K-LegalDeID: A Benchmark Dataset and KLUEBERT-CRF for De-identification in Korean Court Judgments

Wooseok Choi **Hyungbin Kim** **Yon Dohn Chung**
Department of Computer Science and Engineering, Korea University
{woosukqw, hyungbinkim, ydchung}@korea.ac.kr

Abstract

The Korean legal system mandates public access to court judgments to ensure judicial transparency. However, this requirement conflicts with privacy protection obligations due to the prevalence of Personally Identifiable Information (PII) in legal documents. To address this challenge, we introduce **K-LegalDeID**, a large-scale benchmark dataset and an efficient KLUEBERT-CRF model for de-identification for Korean court judgments. Our primary contribution is a new large-scale benchmark dataset spanning 39 legal domains, with its quality is validated by a high inter-annotator agreement (IAA) with Fleiss' Kappa of 0.7352. Our results demonstrate that a lightweight KLUEBERT-CRF model, when trained on our dataset, achieves state-of-the-art performance with an entity-level micro F1 score of 0.9923. Our end-to-end framework offers a practical and computationally efficient solution for real-world legal systems.

1 Introduction

The Korean legal system has established a fundamental principle that court proceedings and judgments must be transparent to ensure public trust and accountability. Article 109 of the Korean Constitution mandates that 'the trial and judgment of courts shall be open to the public,' with an exception allowing trials to be closed only if openness might harm national security, public order, or good morals. The objective of this constitutional provision is to guarantee transparency in judicial proceedings, thereby maintaining public trust in the judiciary while simultaneously safeguarding the rights of the parties involved in litigation.

The necessity of disclosing court judgments extends beyond mere constitutional compliance to encompass broader societal benefits. Public access to these documents enables citizens to understand legal principles and precedents, supports academic research and legal education, and facilitates effective monitoring of judicial decision-making. Furthermore, the disclosure of judgments helps prevent corruption and abuse of power within the judicial system, promotes equal access to legal information regardless of social or economic status, and contributes to the development of legal technology by supplying essential data for AI-based legal services and research.

However, several challenges impede the effective implementation of judgment disclosure in Korean legal documents. First, there exists an inherent conflict between the constitutional mandate for transparency and the obligation to protect Personally Identifiable Information (PII), including names, resident registration numbers, addresses, and other personal identifiers for not only parties but also witnesses, victims, and various stakeholders.

Secondly, most de-identification processes rely on manual methods, which introduce significant bottlenecks. Manual masking typically requires approximately two weeks per document and substantial human resources, severely limiting the scope and speed of judgment disclosure. Despite existing guidelines for personal information protection, manual processing may yield inconsistent results that vary in quality across cases and personnel. Only about 1.6 million judgment documents, approximately 5.97% of the total (Administration, 2019), have undergone de-identification, highlighting the limited extent of automated or systematic anonymization to date.

Korean courts have attempted to address these challenges through automated systems, including an 'intelligent judgment de-identification system.' However, this system performs poorly, achieving only around 8% accuracy in identifying and masking PII (Administration, 2025). The limited effectiveness of current automation stems from several key technical challenges: the unique characteris-

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 2308–2325

tics of Korean legal documents, linguistic challenges such as agglutinative morphology and irregular spacing, and a severe shortage of high-quality training data.

To overcome these challenges, we propose an enhanced framework that combines KLUEBERT with a Conditional Random Field (CRF) layer, significantly improving PII detection and masking in Korean legal documents. To this end, we introduce two new datasets—one comprising case documents from 39 legal categories and another comprising multi-turn conversations from Korean SNS—and establish a unified 11-label PII annotation scheme. Our method addresses the limitations of existing systems by resolving ambiguities in masking guidelines, and implementing an end-to-end framework that ensures consistent entity anonymization while maintaining document coherence and legal relevance.

This paper provides evidence for the value of offering de-identified legal documents and supports further research in related domains. The main contributions of this paper are as follows:

- We introduce the first large-scale, high-quality benchmark dataset specifically designed for de-identification for Korean court judgments, comprising 46,973 annotated sentences from 2,000 cases across 39 diverse legal fields.
- The dataset's integrity is validated by a Fleiss' Kappa score of 0.7352, indicating substantial inter-annotator agreement and ensuring consistent, trustworthy labeling.
- We introduce an enhanced model architecture, KLUEBERT-CRF, specifically designed to handle Korean language and legal text complexities such as agglutinative morphology and intricate sentence structures.
- We present an end-to-end framework that spans the entire de-identification pipeline, demonstrating a practical pathway for deploying automated, high-accuracy de-identification for real-world environments.

2 Related Work

In this section, we review prior research in two key areas relevant to our work: (1) the broader context of de-identification across various domains, and (2) specific challenges and methodologies for de-identification in the legal domain. By analyzing existing approaches, we identify the critical research gaps that our work aims to address.

2.1 De-identification in General Domains

Automated de-identification is a well-established research area, driven by privacy regulations such as HIPAA in healthcare and GDPR in Europe. Early approaches were predominantly rule-based, relying on regular expressions and specialized dictionaries to detect explicit identifiers (Sweeney, 2002; Gupta et al., 2004).

With the development of deep learning, research shifted towards sequence labeling models using contextual embeddings (Dernoncourt et al., 2017), and fine-tuning large pre-trained language models became the de facto standard for Named Entity Recognition (NER) (Lee et al., 2020; Bogdanov et al., 2024). More recently, the paradigm has evolved towards Large Language Models (LLMs), exploring generative approaches via instruction tuning or zero-shot prompting to address data scarcity and generalization challenges (Mayhew et al., 2024; Wang et al., 2025).

Despite these advances, it remains challenging to apply general methodologies to specific domains (Szawerna et al., 2024; Larson et al., 2024). While these approaches are effective in achieving their specific objectives, they are not directly applicable to domains that require fine-grained semantic distinctions and high-quality ground truth.

2.2 De-identification in Legal Domain

Research on de-identification in the legal domain is still underexplored, despite presenting unique challenges posed by complex sentence structures, specialized terminology in documents like court judgments.

Since few studies have addressed these domain-specific characteristics, high-quality benchmark datasets required to train and evaluate de-identification models are also non-existent. Although some studies utilize LLMs to generate synthetic legal text dataset (Savkin et al., 2025), these approaches struggle to replicate the strict formalism and structural integrity of court judgments. This lack of both domain-specific methods and reliable datasets indicates the need for a systematic approach to build a legal-domain benchmark and to design effective de-identification models.

In the context of Korean legal documents, these challenges are compounded by the agglutina-

tive nature of the Korean language, which introduces morphological ambiguity and inconsistent spacing, making accurate tokenization and entity boundary detection particularly difficult. Thunder-DeID (Hahm et al., 2025) is a framework designed specifically for Korean court judgments. It provides a dataset for detecting PII in Korean legal documents. While Thunder-DeID represents a significant step forward, its scope has two notable limitations. First, its dataset is specifically focused on three types of criminal cases, including sexual assault, assault, and fraud. Second, its annotation scheme does not employ a BIO (Beginning, Inside, Outside) tagging, which poses challenges in delineating the boundaries of consecutively occurring entities. This highlights the need for a more comprehensive dataset that covers a wider range of legal domains and an annotation approach suitable for robust entity boundary detection.

Our Position. Based on the limitations in prior research, our work addresses the critical lack of a comprehensive benchmark by introducing a large-scale, high-quality dataset covering a wide spectrum of Korean legal categories. We emphasize the dataset's reliability, validated through rigorous annotation protocols and high inter-annotator agreement. Our approach prioritizes the creation of a foundational public resource and introduces an effective and practical solution for de-identification within the Korean legal system.

3 Preliminaries

3.1 KLUEBERT-NER.

KLUEBERT is a pre-trained BERT model for the Korean language (Park et al., 2021), distributed under the CC BY-NC-SA 4.0 License. KLUEBERT is designed to perform eight Korean natural language understanding tasks, including topic classification, semantic textual similarity, natural language inference, named entity recognition, relation extraction, dependency parsing, machine reading comprehension, and dialogue state tracking. It employs a morpheme-based subword tokenization scheme which first tokenizes raw text into morphemes using a morphological analyzer, followed by applying Byte Pair Encoding (BPE) (Sennrich et al., 2015). After building the vocabulary, KLUEBERT uses only the BPE model during inference, allowing word sequences to be tokenized to reflect morphemes without requiring a morphological analyzer. This approach improves both usability and processing speed.

3.2 Conditional Random Field.

A Conditional Random Field (CRF) is a popular method for sequence labeling tasks. It learns an independent per-position classifier that maps each $\mathbf{x}$ to y_s, where y is a label vector, $y = y_0, y_1, \ldots, y_T$. The CRF models the conditional probability distribution $p(y|\mathbf{x})$ directly. This modeling approach ensures that dependencies involving only variables in $\mathbf{x}$ do not affect the conditional model, allowing for a much simpler structure compared to joint models (Sutton et al., 2012).

4 Proposed Method

4.1 Datasets

Our training and test data are derived from a combination of three distinct sources. The datasets used are a multi-turn SNS conversation dataset, our newly curated Court Judgment Dataset, and publicly available Thunder-DeID Dataset (Hahm et al., 2025). This integration yields a comprehensive PII dataset of approximately one million instances.

A key challenge is that court judgments must be de-identified before public release, as mandated by Supreme Court Regulations [1]. To address this, we develop a pipeline to process these documents into a usable format. In addition, we collect and process the SNS conversation dataset using a data generation logic to further enhance our training data. The following subsections detail our data collection process, the masking rules applied, the synthetic data generator, and the specific adjustments made for each dataset.

4.1.1 Data Collection

We collect a total of 3,246,886 utterances, including an SNS multi-turn conversation dataset, 2,000 case court judgments, and 4,500 sentences from the Thunder-DeID Dataset. The combined dataset consists of 1,091,998 instances, which are 908,422 of SNS, 138,576 of Court Judgement, and 45,000 of Thunder-DeID Dataset.

SNS Dataset This is the 'Korean SNS Multi-turn Conversation Data' published on AI Hub[1], consisting of conversation data built around 9 conversation topics involving 2 or 3 participants in

[1]This research used SNS Multi-turn Conversation datasets from 'The Open AI Dataset Project (AI-Hub, S. Korea)'. This data information can be accessed through 'AI-Hub (www.aihub.or.kr)'.

multi-turn interactions. The topics are Health and Food & Beverage, Economy and Society, Science and Technology, Culture, Lifestyle, and Leisure, Beauty and Fashion, Sports and E-sports, Travel and Attractions, Politics, and Content preference. Politics accounts for 1.84%, Economy and Society for 21.68%, and the remaining topics each comprise approximately 10%. Speaker composition ratios are 90.96% for two-person conversations and 9.04% for three-person conversations.

Court Judgment Dataset We collect court judgment publicly available on the Ministry of Government Legislation's National Law Information Center by legal field. We collect a total of 2,000 cases, with 50 cases each collected from 39 of the 44 classified legal fields (except for civil law, which had 100 cases). To broadly encompass characteristics of case content—such as crime scenarios and frequently occurring case types that may vary by legal field—we collected a mix of lower court and Supreme Court cases across all fields except those with fewer than 50 publicly available cases: Part 2 (National Assembly), Part 12 (Civil Defense · Firefighting), Part 22 (Tobacco · Ginseng), Part 29 (Industrial standards · measurements) and Part 44 (Foreign Affairs). For precedents with overlapping case numbers, the final version from the lower court was selected to maintain consistency.

Thunder-DeID Dataset (Hahm et al., 2025) Thunder-DeID Dataset, created by the Graduate School of Data Science at Seoul National University, is designed for the de-identification within Korean court judgment available under the CC BY-NC-SA 4.0 License. It consists of a labeled court judgment dataset and a named entity list dataset. The Thunder-DeID Dataset comprises a total of 4,500 sentences, with 1,500 sentences for each of the three case types: sexual assault, assault and fraud. This dataset provides 27,402 annotated PII entities labeled with 595 unique placeholders.

4.1.2 Data Masking Rule

To ensure both regulatory compliance and annotation consistency, we develop a unified annotation scheme for de-identification. A primary requirement is adherence to the Supreme Court Administrative Office's 'Standards for Anonymization for Viewing and Copying Judgment'[2]. Because datasets such as SNS conversations from AI Hub use their own distinct PII masking schemes, we implement a standardized process to transform all collected data to conform to our new annotation framework.

We prepare an initial annotation scheme using entity categories that meet the document 'Standards for Anonymization for Viewing and Copying Court Judgments', while also being deemed capable of appropriately segmenting the types of PII within the content of the court judgment data. In addition, through multiclass precision, recall, and F1 score analysis, we merge entity categories that exhibited high confusion, such as school and department, company and business division, and URL and web-mail. Finally, we establish a BIO annotation scheme comprising 11 types of PII, as follows: name, address, number, bank name, account number, security code, school, company, URL, email, ID. This entity category functions as a placeholder, and data corresponding to each category is inserted by the synthetic data generator.

4.1.3 Synthetic Data Generator

To generate appropriate data for each PII category, a dedicated generator is required for each category. Considering that the data consists of Korean court judgment texts, these generators are created based on Korean statistical data. Datasets for surname, given name, address, school, department, company, work department, URL, and bank name are constructed using information collected from sources such as Statistics Korea. Additionally, data generation logic is implemented for each of the following: ID, email, phone number, account number, and security code. This logic ensures that the generated data conforms to formats used in Korea while maintaining randomness.

The data generation process is performed as shown in Figure 1. Structural alignment is first applied to the SNS Dataset via category specialization and to the Thunder-DeID Dataset via category aggregation, ensuring they share the same annotation scheme as the Court Judgment Dataset. When a PII placeholder (#@(\w+)#) is detected in each dataset, contextually appropriate synthetic data is inserted using the generator corresponding to that PII category. Then, the PII category type and its start and end indices are added to the label information. This data is then aligned and BIO-labeled at the token level using the KLUEBERT tokenizer.

[2]Supreme Court Trial Regulation No. 1778 revised on August 9, 2021.

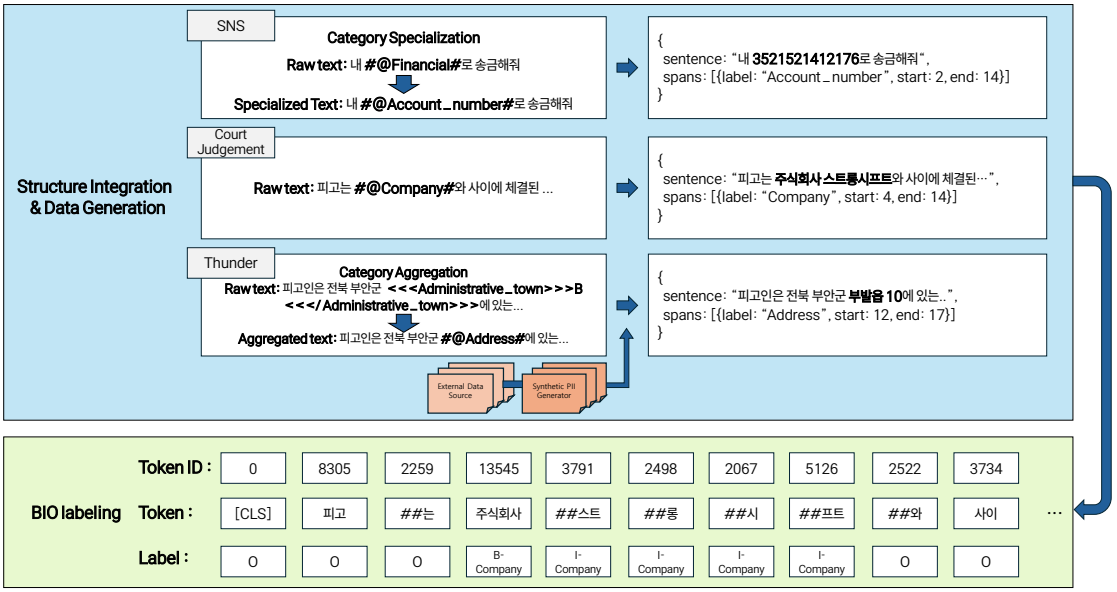


Figure 1: Data Generation and BIO labeling

4.1.4 SNS PII Dataset

The SNS multi-turn dataset has its own annotations, which we transform into a format compatible with our scheme. We retain categories that directly matched our method, such as names and phone numbers. For categories with broader scopes—like finance, affiliation, and account—we subdivide them into our more granular labels. This subdivision is processed in batches using OpenAI's GPT-4o-mini. After filtering out sentences that lacked PII, we collect the correctly processed data along with existing data that did not require subdivision. This process yields a total of 908,422 sentence-level instances, containing 970,129 PII entities.

4.1.5 Thunder-DeID Dataset

The Thunder-DeID Dataset presents several integration challenges, as it used a different annotation scheme and tokenizer.

Therefore, a pre-processing step is required to integrate it into our dataset. The Thunder-DeID Dataset placeholders are {0: 'IT_company', 1: 'O', 2: 'internet_café',..., 594: 'mobile_phone_case_store'}. We classify its 595 placeholder types according to our 11 categories and match them using a custom *'Thunder-To-Ours'* mapping rule (e.g., 'IT_company' → 'company', 'internet_café' → 'address', ..., 'mobile_phone_case_store' → 'address'). Additionally, this integration is necessary because Thunder-DeID uses a custom tokenizer (Mecabko + BPE) and a simple tagging scheme, whereas our approach uses the KLUEBERT wordpiece tokenizer and a BIO annotation scheme.

We align the tokens to match the differences with our tokenizer and restructured the labels of each dataset to the BIO format, making them compatible for integration into our dataset. Finally, a total of 4,500 sentence-level instances are constructed, containing 27,402 PII entities in the dataset. However, the resulting 4,500 sentence-level instances were significantly fewer than the other two datasets. To address this data imbalance, we expanded the dataset to 45,000 instances through data augmentation using placeholder replacement. In contrast to Court Judgement dataset, the entity distribution is concentrated in three specific fields (57.5%, 21.6%, and 20.9%, respectively). About the details about data augmentation of Thunder-DeID Dataset, please refer to Appendix E.

4.1.6 Court Judgment PII Dataset

We create the Court Judgment PII Dataset by collecting court judgments published by the National Law Information Center and processing them according to our annotation scheme. During this process, eight annotators follow annotation guidelines based on the Court Administra-

tion Office's anonymization standards and provide mutual feedback to improve the agreement between annotators. To measure this inter-annotator agreement (IAA), we calculate the Fleiss' Kappa score (Landis and Koch, 1977). This metric statistically measures the agreement among three or more evaluators on categorical data. The measured Fleiss' Kappa score is 0.7352, indicating substantial agreement according to the standard interpretation of the scale. This high level of consistency among annotators signify the reliability and quality of the data labeling. Finally, we construct a dataset of 138,576 sentence-level instances from 2,000 court judgments, containing 51,245 PII entities. The distribution of these entities across the 39 law fields is more balanced and covers more diverse fields compared to the prior work. For detailed statistics and distribution figures, please refer to Figure 3 in Appendix H.

4.2 KLUEBERT-CRF

The de-identification for court judgment can be defined as a Named Entity Recognition (NER) task that accurately detects entity boundaries by understanding the context of entire sentences. Korean, being an agglutinative language, has ambiguous word boundaries, and legal documents feature complex syntactic structures composed of long sentences and specialized terminology. To effectively handle these domain characteristics, we adapt KLUEBERT—a pre-trained transformer encoder model trained on Korean corpora (Park et al., 2021)—as our baseline. However, KLUEBERT alone struggles to sufficiently reflect inter-token dependencies, leading to errors such as impossible tag transitions (e.g., 'I-' tag following 'O') or inconsistent labeling of the same entity. Consequently, even with high token-level micro F1 scores, a single misplaced boundary degrades the Entity-level Micro F1 score.

To address this issue, we add a Conditional Random Field (CRF) layer (Zheng et al., 2015) on top of the final hidden layer. CRF learns not only the classification probability of each token but also the transition probability between adjacent labels, ensuring the consistency of the entire sequence. Furthermore, by employing global decoding using the Viterbi algorithm, we exclude logically impossible label sequences in the BIO tagging scheme (e.g., 'B-name' followed by 'I-address'), reducing errors such as incorrect boundaries ('O' followed by 'I-' tag) and inconsistent labeling. As a result, the entity-level micro F1 score is improved by 2.98%.

Additionally, after training evaluation, we conduct an analysis of cases where the model incorrectly tokenized and resulted in incorrect BIO labeling. We select 26 pieces of vocabulary that could potentially be PII from tokens that caused errors in approximately 5% of cases (600 out of 11,653). These pieces of vocabulary are duplicates in the tokenizer's vocabulary due to incorrect tokenization in the same cases. We add these terms as properly tokenized units to the model's vocabulary and fine-tuned our model accordingly.

Our KLUEBERT-CRF model, with approximately 110 million parameters, features a lightweight structure compared to other models in the legal domain. Compared to Thunder-DeID model (360M parameters), its smaller size approximately 68% offers practical advantages in memory usage during deployment in court systems.

4.3 End-to-End Masking Framework

For each PII symbol, synthetic data is inserted based on statistical data to construct a PII dataset, and the KLUEBERT-CRF model is trained using this dataset. When a court judgment document is provided into the model, it identifies entities within the documents that could be identified as PII. PII with specific patterns (e.g., resident registration numbers, vehicle registration numbers) undergoes secondary filtering using regular expressions to detect PII within the document. The detected PII is then masked in accordance with court anonymization regulations. Specifically, to preserve the semantic consistency of the document, identical PII entities appearing multiple times within the same document are replaced with the same unique masking symbol. Finally, the de-identified court judgment document is provided.

5 Experiments

5.1 Experiment Setting

Dataset. We use our PII dataset described in section 4.1 for training and test. To ensure a fair evaluation and prevent data leakage for the Thunder-DeID dataset which is performed augmentation, we split the dataset based on the unique identifiers of the original source sentences. The dataset is divided into 70% train, 20% validation, and 10% test.

Models and Baselines. We fine-tune five models, KLUEBERT (baseline), KLUEBERT-CRF, with

Model	# of parameters	Token Binary F1	Token Micro F1	Entity Binary F1	Entity Micro F1	Overlap F1	Intermediate F1
KLUEBERT	110M	0.9942	0.9906	0.9509	0.9451	0.9753	0.9739
Kanana-1.5	2.1B	0.5354	0.2889	0.4504	0.2495	0.3148	0.3112
Qwen-2.5	1.5B	0.6931	0.5863	0.5425	0.5233	0.6531	0.6074
Thunder-DeID	360M	0.9970	0.9929	0.9614	0.9608	0.9850	0.9844
KLUEBERT-CRF (Ours)	**110M**	**0.9989**	**0.9988**	**0.9925**	**0.9923**	**0.9952**	**0.9951**

Table 1: Performance comparison on the combined dataset (3 Datasets). Our model, KLUEBERT-CRF, demonstrates superior performance across all metrics.

110M parameters, Kanana-1.5-2.1b (Team et al., 2025), a bilingual LLM, Qwen-2.5-1.5b (Team, 2024), a multilingual LLM, and the Thunder-DeID (Hahm et al., 2025), encoder-only model initially fine-tuned through Thunder-DeID dataset. For more details, please refer to Appendix I.

Evaluation Metrics. We use six metrics to evaluate the performance of the models including token-level binary F1, token-level micro F1, entity-level binary F1, entity-level micro F1 (same as strict F1) (Dernoncourt et al., 2017; Takahashi et al., 2022), overlap F1, and intermediate F1 (Segura-Bedmar et al., 2013).

The binary F1 metrics measure the model's ability to correctly classify whether tokens or entities contain PII, without considering the specific types of PII. In contrast, the micro F1 metrics measure the model's ability to correctly classify the specific types of PII that tokens or entities represent.

These metrics are applied at both the token and entity levels. In the case of entity-level F1 metrics, the micro F1 deems a prediction correct only when both the exact boundary span and the specific entity type match the ground truth, whereas the binary F1 evaluates whether the exact boundary span matches and the entity is correctly classified as PII.

The overlap F1 considers a match valid if there is any boundary overlap with a ground-truth entity of the same type, whereas the intermediate F1 requires at least 50% token overlap for a match.

5.2 Main Result

Table 1 shows the performance of our models compared to other three models, KLUEBERT, Thunder-DeID (360M), Kanana-1.5, and Qwen-2.5. We train each model on a combined training set consisting of three datasets and evaluate them on a test set that also includes all three datasets. The results show that our model consistently outperforms the other models in all metrics we used.

The token-level F1 scores demonstrate that incorporating a CRF layer structure to KLUEBERT and expanding the tokenizer with legal domain-specific vocabulary improves PII token classification performance relative to the baseline KLUEBERT model. Furthermore, examining the entity-level, overlap, and intermediate F1 scores reveals that clearly distinguishing each PII entity and identifying boundaries between entities demonstrates improved performance.

5.3 Robustness to Unseen Data

We further investigate the quality of our proposed dataset by assessing its ability to generalize to unseen data. To this end, we evaluated our model, trained on Court Judgment and SNS datasets (excluding Thunder-DeID), on the Thunder-DeID test set. We compare its performance against the Thunder-DeID model, which serves as a reference benchmark for expected performance having been trained on its native Thunder-DeID dataset. Achieving performance comparable to this native baseline demonstrates that the diversity and quality of our dataset enable models to be robust even on specialized, previously unseen legal texts.

Table 2 presents the performance degradation when tested on Thunder test dataset. This performance gap is natural as the baseline (Thunder-DeID) was trained on its dataset. Furthermore, this gap is also attributable to the interpretations and applications of the masking guidelines. While both our method and the Thunder-DeID aim to follow the official guidelines[3], our approach employs a different annotation scheme and data processing methodology. Despite these constraints, the fact that our model maintains performance comparable to the native baseline validates the robustness and generalization capability of our proposed dataset on unseen data. A score of '-' indicates a value less

[3] Supreme Court Trial Regulation No. 1778 revised on August 9, 2021.

Model	Only Thunder-DeID Dataset						
	# of parameters	Token Binary F1	Token Micro F1	Entity Binary F1	Entity Micro F1	Overlap F1	Intermediate F1
Thunder-DeID	360M	**0.9472**	0.9223	-	-	0.3682	0.3682
KLUEBERT-CRF (Ours)	**110M**	0.9268	**0.9299**	**0.4201**	**0.3658**	**0.4484**	**0.4421**

Table 2: Performance of our models on the Thunder-DeID test set after being trained exclusively on our proposed dataset. The performance drop is due to the differing masking guidelines between the two datasets.

Model	# of parameters	Token Binary F1	Token Micro F1	Entity Binary F1	Entity Micro F1	Overlap F1	Intermediate F1
KLUEBERT	110M	0.9916	0.9861	0.9509	0.9451	0.9753	0.9739
Kanana-1.5	2.1B	0.2870	0.0803	0.0556	0.0192	0.1436	0.1432
Qwen-2.5	1.5B	0.4826	0.3328	0.2636	0.1997	0.4020	0.3900
Thunder-DeID	360M	0.9982	0.9979	0.9052	0.8934	0.9062	0.9044
KLUEBERT-CRF (Ours)	**110M**	**0.9998**	**0.9997**	**0.9935**	**0.9928**	**0.9946**	**0.9946**

Table 3: Performance comparison on the Thunder-DeID dataset only. Our model shows robust performance, outperforming the original Thunder-DeID model on its own data.

than 0.01, due to low entity recognition capability.

5.4 Comparative Analysis

To further analyze the impact of our diverse dataset and to demonstrate the model's generalization capability, we train our KLUEBERT-CRF model on the full combined dataset, and then evaluate the performance exclusively on the Thunder-DeID test dataset. After that, we compare it against other models. More individual test results are in Appendix D.

As shown in Table 3, our model, when trained on the comprehensive combined dataset, significantly outperforms the native Thunder-DeID model on its own test data. This result demonstrates two key points. First, it confirms that the performance drop observed in the robustness test was indeed due to the domain shift and differing annotation schemes, not a fundamental limitation of our dataset. Second, it highlights the benefit of training on a more diverse and larger-scale dataset. By incorporating data from various legal fields and conversational contexts, our model learns a more generalized representation of PII, enabling it to achieve superior performance even on specialized, narrowly focused datasets.

5.5 Qualitative Analysis

The confusion matrices in Appendix C demonstrate the model's robust classification capabilities, indicating consistent accuracy across all datasets with minimal inter-class confusion.

6 Discussion

A key contribution of our work is the introduction of a high-quality benchmark dataset, which addresses a critical data scarcity issue in the Korean legal tech landscape. By enabling researchers and practitioners to develop and validate robust de-identification models, our dataset can help automate a process currently dominated by manual labor. This automation is necessary for increasing the public availability of court judgments, which will enhance judicial transparency.

7 Conclusion

In this paper, we propose a BERT-based framework to address the challenge of automated de-identification for Korean court judgments, balancing judicial transparency with privacy protection. Our primary contribution is the creation of a large-scale PII dataset of approximately one million instances, constructed by combining a new, comprehensive Korean legal document dataset with SNS conversation dataset and the Thunder-DeID dataset. The legal dataset, spanning 39 legal fields, demonstrates high quality and reliability, validated by a Fleiss' Kappa score of 0.7352 for IAA. Experimental results show that our proposed KLUEBERT-CRF model, trained on this diverse dataset, achieves state-of-the-art performance, setting a new benchmark for de-identification for Korean legal documents.

Limitations

Supreme Court Regulation

According to Supreme Court regulations, we cannot access the original (non-anonymized) court judgments. To address this limitation, human annotators processed anonymized court judgment by annotating PII entities and inserting synthetic data into the masked placeholders by our synthetic data generator. However, this synthetic data cannot perfectly replicate the characteristics of actual original court judgments. Furthermore, without access to the original court judgments, we cannot conduct comparative experiments to analyze the differences between the original and our processed court judgments. To minimize these limitations, we plan to analyze dependencies between PII entities and refine our synthetic data generator to achieve greater precision.

Generalizability and Resources Adaptation.

While the proposed data processing pipeline and model architecture are fundamentally language-agnostic, the current implementation utilizes resources specialized for the Korean language, such as KSS for sentence segmentation and KLUE-BERT for embedding. Therefore, extending this framework to other languages does not require structural changes but necessitates the substitution of these language-specific components with equivalent tools suitable for the target language. For example, it is possible to replace tools like the Korean sentence segmenter (KSS) with multilingual tools (e.g., spaCy, NLTK) and swap KLUE-BERT with language-specific encoders suitable for the target language (e.g., UmBERTo(Parisi et al., 2020) for Italian, RoBERTa(Liu et al., 2019) for English, CamemBERT(Martin et al., 2020) for French).

Acknowledgments

This work was supported in part by the Institute of Information & Communications Technology Planning & Evaluation (IITP)-ICT Creative Consilience Program grant funded by the Korea government (MSIT) (IITP-2025-RS-2020-II201819); and in part by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (RS-2021-NR060143)

References

Court Administration. 2019. National assembly of korea.

Court Administration. 2025. National court administration of korea. Technical Report ISP-, Court of Korea, Seoul.

Sergei Bogdanov, Alexandre Constantin, Timothée Bernard, Benoit Crabbé, and Etienne P Bernard. 2024. Nuner: Entity recognition encoder pre-training via llm-annotated data. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11829–11841.

Franck Dernoncourt, Ji Young Lee, Ozlem Uzuner, and Peter Szolovits. 2017. De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association*, 24(3):596–606.

Vipin Gupta, Ian C MacMillan, and Gita Surie. 2004. Entrepreneurial leadership: developing and measuring a cross-cultural construct. *Journal of business venturing*, 19(2):241–260.

Sungeun Hahm, Heejin Kim, Gyuseong Lee, Hyunji Park, and Jaejin Lee. 2025. Thunder-deid: Accurate and efficient de-identification framework for korean court judgments. *arXiv preprint arXiv:2506.15266*.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Stefan Larson, Nicole Cornehl Lima, Santiago Pedroza Diaz, Amogh Manoj Joshi, Siddharth Betala, Jamiu Tunde Suleiman, Yash Mathur, Kaushal Kumar Prajapati, Ramla Alakraa, Junjie Shen, and 1 others. 2024. De-identification of sensitive personal data in datasets derived from iit-cdip. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21494–21505.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

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

Louis Martin, Benjamin Muller, Pedro Ortiz Suarez, Yoann Dupont, Laurent Romary, Éric Villemonte de La Clergerie, Djamé Seddah, and Benoît Sagot. 2020. Camembert: a tasty french language model. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 7203–7219.

Stephen Mayhew, Terra Blevins, Shuheng Liu, Marek Šuppa, Hila Gonen, Joseph Marvin Imperial, Börje F Karlsson, Peiqin Lin, Nikola Ljubešić, Lester James Miranda, and 1 others. 2024. Universal ner: A gold-standard multilingual named entity recognition benchmark. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4322–4337.

Loreto Parisi, Simone Francia, and Paolo Magnani. 2020. Umberto: an italian language model trained with whole word masking. https://github.com/musixmatchresearch/umberto.

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Jiyoon Han, Jangwon Park, Chisung Song, Junseong Kim, Yongsook Song, Taehwan Oh, and 1 others. 2021. Klue: Korean language understanding evaluation. *arXiv preprint arXiv:2105.09680*.

Maksim Savkin, Timur Ionov, and Vasily Konovalov. 2025. Spy: Enhancing privacy with synthetic pii detection dataset. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 236–246.

Isabel Segura-Bedmar, Paloma Martínez, and María Herrero-Zazo. 2013. Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013). In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 341–350.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Charles Sutton, Andrew McCallum, and 1 others. 2012. An introduction to conditional random fields. *Foundations and Trends® in Machine Learning*, 4(4):267–373.

Latanya Sweeney. 2002. k-anonymity: A model for protecting privacy. *International journal of uncertainty, fuzziness and knowledge-based systems*, 10(05):557–570.

Maria Irena Szawerna, Simon Dobnik, Ricardo Muñoz Sánchez, Therese Lindström Tiedemann, and Elena Volodina. 2024. Detecting personal identifiable information in swedish learner essays. In *Proceedings of the Workshop on Computational Approaches to Language Data Pseudonymization (CALD-pseudo 2024)*, pages 54–63.

Kanae Takahashi, Kouji Yamamoto, Aya Kuchiba, and Tatsuki Koyama. 2022. Confidence interval for micro-averaged f 1 and macro-averaged f 1 scores. *Applied Intelligence*, 52(5):4961–4972.

Kanana LLM Team and 1 others. 2025. Kanana: Compute-efficient bilingual language models. *arXiv preprint arXiv:2502.18934*.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, Guoyin Wang, and Chen Guo. 2025. Gpt-ner: Named entity recognition via large language models. In *Findings of the association for computational linguistics: NAACL 2025*, pages 4257–4275.

Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip H. S. Torr. 2015. Conditional random fields as recurrent neural networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE.

Appendix

A Prompt on Category Specialization

A.1 Scheme of Prompt on GPT-4o-mini

(a) Original Prompt (Korean)

```
{
   "role": "system",
      "content": "너는 주어진 대화에서 마스킹된 위치의 가명정보 문자열을 생성하는 AI assistant야. 마스킹된 원래 단어는 구체적인 정보입니다. 전체 대화 맥락을 고려하여 자연스러운 가명정보를 단일 단어로 생성하세요."
},
{
   "role": "user",
   "content":
      "T1: 카드사에 전화를 해봐
      T1: 지금
      T2: 홈페이지에적혀있었어요
      T2: 영업점 직접방문
      T3: 저나해봐
      T3: 안풀린건지
      T4: 고객번호. #@소속#"
}
```

(b) Translated Prompt (English)

```
{
   "role": "system",
  "content": "You are an AI assistant that generates pseudonymized information strings for masked positions in a given conversation. The original masked words represent specific information. Generate the pseudonymized information as a natural single word, considering the entire conversation context."
},
{
   "role": "user",
   "content":
      "T1: Call the credit card company.
      T1: Now
      T2: It was written on the website.
      T2: Visit a branch in person
      T3: Give them a call
      T3: To check whether the suspension has been lifted yet.
      T4: Customer Number. #@affiliation#"
}
```

A.2 Example of Prompt Response on GPT-4o-mini


```
{
  "id": "batch_req_68354c936464-8190bba6-40c370e198e2",
   "custom_id": "request-18",
   "response": {
     ...
     "body": {
       ...
       "choices": [{
       "index": 0,
       "message": {
         "role": "assistant",
         "content": "'department'",
         "refusal": null,
       },
       ...
       }],
     ...
     }
   },
  "error": null
}
```


B Prompt on Kanana-1.5

B.1 Scheme of Prompt on Kanana-1.5

(a) Original Prompt (Korean)

```
### 지시: 주어진 문장에서 모든 개인 식별 정보(PII)를 찾아서, 각 PII의 종류, 시작 인덱스, 끝 인덱스를 JSON 형식으로 추출하세요.

### 입력: {}

### 답변: {}
```

(b) Translated Prompt (English)

```
### Instruction: Find all personally identifiable information (PII) in the given sentence and extract each PII's type, start index, and end index in JSON format.

### Input: {}

### Response: {}
```

B.2 Example of Prompt Response on Kanana-1.5

(a) Original Prompt (Korean)

지시: 주어진 문장에서 모든 개인 식별 정보(PII)를 찾아서, 각 PII의 종류, 시작 인덱스, 끝 인덱스를 JSON 형식으로 추출하세요.

입력: 공소사실의 요지. 피고인은 경기도 이천시 부발읍 10에서 '주식회사더블에스메디칼'을 운영하는 사람이다.

답변: [{"label": address, "start": 23, "end": 29}, {"label": company, "start": 33, "end": 44}]

(b) Translated Prompt (English)

Instruction: Find all personally identifiable information (PII) in the given sentence and extract each PII's type, start index, and end index in JSON format.

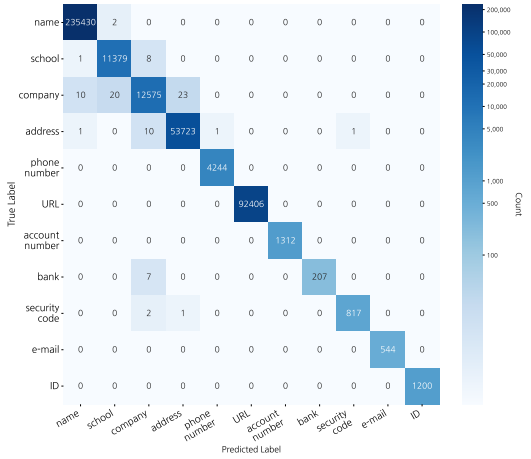
Input: Summary of the Indictment. The defendant is the person operating 'Double S Medical Corporation' at 10 Bubal-eup, Icheon-si, Gyeonggi-do.

Response: [{"label": address, "start": 23, "end": 29}, {"label": company, "start": 33, "end": 44}]

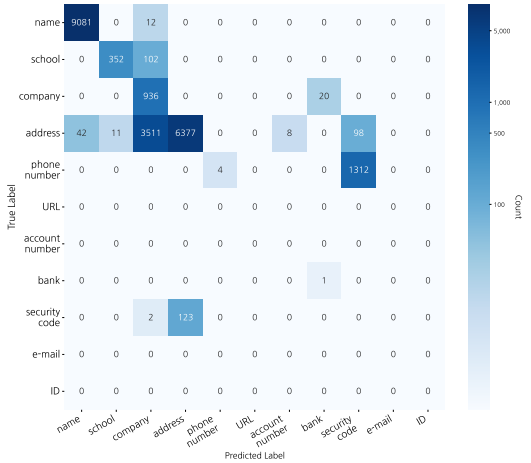
C Qualitative Analysis

Confusion matrix of KLUEBERT-CRF on the main result, the robustness result and the ablation study result. The main results were tested on a mixed dataset comprising all three datasets for both the train and test datasets. The robustness results were tested on a mixed dataset of the two datasets excluding the Thunder-DeID Dataset for the train dataset, and the Thunder-DeID Dataset for the test dataset. The ablation study results were tested on all three datasets for the train dataset and the Thunder-DeID Dataset for the test dataset.

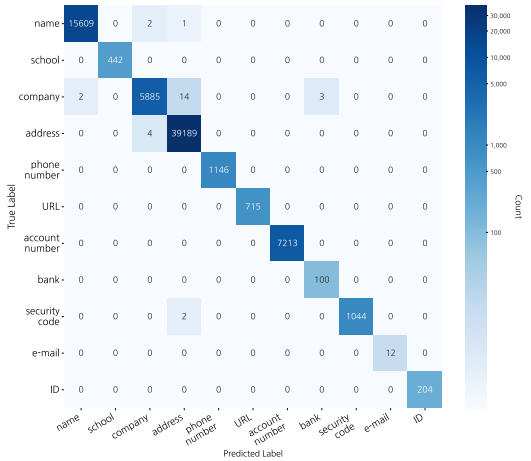
In the confusion matrix for figure 2a, 2b, and 2c, respectively, the vertical axis represents the actual labels, and the horizontal axis represents the predicted labels. Therefore, the numbers on the diagonal indicate cases where the predictions match the actual labels, while the other numbers indicate cases where the predictions do not match the actual labels. This helps to analyze how each label tends to be incorrectly classified into other labels.



(a) In the case of train on three dataset and test on three dataset.



(b) In the case of train on two dataset and test on the left dataset (Thunder-DeID).



(c) In the case of train on three dataset and test on the Thunder-DeID dataset.

Figure 2: Per class analysis.

Model	# of parameters	Token Binary F1	Token Micro F1	Entity Binary F1	Entity Micro F1	Overlap F1	Intermediate F1
KLUEBERT	110M	0.9946	0.9924	0.8208	0.8110	0.9109	0.9015
Thunder-DeID	360M	0.9964	0.9952	0.9297	0.9255	**0.9561**	**0.9552**
KLUEBERT-CRF (Ours)	**110M**	**0.9981**	**0.9981**	**0.9447**	**0.9438**	0.9536	0.9518

Table 4: Performance comparison on the individual dataset (Court Judgement PII Dataset).

Model	# of parameters	Token Binary F1	Token Micro F1	Entity Binary F1	Entity Micro F1	Overlap F1	Intermediate F1
KLUEBERT	110M	0.9961	0.9928	0.9813	0.9810	0.9926	0.9922
Thunder-DeID	360M	0.9964	0.9871	0.9645	0.9643	0.9882	0.9877
KLUEBERT-CRF (Ours)	**110M**	**0.9988**	**0.9986**	**0.9945**	**0.9945**	**0.9973**	**0.9972**

Table 5: Performance comparison on the individual dataset (SNS PII Dataset).

Model	# of parameters	Token Binary F1	Token Micro F1	Entity Binary F1	Entity Micro F1	Overlap F1	Intermediate F1
KLUEBERT	110M	0.9916	0.9861	0.8047	0.7679	0.8903	0.8848
Thunder-DeID	360M	0.9982	0.9979	0.9052	0.8934	0.9062	0.9044
KLUEBERT-CRF (Ours)	**110M**	**0.9998**	**0.9997**	**0.9935**	**0.9928**	**0.9946**	**0.9946**

Table 6: Performance comparison on the individual dataset (Thunder-DeID Dataset).

D Experiment Results for Individual Dataset

We conducted a detailed performance breakdown for each individual dataset to ensure transparency and reproducibility. Table 4, 5, and 6 show the individual test results for the Court Judgment, SNS, and Thunder-DeID datasets, respectively.

E Performance comparison of data augmentation

We would like to clarify our data processing strategy regarding "augmentation." For our primary datasets (Court Judgment and SNS), we did not apply data augmentation (i.e., generating multiple synthetic variations for a single instance to increase dataset size). Instead, we performed a 1:1 replacement, where each masked placeholder in the source text was replaced with a single contextually appropriate synthetic entity to construct the training data. This approach was chosen to prevent the model from overfitting to specific sentence structures, which can occur with excessive augmentation.

However, for the Thunder-DeID dataset, we applied data augmentation by generating multiple synthetic variations for the PII placeholders to enhance the diversity of this relatively smaller dataset. So, we conducted an comparative analysis to analyze the performance change on the Thunder-DeID dataset with and without this augmentation. The experimental results are currently shown in Table 7, 8, 9, and 10.

As demonstrated in these tables, applying data augmentation generally yields consistent performance improvements. However, an exception is observed in the individual test on the Court Judgment PII Dataset (Table 8), where the performance slightly decreases after data augmentation. This specific gap is attributable to the interpretations and applications of the masking guidelines between Court Judgement PII Dataset and Thunder-DeID Dataset. While both our method and the Thunder-DeID aim to follow the official guidelines[4], our approach employs a different annotation scheme and data processing methodology.

F KSS

KSS is a Korean string processing suite that provides various functions for processing Korean strings. We used this to segment the annotated case court judgement data from the case level to the

[4]Supreme Court Trial Regulation No. 1778 revised on August 9, 2021.

Model	# of parameters	Token Binary F1	Token Micro F1	Entity Binary F1	Entity Micro F1	Overlap F1	Intermediate F1
KLUEBERT-CRF w/o augmentation	110M	0.9977	0.9970	0.9836	0.9808	0.9892	0.9887
KLUEBERT-CRF w/ augmentation	110M	**0.9989**	**0.9988**	**0.9925**	**0.9923**	**0.9952**	**0.9951**

Table 7: Performance comparison of data augmentation on Combined Dataset.

Model	# of parameters	Token Binary F1	Token Micro F1	Entity Binary F1	Entity Micro F1	Overlap F1	Intermediate F1
KLUEBERT-CRF w/o augmentation	110M	**0.9986**	**0.9985**	**0.9711**	**0.9703**	**0.9792**	**0.9771**
KLUEBERT-CRF w/ augmentation	110M	0.9981	0.9981	0.9447	0.9438	0.9536	0.9518

Table 8: Performance comparison of data augmentation on Court Judgement PII Dataset.

Model	# of parameters	Token Binary F1	Token Micro F1	Entity Binary F1	Entity Micro F1	Overlap F1	Intermediate F1
KLUEBERT-CRF w/o augmentation	110M	0.9987	0.9985	0.9943	0.9943	0.9971	0.9971
KLUEBERT-CRF w/ augmentation	110M	**0.9988**	**0.9986**	**0.9945**	**0.9945**	**0.9973**	**0.9972**

Table 9: Performance comparison of data augmentation on SNS PII Dataset.

Model	# of parameters	Token Binary F1	Token Micro F1	Entity Binary F1	Entity Micro F1	Overlap F1	Intermediate F1
KLUEBERT-CRF w/o augmentation	110M	0.9955	0.9934	0.9231	0.9031	0.9451	0.9422
KLUEBERT-CRF w/ augmentation	110M	**0.9998**	**0.9997**	**0.9935**	**0.9928**	**0.9946**	**0.9946**

Table 10: Performance comparison of data augmentation on Thunder-DeID Dataset.

sentence level. The parameter settings for the sentence segmentation function used are as follows.

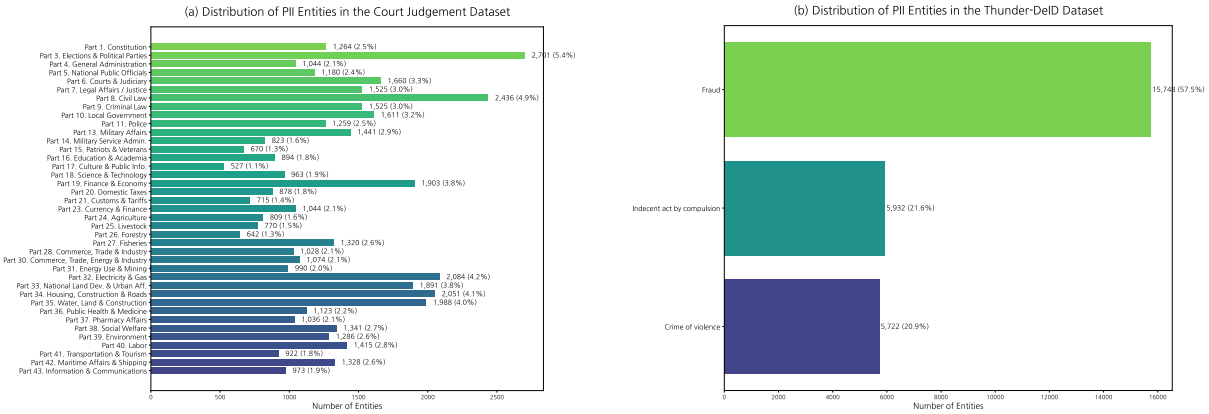
```
split_sentences(
    text: str,
    backend: str = "mecab",
    num_workers: str = "auto",
    strip: bool = True,
    ignores: List[str] = None,
)
```

G Annotators

Eight annotators, including the author, contributed to the annotation process over 2 weeks. We informed the annotators that the processed data would be used for court judgment de-identification. To determine appropriate compensation, the author completed preliminary tasks before employment to estimate the time requirements, and the pay was set based on this assessment. Annotators received 1,000 KRW per court judgment case. Considering their average working hours, their pay was higher than the legal minimum wage (10,030 KRW per hour), so we consider it appropriate.

H Distribution Details

We provide the specific distribution of PII entities across each legal domain. Figure 3a presents the PII entity distribution across the 39 legal domains in our court judgment dataset. Figure 3b presents the entity distribution across the three types of court cases in the Thunder-DeID dataset.



(a) The entry distribution in our Court Judgment dataset.

(b) The entry distribution in the Thunder-DeID dataset.

Figure 3: Distribution of PII entities in the Court Judgment and Thunder-DeID datasets.

I Fine-tuning

We fine-tune KLUEBERT-CRF, KLUEBERT, Thunder-DeID, Kanana-1.5, and Qwen-2.5 on PII Entity Recognition task using the dataset split for train, validation, and test described in section 4.1. We fully fine-tune KLUEBERT-CRF, KLUEBERT, and Thunder-DeID and we LoRA fine-tune Kanana-1.5 and Qwen-2.5, setting hyperparameters while considering hardware capacity. More details about model specifications and hyperparameter are described in Table 11

J Annotated Data Sample

In this section, we present the Figure 4 and 5 the sample of both publicly available anonymized court judgement and the court judgement annotated by our annotators. We convert PII to unique symbols in our annotation scheme. The attached court judgment is an excerpt from "2014 가합38116". In publicly available anonymized court judgments, words that would be PII if not anonymized were replaced with symbols according to the annotation scheme.

Aspect	KLUEBERT-CRF (ours)	KLUEBERT	Thunder-DeID	Kanana-1.5	Qwen-2.5
Model Specification					
# of Parameters	110M	110M	360M	2.1B	1.5B
Hidden Dimension	768	768	1,024	1,792	1,536
Hidden layers	12	12	24	32	28
Attention Head	12	12	16	24	14
Vocabulary Size	32,026	32,000	32,000	128,259	151,936
Fine-tuning					
Hardware	2x RTX 8000	2x RTX 8000	2x RTX 8000	2x RTX 8000	2x RTX 8000
Duration	24 hours	24 hours	24 hours	3 days	4 days
Learning Rate	3e-5	3e-5	3e-5	2e-5	3e-5
Batch Size	32	32	8	4	16
Seq Length	512	512	512	512	512
AdamW Weight Decay	0.02	0.01	0.01	0.01	0.01
AdamW Betas	β = (0.9, 0.999)	β = (0.9, 0.999)	β = (0.9, 0.999)	β = (0.9, 0.999)	β = (0.9, 0.999)
LoRA tuning					
LoRA r	-	-	-	16	8
LoRA Alpha	-	-	-	32	16
LoRA Dropout	-	-	-	0.05	0.05
LoRA Target Modules	-	-	-	q, v, k, o, gate, up, down_proj	q, v, k, o, gate, up, down_proj

Table 11: Detailed report of used models.

An example of publicly available anonymized court judgement

【판시사항】

갑 외국법인이 인터넷을 기반으로 하여 전 세계적으로 제공하는 검색, 이메일 등의 서비스에 가입한 을 등이 갑 법인을 상대로 정보통신망 이용촉진 및 정보보호 등에 관한 법률 제30조 제2항, 제4항에 따라 갑 법인이 을 등의 개인정보 및 서비스 이용 내역을 제3자에게 제공한 현황의 공개 등을 구한 사안에서, 을 등과 갑 법인 사이의 서비스 이용에 관한 법률관계에는 서비스 약관상 준거법 합의가 있더라도 정보통신망 이용촉진 및 정보보호 등에 관한 법률상 이용자의 권리보호에 관한 규정들이 적용되고, 갑 법인은 법령에 의하여 비공개 의무가 부과된 사항을 제외하고 을 등의 개인정보 및 서비스 이용 내역을 제3자에게 제공하였는지와 그 내용을 공개할 의무가 있다고 한 사례

【판결요지】

갑 외국법인이 인터넷을 기반으로 하여 전 세계적으로 제공하는 검색, 이메일 등의 서비스에 가입한 을 등이 갑 법인을 상대로 정보통신망 이용촉진 및 정보보호 등에 관한 법률(이하 '정보통신망법'이라 한다) 제30조 제2항, 제4항에 따라 갑 법인이 을 등의 개인정보 및 서비스 이용 내역을 제3자에게 제공한 현황의 공개 등을 구한 사안에서, 정보통신망법 제30조에서 정한 정보통신서비스 이용자의 권리는 국제사법 제27조 제1항의 '준거법 선택에 의하더라도 박탈할 수 없는 소비자에게 부여되는 보호에 관한 강행규정'에 해당하고, 당사자가 준거법으로 외국법을 적용하는 것에 대한 합의를 하였더라도 이용자가 정보통신망법에 근거한 권리를 행사할 수 없도록 하는 것은 우리나라 강행규정에 의하여 소비자에게 부여되는 보호를 박탈하는 것으로서 그 범위 내에서는 외국법을 준거법으로 하는 합의의 효력을 인정할 수 없으므로, 을 등과 갑 법인 사이의 서비스 이용에 관한 법률관계에는 서비스 약관상 준거법 합의가 있더라도 정보통신망법상 이용자의 권리보호에 관한 규정들이 적용되고, 다만 정보통신망법 제30조 제4항이 정보통신서비스 제공자에게 어떤 경우이든지 예외 없이 개인정보를 제3자에게 제공한 현황을 공개하도록 하는 의무를 부담시키고 있다고 보기 어려우므로, 갑 법인은 법령에 의하여 비공개 의무가 부과된 사항을 제외하고 을 등의 개인정보 및 서비스 이용 내역을 제3자에게 제공하였는지와 그 내용을 공개할 의무가 있다고 한 사례.

An example court judgment annotated according to our annotation scheme

【판시사항】

#@company#이 인터넷을 기반으로 하여 전 세계적으로 제공하는 검색, 이메일 등의 서비스에 가입한 #@name# 등이 #@company#을 상대로 정보통신망 이용촉진 및 정보보호 등에 관한 법률 제30조 제2항, 제4항에 따라 #@company#이 #@name# 등의 개인정보 및 서비스 이용 내역을 제3자에게 제공한 현황의 공개 등을 구한 사안에서, #@name# 등과 #@company# 사이의 서비스 이용에 관한 법률관계에는 서비스 약관상 준거법 합의가 있더라도 정보통신망 이용촉진 및 정보보호 등에 관한 법률상 이용자의 권리보호에 관한 규정들이 적용되고, #@company#은 법령에 의하여 비공개 의무가 부과된 사항을 제외하고 #@name# 등의 개인정보 및 서비스 이용 내역을 제3자에게 제공하였는지와 그 내용을 공개할 의무가 있다고 한 사례

【판결요지】

#@company#이 인터넷을 기반으로 하여 전 세계적으로 제공하는 검색, 이메일 등의 서비스에 가입한 #@name# 등이 #@company#을 상대로 정보통신망 이용촉진 및 정보보호 등에 관한 법률(이하 '정보통신망법'이라 한다) 제30조 제2항, 제4항에 따라 #@company#이 #@name# 등의 개인정보 및 서비스 이용 내역을 제3자에게 제공한 현황의 공개 등을 구한 사안에서, 정보통신망법 제30조에서 정한 정보통신서비스 이용자의 권리는 국제사법 제27조 제1항의 '준거법 선택에 의하더라도 박탈할 수 없는 소비자에게 부여되는 보호에 관한 강행규정'에 해당하고, 당사자가 준거법으로 외국법을 적용하는 것에 대한 합의를 하였더라도 이용자가 정보통신망법에 근거한 권리를 행사할 수 없도록 하는 것은 우리나라 강행규정에 의하여 소비자에게 부여되는 보호를 박탈하는 것으로서 그 범위 내에서는 외국법을 준거법으로 하는 합의의 효력을 인정할 수 없으므로, #@name# 등과 #@company# 사이의 서비스 이용에 관한 법률관계에는 서비스 약관상 준거법 합의가 있더라도 정보통신망법상 이용자의 권리보호에 관한 규정들이 적용되고, 다만 정보통신망법 제30조 제4항이 정보통신서비스 제공자에게 어떤 경우이든지 예외 없이 개인정보를 제3자에게 제공한 현황을 공개하도록 하는 의무를 부담시키고 있다고 보기 어려우므로, #@company#은 법령에 의하여 비공개 의무가 부과된 사항을 제외하고 #@name# 등의 개인정보 및 서비스 이용 내역을 제3자에게 제공하였는지와 그 내용을 공개할 의무가 있다고 한 사례.

Figure 4: A sample data of annonymized court judgement and court judgement annotated according to our annotation scheme.

An example of publicly available anonymized court judgement (translated in English)

【Holding】

In a case where Party B, who subscribed to services such as search and email provided globally by Foreign Corporation A via the internet, requested disclosure of the status of Party A's provision of Party B's personal information and service usage records to third parties pursuant to Article 30(2) and (4) of the Act on Promotion of Information and Communications Network Utilization and Information Protection, etc., Even if the legal relationship between Party B and the Corporation A regarding service use contains a governing law agreement in the service terms, the provisions protecting user rights under the Act on Promotion of Information and Communications Network Utilization and Information Protection apply. The Corporation A has an obligation to disclose whether it provided Party B's personal information and service usage details to third parties, and the content of such disclosure, except for matters subject to a non-disclosure obligation imposed by law.

【Abstract】

Party A, a foreign corporation, provides search, email, and other services globally via the internet. Party B and others, who subscribed to these services, requested Party A to disclose the status of providing their personal information and service usage details to third parties pursuant to Article 30, Paragraphs 2 and 4 of the Act on Promotion of Information and Communications Network Utilization and Information Protection, etc. (hereinafter referred to as the "Information and Communications Network Act"). In this case, the rights of users of information and communications services stipulated in Article 30 of the Act constitute a mandatory provision concerning the protection granted to consumers that cannot be deprived even by choice of law under Article 27(1) of the International Private Law Act. Therefore, even if the parties agreed to apply foreign law as the governing law, preventing users from exercising their rights under the Information and Communications Network Act would deprive consumers of the protection afforded by Korea's mandatory provisions. Consequently, within that scope, the validity of an agreement designating foreign law as the governing law cannot be recognized. Therefore, even if there is an agreement on the governing law in the service terms between Party B and Company A regarding the legal relationship concerning the use of the service, the provisions of the Information and Communications Network Act concerning the protection of the user's rights apply. However, it is difficult to interpret Article 30(4) of the Information and Communications Network Act as imposing an obligation on information and communications service providers to disclose the status of personal information provided to third parties in all cases without exception. Therefore, there is a case where Company A was found to have an obligation to disclose whether it provided the personal information and service usage details of Party B and others to third parties, and the content thereof, except for matters subject to a non-disclosure obligation imposed by law.

An example court judgment annotated according to our annotation scheme (translated in English)

【Holding】

In a case where #@name#, who subscribed to services such as search and email provided globally by #@company# via the internet, requested disclosure of the status of #@company#'s provision of #@name#'s personal information and service usage records to third parties pursuant to Article 30(2) and (4) of the Act on Promotion of Information and Communications Network Utilization and Information Protection, etc., Even if the legal relationship between Party #@name# and the #@company# regarding service use contains a governing law agreement in the service terms, the provisions protecting user rights under the Act on Promotion of Information and Communications Network Utilization and Information Protection apply. #@company# has an obligation to disclose whether it provided #@name#'s personal information and service usage details to third parties, and the content of such disclosure, except for matters subject to a non-disclosure obligation imposed by law.

【Abstract】

#@company#, a foreign corporation, provides search, email, and other services globally via the internet. #@name# and others, who subscribed to these services, requested #@company# to disclose the status of providing their personal information and service usage details to third parties pursuant to Article 30, Paragraphs 2 and 4 of the Act on Promotion of Information and Communications Network Utilization and Information Protection, etc. (hereinafter referred to as the "Information and Communications Network Act"). In this case, the rights of users of information and communications services stipulated in Article 30 of the Act constitute a mandatory provision concerning the protection granted to consumers that cannot be deprived even by choice of law under Article 27(1) of the International Private Law Act. Therefore, even if the parties agreed to apply foreign law as the governing law, preventing users from exercising their rights under the Information and Communications Network Act would deprive consumers of the protection afforded by Korea's mandatory provisions. Consequently, within that scope, the validity of an agreement designating foreign law as the governing law cannot be recognized. Therefore, even if there is an agreement on the governing law in the service terms between #@name# and #@company# regarding the legal relationship concerning the use of the service, the provisions of the Information and Communications Network Act concerning the protection of the user's rights apply. However, it is difficult to interpret Article 30(4) of the Information and Communications Network Act as imposing an obligation on information and communications service providers to disclose the status of personal information provided to third parties in all cases without exception. Therefore, there is a case where #@company# was found to have an obligation to disclose whether it provided the personal information and service usage details of #@name# and others to third parties, and the content thereof, except for matters subject to a non-disclosure obligation imposed by law.

Figure 5: A translated sample data of annonymized court judgement and court judgement annotated according to our annotation scheme.