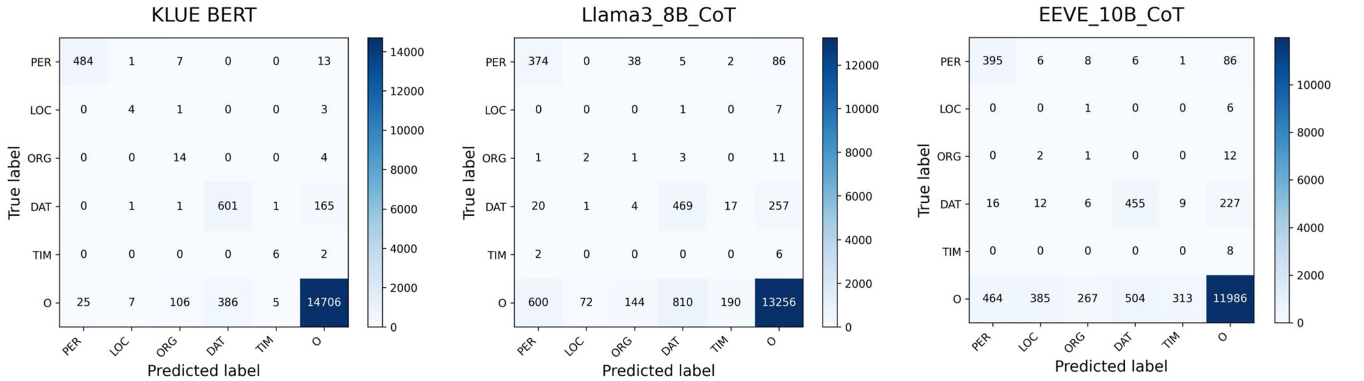


Fig. 3 | Entity-wise confusion matrix per model.

results. In the case of open LLMs, performance was evaluated under two conditions: general prompting and CoT prompting. The CoT approach consistently yielded higher performance. For example, Llama3-Open-Ko-8B achieved 347 matching tags with general prompting, but improved to 366 tags when CoT prompting was applied. Similarly, EEVE-Korean-Instruct-10.8B initially produced 182 matching tags, which increased significantly to 327 with the application of CoT prompting.

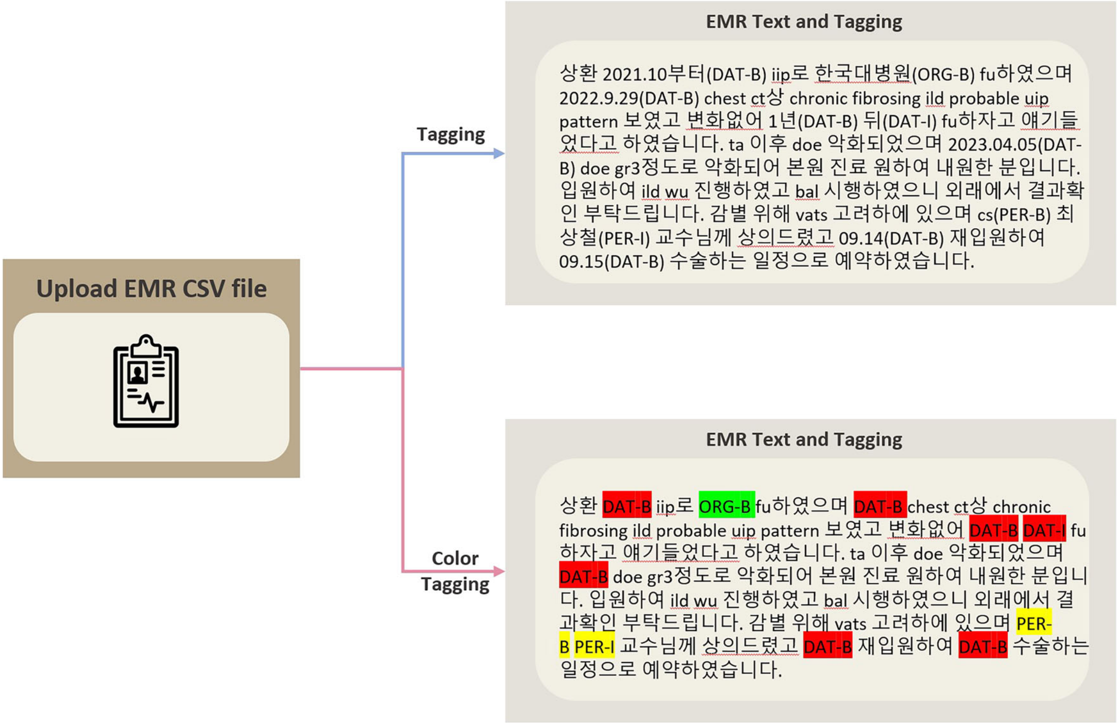
We conducted an error analysis on KLUE BERT, which was trained using base+AD, as well as on Llama3-Open-Ko-8B and EEVE-Korean-Instruct-10.8B, both of which incorporated the CoT prompting technique, to provide a more detailed understanding of each model's behavior and error patterns based on their performance on the test dataset. To this end, we analyzed the entity-level tagging results using the confusion matrices shown in Fig. 3. The results showed that all models exhibited relatively low confusion for the "PER", "DAT", and "O" entities. In particular, KLUE BERT demonstrated the highest overall consistency and accuracy, with minimal confusion between entity categories, highlighting its robustness. In contrast, Llama3-Open-Ko-8B and EEVE-Korean-Instruct-10.8B tended to confuse various entities with the "O" category more frequently. Overall, compared to KLUE BERT, these two models showed lower recall but higher precision, suggesting a conservative tendency to predict "O" in ambiguous situations.

Software for de-identification

The aim of the present study was to develop an affordable and highly usable EMR de-identification software that can be used with ease at any hospital. The proposed NER model was leveraged to automatically de-identify personal information, and the software was implemented with an intuitive user interface using Gradio[30]. This software accurately analyzed EMR data, protected sensitive patient information, and facilitated the secure sharing of data between hospitals while supporting research and analysis.

The software enabled the input of unstructured text EMR data in CSV format for de-identification. The pre-trained NER model could identify and tag personal information within the text. Two buttons, the "Tagging" and "Color Tag" buttons, were provided during this process. The "Tagging" button enabled the verification of whether the tagging and text processing were accurate. The "Color Tag" ensured that the personal information was properly highlighted and displayed with the correct NER tags.

Figure 4 illustrates the Gradio interface for de-identifying EMRs using the NER model. This interface visually demonstrates the process of identifying and tagging personal information within sample EMR text. For convenience, Table 3 presents both the original Korean EMR text and its de-identified, tagged version translated into English.



상환 2021.10부터(DAT-B) iip로 한국대병원(ORG-B) fu하였으며 2022.9.29(DAT-B) chest ct상 chronic fibrosing ild probable uip pattern 보였고 변화없어 1년(DAT-B) 뒤(DAT-I) fu하자고 얘기들었다고 하였습니다. ta 이후 doe 악화되었으며 2023.04.05(DAT-B) doe gr3정도로 악화되어 본원 진료 원하여 내원한 분입니다. 입원하여 ild wu 진행하였고 bal 시행하였으니 외래에서 결과확인 부탁드립니다. 감별 위해 vats 고려하에 있으며 cs(PER-B) 최상철(PER-I) 교수님께 상의드렸고 09.14(DAT-B) 재입원하여 09.15(DAT-B) 수술하는 일정으로 예약하였습니다.

상환 DAT-B iip로 ORG-B fu하였으며 DAT-B chest ct상 chronic fibrosing ild probable uip pattern 보였고 변화없어 DAT-B DAT-I fu하자고 얘기들었다고 하였습니다. ta 이후 doe 악화되었으며 DAT-B doe gr3정도로 악화되어 본원 진료 원하여 내원한 분입니다. 입원하여 ild wu 진행하였고 bal 시행하였으니 외래에서 결과확인 부탁드립니다. 감별 위해 vats 고려하에 있으며 PER-B PER-I 교수님께 상의드렸고 DAT-B 재입원하여 DAT-B 수술하는 일정으로 예약하였습니다.

Fig. 4 | EMR de-identification input in Gradio and outcomes (Korean version).

Table 3 | EMR de-identification input in Gradio and outcomes (English version)

Original Data	EMR Text and Tagging
From 2021.10, IIP was performed at Korea University Hospital and chronic fibrosing ild probable uip pattern was shown on chest CT on 2022.9.29 and there was no change, so they said to do FU after 1 year. After TA, DOE got worse and on 2023.04.05, DOE got worse to gr3 and visited our hospital for treatment. I was admitted and underwent ild wu and bal, so please check the results at the outpatient clinic. VATS is being considered for differentiation, and Professor Choi Sang-cheol of CS was consulted and readmitted on 09.14 and surgery was scheduled for 09.15.	From 2021.10 **(DAT-B)**, IIP was performed at Korea University Hospital **(ORG-B)** and chronic fibrosing ild probable uip pattern was shown on chest CT on 2022.9.29 **(DAT-B)** and there was no change, so they said to do FU after **(DAT-I)** 1 year **(DAT-B)**. After TA, DOE got worse and on 2023.04.05 **(DAT-B)**, DOE got worse to gr3 and visited our hospital for treatment. I was admitted and underwent ild wu and bal, so please check the results at the outpatient clinic. VATS is being considered for differentiation, and Professor Choi Sang-cheol **(PER-I)** of CS **(PER-B)** was consulted and readmitted on 09.14 **(DAT-B)** and surgery was scheduled for 09.15 **(DAT-B)**.

Discussion

This study presents a lightweight and practical de-identification software designed to protect patient privacy in unstructured Korean EMR text while preserving critical clinical information—such as diagnoses, surgical schedules, and medication prescriptions—essential for research purposes. To achieve this, we proposed a novel preprocessing method for effectively identifying personal information in unstructured EMR text. Furthermore, we enhanced model performance by augmenting the existing Korean-specific NER dataset with newly added data to overcome limitations in coverage and diversity. Most existing NER models are specialized for languages such as English or Chinese[12–14], or are trained on pre-defined personal information categories. To address this, we redefined de-identification standards suitable for the Korean healthcare environment, based on the PIPA and relevant national guidelines[9], and constructed training data accordingly.

The KLUE BERT model trained on the augmented dataset achieved the best performance with an F1-score of 94.30%, and maintained a high accuracy of 93.74% when applied to real discharge summary data. These results demonstrate that incorporating data augmentation into a Korean-specialized model significantly improves entity recognition performance and enables accurate de-identification even in complex clinical documents containing a mixture of Korean, English, and special characters. This indicates that the proposed model can serve as an effective tool for NER tasks on Korean EMR data.

This KLUE BERT model was evaluated on real discharge summary data in comparative experiments with open-source LLMs, and it demonstrated over 200 more matching tags than the open-source LLMs when assessed on a gold-standard dataset of 1000 samples. These results suggest that lightweight language models, when fine-tuned for a specific language and trained on sufficient domain-specific data, can achieve higher accuracy and greater flexibility than large-scale, computationally intensive LLMs.

Finally, a Gradio-based software prototype was implemented, demonstrating an automated de-identification process wherein users can input patient data and receive anonymized outputs in real time. This prototype highlights the feasibility of deploying the proposed solution in diverse hospital environments as an accessible and practical privacy-preserving tool.

A limitation of the present study is that, rather than training the model on a dataset directly tagged from discharge summary data, de-identification was performed using a model trained on external Korean data. Manual tagging of the discharge summaries after performing NER tasks would enable customized de-identification for each hospital. However, this is a cumbersome process that requires manual effort and consumes a significant amount of time and cost, making it inefficient. Furthermore, by incorporating external data, the model has the advantage of enabling robust identification of personal information not only within a single institution but also across various external hospitals, demonstrating its generalizability and

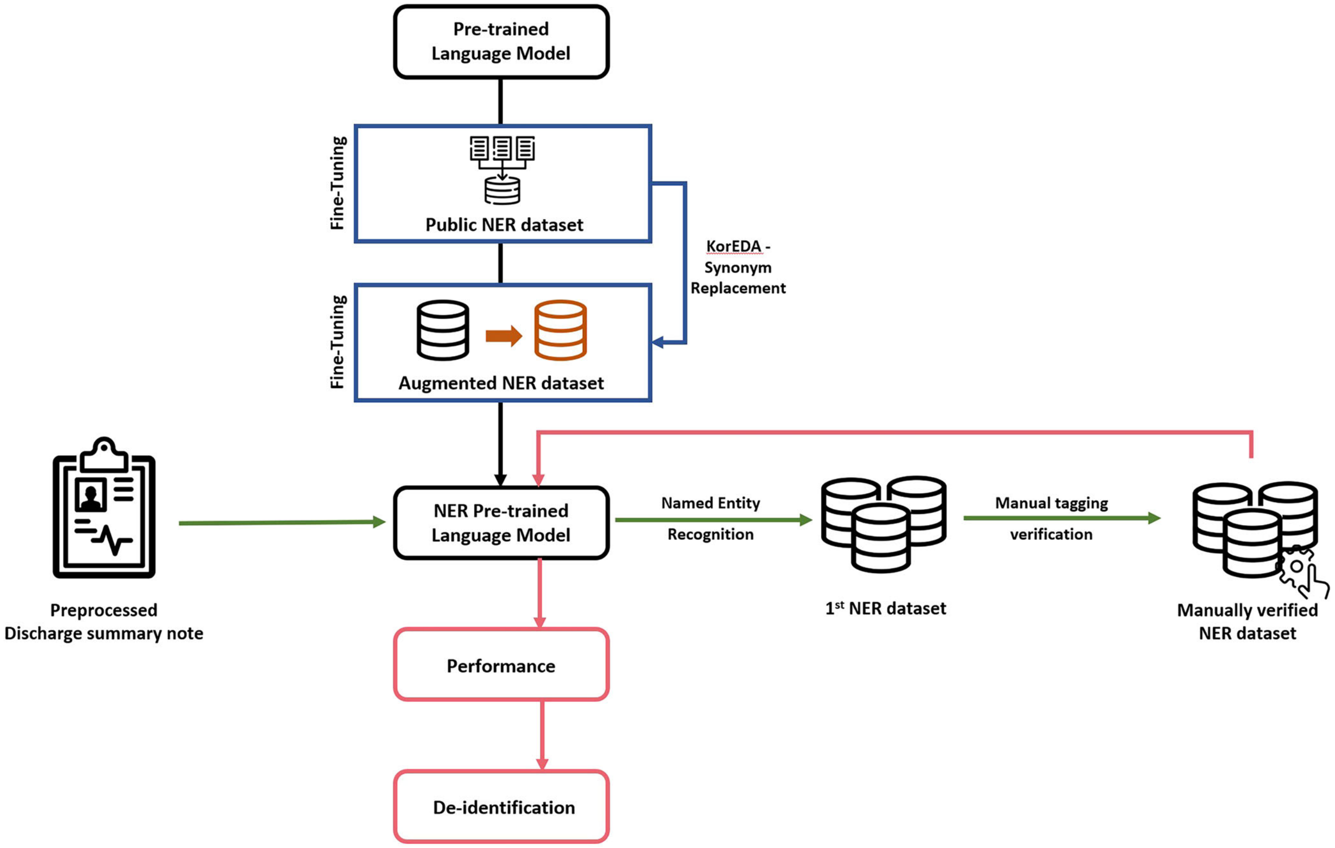


Fig. 5 | Overall flow in this study.

practical applicability. Therefore, a simple and fast learning method was proposed in the present study to address this limitation.

This study applied NER to de-identify personal information in unstructured discharge summaries from a single medical institution in AMC, and thus has limitations in terms of generalizability. However, since the training data used in this study included not only data from our institution but also publicly available Korean datasets, it is expected that identifying common sensitive information in discharge summaries from other Korean medical institutions would not pose significant difficulties. Nevertheless, variations in document formats across different hospitals may inevitably lead to differences in de-identification performance. To address this issue, our research team is collaborating with the Ministry of Health and Welfare of Korea to develop standardized templates based on Fast Healthcare Interoperability Resources (FHIR). Through this initiative, we aim to enhance the applicability of de-identification methods and promote the generalizability of our research outcomes.

Another limitation of the present study is the absence of large-scale parameter LLM models, such as Llama3-70B. Compared with that achieved with the Llama3-8B model, the use of large-scale LLMs specialized for Korean facilitated more precise tagging. However, the BERT model developed herein comprised 110 million parameters, in contrast to the 8 billion parameters of the Llama3-8B model. When compared to the time applied to EMR, a significant difference is observed. In this study, de-identification was performed on 1000 lines of unstructured text data using the BERT model, Llama3-Open-Ko-8B, and EEVE-Korean-Instruct-10.8B. As a result, the BERT model took 106 s, while the Llama3-Open-Ko-8B and EEVE-Korean-Instruct-10.8B models took 292 s and 429 s, respectively. In the case of applying CoT, additional processing time was required. The primary goal of this study was to provide a universal and lightweight solution for efficiently utilizing medical data in real-time research and analysis, supporting doctors, nurses, and engineers within hospitals. At Seoul AMC, an average of 12,000 outpatient visits and 2600 inpatient admissions occur daily, generating vast amounts of EMR data. To handle this efficiently, leveraging BERT-based models for de-identification proves to be an effective approach, ensuring both high performance and fast processing speeds.

Despite these limitations, a high-performance NER model for EMR de-identification was developed in the present study by overcoming the lack of Korean data and enhancing the training dataset through pre-processing tasks and data augmentation for discharge summaries. The proposed model can address these issues in the future by manually modifying physician-labeled notes; however, additional studies must be conducted to provide an automated solution without human intervention. The performance of the KLUE BERT model was superior to that of open LLMs with larger parameters. This finding suggests that the model can effectively de-identify personal information in Korean EMRs, thereby expanding the potential use of clinical texts and broadening the scope of data utilization. Thus, the proposed model could contribute to the advancement of medical research.

Methods

This retrospective study adhered to the tenets of the Declaration of Helsinki (2008) and was approved by the Institutional Review Board of the Seoul AMC (IRB-2023-1293). The data were extracted from the "data" and "clinical research data" warehouses. No patient-identifying information was provided to the models used in the present study: Llama3-Open-Ko-8B, EEVE-Korean-Instruct-10.8B, Llama3-8B-Instruct. The entire research process is visually summarized in Fig. 5 for clarity.

Study design

The "Guidelines for the Utilization of Health and Medical Data"[9] states that personal information refers to information related to living individuals, including details such as name, resident registration number, and images, that can be used on their own or when combined with other information to identify an individual. Therefore, the present study focused on key entity

Table 4 | Tag and definition of entity name category class

	Entity Name Category	Tag	Definition
1	PERSON	PER	Items such as real or virtual names of individuals
2	ORGANIZATION	ORG	Items such as organizations and institutions, as well as meetings and conferences
3	LOCATION	LOC	Items such as names of regions and administrative divisions
4	TIME	TIM	Time
5	DATE	DAT	Date

Table 5 | Final entity category

Total Entity Tagging Category	Number
PER-B, PER-I, ORG-B, ORG-I, LOC-B, LOC-I, TIM-B, TIM-I, DAT-B, DAT-B, O	11

name tagging to extract information in clinical text data that can be used to identify individuals.

Five categories were established for the entity names used for de-identification: person (PER), organization (ORG), location (LOC), time (TIM), and date (DAT) (Table 4). The entity name tagging categories were created based on meaningful words. For instance, the term “한국사람“ (Korean person) was split into “한국“ (Korea) and “사람“ (person), such that “한국“ could be tagged as a location (LOC) or country (ORG). However, the tagging process was conducted based on the meaning of the word, as “한국사람“ is used as a common noun.

NER divided a sentence into tokens and assigned tags to each token to distinguish named entities. A single named entity can consist of multiple tokens; therefore, a tagging system was used to correctly group them. BIO, a commonly used tagging system, is characterized by its ease of implementation and interpretation, as well as its ability to ensure a consistent entity structure. This system provided a structure that can distinguish between the beginning and internal tokens of an entity[31].

- **B**(Beginning): Beginning of the entity
- **I**(Inside): Inside of the entity
- **O**(Outside): Tokens, not entities

BIO can recognize meaningful words and identify entity boundaries. The segmentation of labels enables the model to learn specific tagging patterns. These features of BIO have led to its use in fields such as healthcare and law that require the processing of complex entities. For this reason, the BIO tagging system was used in the present study to assign -B and -I to the five categories. A total of 11 tagging labels were obtained with the inclusion of O (Table 5). The simplicity of BIO tags and the clear boundary setting enhance the accuracy and consistency of learning, in addition to ensuring compatibility with various datasets.

EMR discharge note preprocessing

The dataset used to apply the trained NER model for de-identification comprises the discharge summaries of patients admitted to Seoul Asan Medical Center between January 2019 and October 2023. The dataset includes 611,681 rows and 22 columns. Eight columns with missing data were removed, yielding a final dataset with 14 columns. The discharge summaries primarily included information related to discharge, such as primary symptoms, allergy symptoms, primary diagnosis, other diagnoses, and surgical or procedure information. These texts were primarily written by administrators, nurses, and physicians. The text contained a mix of Korean, English, stopwords, and special characters. De-identification of this text is essential, given that it includes personally identifiable information such as names of patients and healthcare providers, as well as regions and locations.

An appropriate pre-processing process for the discharge summaries was commenced to apply the trained NER model. The entirety of the text was converted to lowercase. Subsequently, the characters were removed such that only English letters, Korean characters, and numbers were retained. This preprocessing approach differed from standard text data preprocessing. Special characters are typically removed from text to improve text utility in general; however, the correlation between the context and numbers can influence the prediction performance of the model for medical data. Furthermore, the meaning of special characters must be preserved to facilitate precision medicine, personalized treatments, and comparisons with other research.

The data pre-processing steps applied in the present study were as follows. Special characters such as “.”, “:”, “+“, “-“, “>“, “<“, “~“, and “%“ were retained by using regular expressions. The period symbol (“.”) is used in unit notation; thus, it was retained during the pre-processing steps as removing it could lead to incorrect interpretations, such as “2.48 kg” being misread as “248 kg.” The symbols “+“ and “-“ were also retained, given that they may indicate additional information following numbers. The symbols “>“ and “<“ were also retained as they can represent steps in a sequence or size comparisons. The tilde symbol (“~”) was also retained as it is used to represent ranges, e.g., “6–7 months old.” The percentage symbol (“%”) was also retained as it indicates percentages, e.g., “80–90%”.

Consecutive spaces were unified into a single space during the pre-processing process, and newline characters were removed. The NER dataset was tokenized based on simple word spacing, and all data types were converted to strings. The missing values were filled with the text “missing,” and categorical variables were standardized to the most frequent value based on their frequency. For instance, “호전됨“ and “호전됨(improved),” if present, were unified under the variable name “호전됨”.

Training data and data augmentation

The Korean NER dataset used for training in the present study was sourced from the Naver NLP Challenge 2018 dataset. This dataset, comprising 90,000 sentences labeled with 14 entity categories using the BIO tagging method, was derived from the Korean Wikipedia, processed into text form, and made publicly available. Text data augmentation was applied based on this dataset to obtain a Korean NER dataset with the entity names tagged at the word and spacing token levels.

Data regarding the names of medical institutions for patients transferred from the emergency department to other hospitals at Seoul AMC were extracted to enhance the training for accurate detection of organization names (ORG). The names of 221 medical institutions were extracted, and these data were tagged as ORG-B for the names of other hospitals to which the patients were transferred from AMC.

Improving the performance of the model is one approach to developing an accurate DL model; however, obtaining vast amounts of data through diverse data learning is essential. Data augmentation artificially generates new data based on existing data. This method is often used in the image domain to expand the dataset through the addition of noise while maintaining the core features of the data. Image augmentation, which involves transformations such as rotation or brightness adjustments, is performed to expand the data and enhance the prediction performance of the model when utilized as training data[32].

Unlike images, in text data, even a single word modification or a change in word order can alter the meaning of a sentence, making data augmentation more challenging. In this study, additional Korean NER tagging datasets were needed to de-identify personally identifiable information

Table 6 | Count of augmented data by entity type

Entity Name Category	Tag	Counts
PERSON	PER-B	10,201
	PER-I	1236
ORGANIZATION	ORG-B	11,328
	ORG-I	1293
LOCATION	LOC-B	6323
	LOC-I	89
TIME	TIM-B	855
	TIM-I	290
DATE	DAT-B	7331
	DAT-I	2443
etc.	O	255,739

Table 7 | Description of Base and Base + AD Data

Type	Counts	
	Base	Base + AD
Average sentence length	51.79	54.02
Average number of tokens	28.37	29.64
Number of entities by category		
PER-B/-I	43,034/5465	53,235/6401
ORG-B/-I	41,078/4723	52,406/6016
DAT-B/-I	25,837/8107	33,168/10,550
LOC-B/-I	20,885/211	27,208/300
TIM-B/-I	3263/1074	4118/1364
O	910,471	1,166,210
Average number of entities per sentence by category		
PER-B/-I	0.29/0.09	0.31/0.10
ORG-B/-I	0.46/0.05	0.48/0.05
DAT-B/-I	0.29/0.09	0.30/0.10
LOC-B/-I	0.23/0.002	0.25/0.002
TIM-B/-I	0.04/0.01	0.04/0.01
O	10.10	10.64

within sentences containing diverse meanings. Therefore, we applied the Easy Data Augmentation (EDA)[33] technique to the existing Korean dataset from Naver, and used KorEDA (GitHub—catSirup/KorEDA), which leverages the Korean WordNet (KWN) for augmentation specialized in Korean.

There are four main techniques for text data augmentation: Synonym Replacement (SR), Random Insertion (RI), Random Swap (RS), and Random Deletion (RD)[33]. We chose the SR technique to secure additional words and sentences while preserving the originally tagged dataset. This method randomly selects N non-stop words from a sentence and replaces each with one of its synonyms. KorEDA is a modified version of the original EDA, replacing the WordNet component with Korean WordNet, so the selected words in the original sentence are replaced with semantically similar synonyms. Here is the example provided below. In this way, the position and meaning of each word in the original sentence are preserved as much as possible while augmenting the data.

- Original: “환자가 병원에서 치료를 받았습니다.” * The patient received treatment at the hospital.
- After SR: “환자가 의료기관에서 치료를 받았습니다.” * The patient received treatment at the medical institution.

The final augmented dataset includes an additional 19,371 rows where the sentence length remained the same, but words were replaced with synonyms. In this dataset, unnecessary classes except for the personally identifiable entity categories “PER,” “ORG,” “LOC,” “TIM,” and “DAT” were replaced with “O,” and the number of augmented samples per entity type is summarized in Table 6. Since augmentation was applied uniformly across the entire dataset, augmentation rates were not calculated separately for each entity type.

The combined dataset, which integrates all data, consists of a total of 109,589 rows. Table 7 provides detailed information on the average sentence length, number of tokens, and tag counts for the Base+AD dataset, which comprises the original Base data set along with the augmented data.

Pretrained language models

Tokenization and modeling with Bert base models. BERT, a transformer-based model, processes bidirectional input sequences to predict randomly masked tokens and understand general patterns and structures of the language. BERT can capture the meaning of words in context, thereby providing contextual embeddings based on the usage within a sentence. This bidirectional context understanding capability of BERT is particularly strong for context-dependent tasks such as NER[25]. In contrast to the original BERT, which is trained only in English, BERT-base-multilingual-cased[25] is a multilingual BERT model trained on the top 104 languages available through Wikipedia. It supports an extensive tokenizer that can handle multiple languages and is optimized for multilingual NLP tasks, indicating its utility for de-identifying personal information in text containing a mix of Korean and English. KLUE BERT, a BERT architecture-based model trained using 62 gigabytes of Korean-specific data, can effectively understand Korean grammar, vocabulary, and context[24]. KLUE BERT has been further fine-tuned with data sourced from news and Wikipedia, thereby optimizing its ability to understand Korean. Thus, the use of KLUE BERT was deemed appropriate for processing the de-identification of personal information in Korean discharge summaries.

The pre-trained weights of the two models were maintained during fine-tuning using the constructed datasets. The datasets were divided into the Naver dataset (Base) and the added augmented dataset (Base+AD), and training was conducted twice. The structure of each dataset was designed such that it was suitable for the NER task. Entity recognition labels were tagged at the word level rather than the morpheme level to enhance the simplicity of implementation and readability. This tagging method improved the accuracy of entity recognition based on the context. Each entity name was mapped with entity tagging information and integers.

The Korean tokenizer based on BERT performs tokenization into optimized subword units that reflect the morphological characteristics of the Korean language[24]. However, the tokenizer's vocabulary also includes the English alphabet, numbers, and various special characters, enabling effective segmentation and encoding of not only Korean but also English words and special characters. English words included in the vocabulary are treated as single tokens, while those not included are segmented into subwords. Special characters are similarly handled as separate tokens or subwords. Therefore, the presence of English words and special characters does not significantly affect the model's performance or preprocessing pipeline. The tokenizer applied to each model split words into subwords, resulting in longer sentence lengths and varying label lengths. Only the first subword of the generated subwords was assigned the original label; the remaining subwords were assigned –99, indicating that no label was provided. The labels corresponding to –99 were ignored during model training, and padding tokens [PAD] were used to maintain a consistent sentence length. [CLS] was added at the beginning, and [SEP] was added at the end to differentiate sentences. Integer encoding was applied for BERT input. In addition, segment encoding and attention masks were also applied. The NER task does not require distinguishing between multiple sentences; consequently, segment encoding filled all tokens with “0”. The value of “1” was assigned from [CLS] to [SEP] for the attention mask, and the value of “0” was assigned after [SEP]

Table 8 | Hyperparameters for BERT and KLUE BERT

	Value	
	BERT-base-multilingual-cased	KLUE BERT
Hyperparameter		
Input sequence length	256 tokens	256 tokens
Optimizer	Adam	Adam
Learning rate	2e-5	5e-5
Loss function	SparseCategoricalCrossentropy	SparseCategoricalCrossentropy

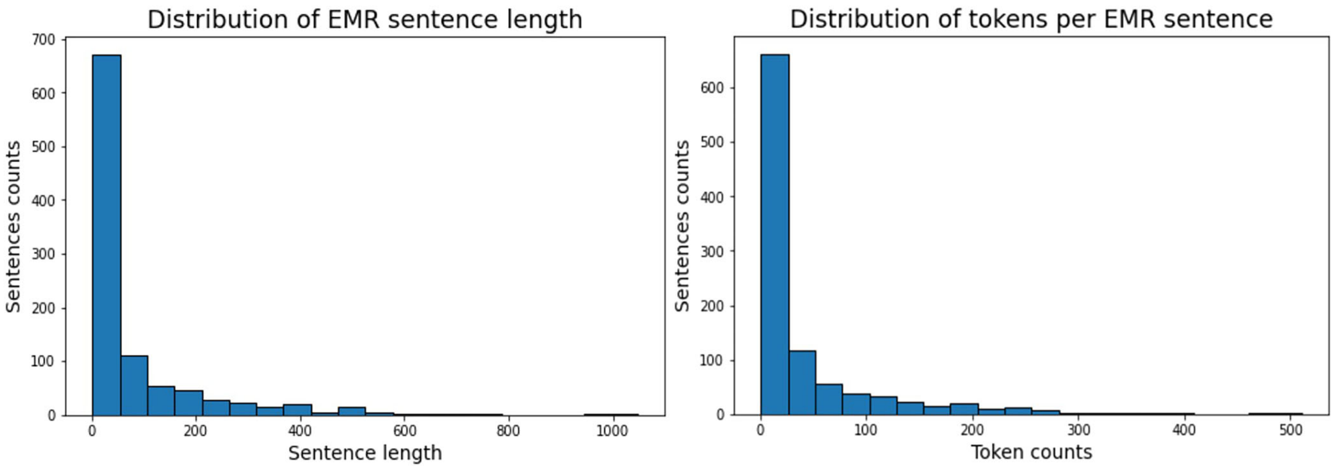


Fig. 6 | Distribution of Sentence Lengths and Token Counts in EMR of the test dataset.

to indicate the tokens that the model should focus on. NER is a many-to-many problem wherein each word must be assigned a tag for personal information recognition. The loss function was designed to ignore the error and not compute the loss when the label was –99. This ensured that the –99 value was treated as the [PAD] token in the label and no loss was calculated for it. This processing method improved the efficiency of model training and reduced unnecessary computations.

The hyperparameters used for each model are summarized in Table 8. They were determined based on prior research and empirical experience, with experimental tuning conducted to achieve optimal performance. Considering the sentence length and token counts in the EMR discharge summaries, the input sequence length was set to 256 tokens, and the Adam optimizer was employed. Learning rates were adjusted individually for each model to optimize performance. For the loss function, SparseCategoricalCrossentropy, suitable for multi-class classification, was used. Additionally, a custom implementation was applied to exclude the masked value of -100 from loss calculation, as processed by the tokenizer. The related formula is presented below.

$$L = -\frac{1}{N}\sum_{i=1}^{N} 1_{[y_i \neq -100]} \sum_{c=1}^{C} 1_{[y_i = C]} \log\left(\frac{\exp(z_{i,c})}{\sum_{j=1}^{C}\exp(z_{i,j})}\right) \tag{1}$$

Here, N denotes the total number of tokens, and C represents the number of classes (or labels). The term y_i refers to the ground-truth label of the i-th token, while $z_{i,c}$ indicates the logit corresponding to the c-th class for the i-th token. In addition, $1_{[y_i \neq -100]}$ is an indicator function that equals 1 if the token is not padding (i.e., the label is not -100), and 0 otherwise. Similarly, $1_{[y_i = C]}$ is an indicator function that equals 1 if the ground-truth label of the i-th token is c, and 0 otherwise.

Model training was conducted with a 90% training ratio and a 10% test ratio using the custom dataset. The Precision, Recall, and F1 scores were evaluated to determine model performance. The F1 score was calculated at the end of each epoch to assess overfitting, and the generalization ability of the model was continuously monitored.

Construction of discharge note test dataset. To construct a test dataset (gold standard) for model performance evaluation, 1000 rows were randomly sampled from the "Treatment Plan" and "Key Notes" columns of the AMC discharge summary data. The most performant model among our four internally developed models was then used to generate tagging results on this data. The tagging results were subsequently reviewed by healthcare professionals with over three years of experience at AMC to ensure accuracy. The finalized dataset was ultimately used as a test set for performance comparison with open LLM models.

The average sentence length in the test dataset was 84.75, with an average token count of 43.80. Through Fig. 6, we confirmed that most sentence lengths and tokens do not exceed 200.

Performance evaluation

Two BERT-based models were fine-tuned using the dataset, resulting in a total of four models. The F1-score, a standard metric used to evaluate classification model performance, is particularly useful when there is class imbalance in the data. The F1-score is defined as the harmonic mean of precision and recall, reflecting the balance between the two metrics. The F1-score is used to assess whether the model can accurately recognize and classify entities in NER tasks. Errors such as incorrect entity identification or omission may occur in NER. The F1-score provides a balanced evaluation of these errors, thereby offering a comprehensive assessment of the performance of the model. The F1-score offers the advantage of objectively evaluating the performance of specific entity classes, particularly in cases wherein the "O" class dominates in the tagging results, causing significant data imbalance.

Open source LLM & prompt engineering

API-based closed LLMs[34,35] pose a risk of data leakage. Thus, publicly available models for academic and research purposes were used to handle EMR data containing personal information. The latest Llama3 series provided by Meta AI[28] and the Solar model from Upstage were used in the present study. LLMs tailored for the Korean language were used for Korean-specific NER tagging.

The present study aimed to build an LLM that can be easily used in hospitals. Large parameter models such as Llama3 70B did not align with the research intent. Therefore, Llama-3-8B-Instruct and Llama-3-open-ko-8B[26] based on the LLaMA series by Meta, which comprise 8 billion parameters each, were used for Korean-specific NER tagging. This is significantly higher than the 110 million parameters of the BERT model used for training, allowing for a more complex understanding of language. Llama-3-open-ko-8B is a language model pre-trained specifically for Korean, based on Llama-3-8B. Llama-3-8B-Instruct, which focuses on instruction-based learning, responds naturally and appropriately to user queries. Notably, it supports multiple languages, making it a suitable comparison model for the present study. EEVE-Korean-Instruct-10.8B[27], a Korean-specific instruction-based language model with 10.8 billion parameters, is a fine-tuned version of Upstage's Solar model[36], extended with Korean vocabulary and is adjusted to better meet the requirements of Korean users.

LLMs exhibit outstanding zero-shot performance; however, additional improvements are often required for more complex tasks. Few-shot prompting is a technique that incorporates example instances into the prompt, enabling the model to better understand the desired output through contextual learning[37]. This approach helps reduce ambiguity, improve clarity, and enhance overall efficiency. In particular, we adopt the CoT prompting strategy to guide the model to solve complex problems in a step-by-step and logical manner[29]. CoT prompting explicitly encourages the model to generate intermediate reasoning steps, which leads to more systematic and reliable outputs. For example, when instructing the model on NER tagging, we provide explicit examples that include not only the correct tag (e.g., "PER-B" to denote the beginning of a person name) but also detailed explanations of the reasoning process behind each tagging decision. In this study, the LLM was guided using a combination of example-based few-shot prompts and CoT prompting, enabling it to perform fine-grained entity tagging on EMR data accurately.

Outputs comparison

The tagging results of each model were evaluated using Python scripts based on sentence length, token count, and character-level matching. A value of "1" was assigned when all three criteria were fully satisfied, and "0" was assigned if any criterion was not met.

We aimed to gain a deeper understanding of the dataset complexity for each model and to present their error patterns in more detail. To this end, we conducted an in-depth analysis of entity-level tagging results using confusion matrices.

Gradio software

Gradio, an open-source Python package, facilitates the sharing and testing of machine learning (ML) and DL models[30]. Traditional ML models, even those built by experts, often present challenges for domain experts or general users to test directly. However, Gradio generates a web-based visual interface that enables non-experts to use ML models and provide feedback without requiring coding, thereby enhancing the ease of use[29].

Data availability

The data resources generated and analyzed during this study are not publicly available but can be shared upon reasonable request. Interested researchers are encouraged to contact the corresponding author to discuss potential access and usage conditions.

Code availability

The code supporting this study is available upon request from the corresponding author.

Received: 12 May 2025; Accepted: 4 August 2025;
Published online: 04 September 2025

References

1. Hanson, J. L. et al. Quality of outpatient clinical notes: a stakeholder definition derived through qualitative research. *BMC Health Serv. Res.* **12**, 1–12 (2012).
2. Kierkegaard, P. Electronic health record: wiring Europe's healthcare. *Comput. Law Secur. Rev.* **27**, 503–515 (2011).
3. Fernández-Alemán, J. L. et al. Security and privacy in electronic health records: a systematic literature review. *J. Biomed. Inform.* **46**, 541–562 (2013).
4. NHS Lothian Communications Office. NHS Lothian staff member loses patient data. *NHS Lothian* http://www.nhslothian.scot.nhs.uk/MediaCentre/PressReleases/2008/Pages/0307PatientData.aspx/ (2008).
5. Department of Veterans Affairs Office of Inspector General. Review of issues related to the loss of VA information involving the identity of millions of veterans. *Department of Veterans Affairs* http://www.va.gov/oig/apps/info/OversightReports.aspx?igRT=ai/&igPG=4 (2006).
6. Kim, M. et al. Fine-Tuning LLMs with Medical Data: Can Safety Be Ensured? *NEJM AI* 2.1 (2025): AIcs2400390.
7. European Medicines Agency. External guidance on the implementation of the European Medicines Agency policy on the publication of clinical data for medicinal products for human use. *European Medicines Agency* (2017).
8. Dorr, D. A., Phillips, W. F., Phansalkar, S., Sims, S. A. & Hurdle, J. F. Assessing the difficulty and time cost of de-identification in clinical narratives. *Meth. Inf. Med.* **45**, 246–252 (2006).
9. Personal Information Protection Commission, Ministry of Health and Welfare. Guidelines on the Utilization of Healthcare Data. Korea Health Information Service Notice. https://k-his.or.kr/board.es?mid=a10301000000&bid=0001&list_no=1538&act=view (2024).
10. Office for Civil Rights, HHS. Standards for privacy of individually identifiable health information. *Fed. Regist*. **67**, 53182–53273 (2002).
11. Shin, T.-S. The meaning and tasks of guidelines for utilization of healthcare data. *Korean Soc. Law Med.* **22**, 31–55 (2021).
12. Hardeniya, N. et al. *Natural Language Processing: Python and NLTK*. (Packt Publishing Ltd., 2016).
13. Liu, P. et al. Chinese named entity recognition: the state of the art. *Neurocomputing* **473**, 37–53 (2022).
14. Zaratiana, U. et al. Gliner: generalist model for named entity recognition using bidirectional transformer. In *Proc. Conf. North Am. Chapter Assoc. Comput. Linguist.: Hum. Lang. Technol.* (Vol. 1: Long Papers) 5364–5376 (Association for ComputationalLinguistics, 2024).
15. Ahmed, T., Al Aziz, M. M. & Mohammed, N. De-identification of electronic health record using neural network. *Sci. Rep.* **10**, 18600 (2020).
16. Yang, X. et al. A study of deep learning methods for de-identification of clinical notes in cross-institute settings. *BMC Med. Inform. Decis. Mak.* **19**, 1–9 (2019).
17. Liu, L. et al. De-identifying Australian hospital discharge summaries: An end-to-end framework using ensemble of deep learning models. *J. Biomed. Inform.* **135**, 104215 (2022).
18. Kim, Y.-M. & Lee, T.-H. Korean clinical entity recognition from diagnosis text using BERT. *BMC Med. Inform. Decis. Mak.* **20**, 1–9 (2020).
19. Kim, Y. et al. A pre-trained BERT for Korean medical natural language processing. *Sci. Rep.* **12**, 13847 (2022).
20. Byun, S. et al. Korean Bio-Medical Corpus (KBMC) for Medical Named Entity Recognition. In *Proc. Joint Int. Conf. Comput. Linguist., Lang. Resour. Eval. (LREC-COLING 2024).* 9941–9947 (European Language Resources Association, 2024).

21. An, J. et al. De-Identification of Clinical Notes with Pseudo-labeling using Regular Expression Rules and Pre-trained BERT. (2023).
22. Shin, S.-Y. et al. A de-identification method for bilingual clinical texts of various note types. *J. Korean Med. Sci.* **30**, 7–15 (2015).
23. Park, Y.-J. et al. Assessing the research landscape and clinical utility of large language models: a scoping review. *BMC Med. Inform. Decis. Mak.* **24**, 72 (2024).
24. Park, S. KLUE: Korean Language Understanding Evaluation. In **Adv. Neural Inf. Process. Syst.** (NeurIPS Datasets &Benchmarks Track, 2021).
25. Devlin, J. BERT: Pre-training of deep bidirectional transformers for language understanding. In **Proc. Conf. North Am. Chapter Assoc. Comput. Linguist.: Hum. Lang. Technol.** (Long Papers) 4171–4186 (Association for Computational Linguistics, 2019).
26. Llama-3-Open-Ko L. *Llama-3-Open-Ko* https://huggingface.co/beomi/Llama-3-Open-Ko-8B (2024).
27. Kim, S., Choi, S. & Jeong, M. Efficient and effective vocabulary expansion towards multilingual large language models. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2402.14714 (2024).
28. Dubey, A. et al. The Llama 3 herd of models. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2407.21783 (2024).
29. Wei, J. et al. Chain-of-thought prompting elicits reasoning in large language models. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2201.11903 (2022).
30. Abid, A. et al. Gradio: Hassle-free sharing and testing of ML models in the wild. In **Proc. ICML Workshop Human in the Loop Learning (HILL 2019)** (Long Beach, CA, 2019).
31. Ramshaw, L. A. & Marcus, M. P. Text chunking using transformation-based learning. *Nat. Lang. Process. Using Very Large Corpora* 157–176 (Springer Netherlands, 1999).
32. Taylor, L. & Nitschke, G. Improving deep learning with generic data augmentation. In *2018 IEEE Symposium Series on Computational Intelligence (SSCI)* 1542–1547 (IEEE, 2018).
33. Wei, J. & Zou, K. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In **Proc. Conf. Empir. Methods Nat. Lang. Process. & Int. Joint Conf. Nat. Lang. Process.** 6383–6389 (Association for Computational Linguistics, 2019).
34. Brown, T. B. et al. Language models are few-shot learners. In **Adv. Neural Inf. Process. Syst.** (NeurIPS 2020) (Curran Associates, Inc., 2020).
35. Achiam, J. et al. GPT-4 technical report. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2303.08774 (2023).
36. Kim, D. et al. Solar 10.7 B: Scaling large language models with simple yet effective depth up-scaling. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2312.15166 (2023).
37. Gu, Y. et al. PPT: Pre-trained prompt tuning for few-shot learning. In **Proc. Annual Meeting Assoc. Comput. Linguist.** (Vol. 1: Long Papers) 8410–8423 (Association for Computational Linguistics, 2022).

Acknowledgements

This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: RS-2025-02213061) and by a grant [grant number: 2023IP0042] from the Asan Institute for Life Sciences, Asan Medical Center, Seoul, Korea. The funder played no role in study design, data collection, analysis and interpretation of data, or the writing of this paper.

Author contributions

J.S. conceived the research, developed the proposed methods, performed data analysis, experiments, and evaluations, and prepared the tables and figures of the paper. J.S., Y.K., H.C., M.K., J.H., G.K., S.K., H.J., B.K., and B.C. contributed to data collection and preprocessing. S.P. provided a paper review, while T.J.J. and Y.J.K. offered research direction, ideas, and assisted in the final paper review. All authors have read and approved the revised paper.

Competing interests

All authors declare no financial or non-financial competing interests.

Additional information

Correspondence and requests for materials should be addressed to Tae Joon Jun or Young-Hak Kim.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.