Thunder-DeID: Accurate and Efficient De-identification Framework for Korean Court Judgments

Sungeun Hahm[*1] **Heejin Kim**[*1] **Gyuseong Lee**[*1] **Hyunji M. Park**[1] **Jaejin Lee**[1,2]
[1]Graduate School of Data Science, Seoul National University
[2]Dept. of Computer Science and Engineering, Seoul National University
{isungeuni, kheejin, ksnannaya, mhj233, jaejin}@snu.ac.kr

Abstract

To ensure a balance between open access to justice and personal data protection, the South Korean judiciary mandates the de-identification of court judgments before they can be publicly disclosed. However, the current de-identification process is inadequate for handling court judgments at scale while adhering to strict legal requirements. Additionally, the legal definitions and categorizations of personal identifiers are vague and not well-suited for technical solutions. To tackle these challenges, we propose a de-identification framework called Thunder-DeID, which aligns with relevant laws and practices. Specifically, we (i) construct and release the first Korean legal dataset containing annotated judgments along with corresponding lists of entity mentions, (ii) introduce a systematic categorization of Personally Identifiable Information (PII), and (iii) develop an end-to-end deep neural network (DNN)-based de-identification pipeline. Our experimental results demonstrate that our model achieves state-of-the-art performance in the de-identification of court judgments.

1 Introduction

Generally, court proceedings are open and accessible to the public. It is one of the key democratic principles enshrined in the constitutions of many countries, including South Korea[1]. South Korea is one of the countries with more stringent conditions that cover a broader range of personal identifiers to be anonymized in the court setting.

Before the publication of court decisions, the Korean National Court Administration uses both manual and automated de-identification methods throughout four stages of processing and review (Judicial Policy Research Institute of Korea, 2021). However, the current state of the de-identification procedure is not capable of handling court judgments at scale.

We want to address the following three problems of the current state of the de-identification procedure in South Korea. First, over-reliance on the manual method has been a major bottleneck, causing administrative strain and delaying publication of judgments. Public accessibility of judgments has been significantly low in South Korea, and the stagnant de-identification procedure is one of the reasons (National Court Administration of Korea, 2025). Second, the automatic de-identification tool's performance is surprisingly low. From 2019 to 2025, their overall accuracy merely spans 8 to 15% (National Assembly of Korea, 2019; National Court Administration of Korea, 2025). Finally, while existing law lays out the scope of de-identification, how personal identifiers are categorized and defined for administrative practice at the court is vague and especially unsuitable to be used for automated technical solutions.

To overcome the above problems, this paper proposes Thunder-DeID, a DNN- and NER-based framework, which improves the accuracy, efficiency, and consistency of de-identifying court judgments. Unlike a prompt-based approach using a large language model (LLM), which often alters the original sentence structure in the process of de-identification task (e.g., "총 3명 (a total of three people)" altered to "총 명수 1 (a total of one person)"), the token-level classification method of Thunder-DeID eliminates such risks of sentence and context distortion (see Appendix H). Moreover, due to privacy and information security concerns, the use of API-based LLM services, such as ChatGPT, is restricted in many of the key government institutions in Korea (National Intelligence Service, 2023). To create a trainable dataset from anonymized and unannotated court judgment data, we first manually label 6,700 civil, criminal, and administrative law cases that cover a broad spectrum of scenar-

[*]These authors contributed equally to this research.
[1]Constitution of South Korea, Art. 109

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 12728–12755
November 4-9, 2025

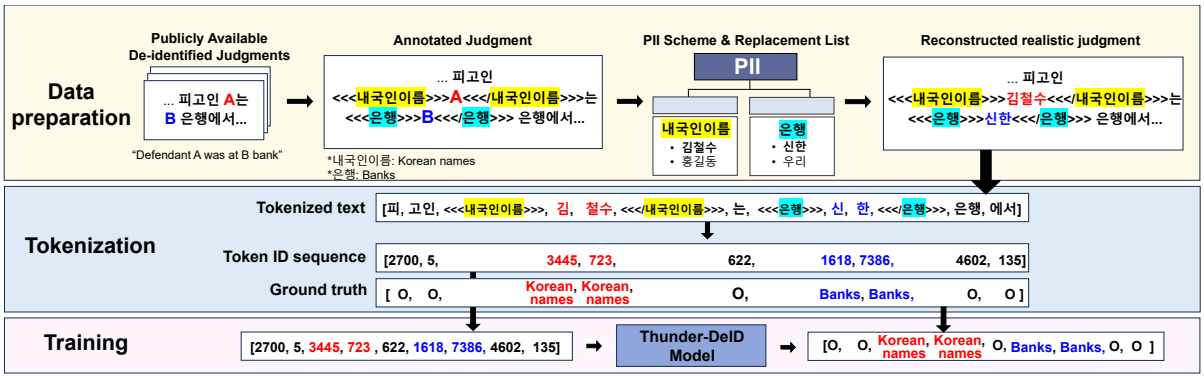

Figure 1: Overview of Thunder-DeID.

ios in civil, criminal, and administrative law. From these annotations, which identified 48,306 named entities, we establish a hierarchical categorization scheme for PII that aligns with relevant laws and practices and is suitable for model training. For each of the 729 labels in the PII scheme, we curate a corresponding list of entity mentions to generate model training data, as illustrated in Figure 1. Furthermore, we design a de-identification pipeline for the DNN-based language model, incorporating a specialized tokenizer that leverages the unique characteristics of the Korean language.

The approach used in this paper may offer valuable insights for other jurisdictions looking to efficiently anonymize large volumes of court decisions. The contributions of this paper are summarized as follows:

- We have created a two-part dataset that consists of 6,700 labeled judgments from three kinds of cases: civil, criminal, and administrative cases and a list of actual entity mentions to replace the labels. The labeled judgments are created from publicly available anonymized court judgments.
- We propose a three-tiered PII framework based on an inductive analysis of 48,306 named entities identified in our dataset.
- We propose a tokenizer that integrates a morphological analyzer, Mecab-ko, with Byte Pair Encoding (BPE) to leverage the unique features of the Korean language. Using this tokenizer, we also propose a method for generating training data from our labeled dataset and replacement list.
- We evaluate Thunder-DeID and it achieves the highest performance among existing de-identification models for court judgments.

2 Related Work

Among others, there are many de-identification studies in health information. In the USA, de-identification in the medical field is guided by the Health Insurance Portability and Accountability Act (HIPAA) (U.S. Department of Health and Human Services, 1996), which defines two main strategies for compliance: the Safe Harbor method and Expert Determination (Meystre et al., 2010; Emelyanov, 2021). The Safe Harbor method requires the removal of 18 identifiers called Personal Health Information (PHI). Alternatively, Expert Determination relies on a statistical or scientific method to ensure minimal re-identification risk. In Europe, the General Data Protection Regulation (GDPR) (European Parliament and Council, 2016) guides the de-identification of personal information in medical data. In this paper, we propose a three-tiered PII scheme for the de-identification of court judgment.

Medical de-identification. Research in medical de-identification has evolved through three major technical approaches. Early efforts primarily relied on rule-based systems (Uzuner et al., 2007). With the advancement of deep learning, learning-based de-identification approaches, such as BiLSTM-CRF (Liu et al., 2017) and BERT-based NER models (Berg et al., 2020; An et al., 2025), were introduced. Large language models (LLMs) have been recently explored for de-identification in zero-shot or few-shot settings (Liu et al., 2023; Altalla' et al., 2025). However, practical deployment is very limited because HIPAA regulations can be violated.

De-identification of court judgments. In recent years, there has been growing interest in automating the de-identification of court judgments based on NER. Many countries have launched government-led initiatives to adopt technical so-

Domain	Case type	Documents	**Entities**
Civil	Compensation for damage	901	9,223
	Security deposit disputes	696	5,187
	Payment of purchase price	557	4,983
	Eviction	846	6,816
	Subtotal	**3,000**	**26,209**
Criminal	Bodily injury	600	2,562
	Violence	600	2,583
	Sexual misconduct	600	2,732
	Property theft & deception	600	4,376
	Drunk driving	600	2,354
	Subtotal	**3,000**	**14,607**
Administrative	Administrative litigation	700	7,490
	Subtotal	**700**	**7,490**
Total		**6,700**	**48,306**

Table 1: Number of documents and entities for each case type in the dataset.

lutions to tackle problems with the labor-intensive de-identification procedure. The manual processing has been highlighted as delaying public disclosure and publication of judgments in Italy and Uruguay (Salierno et al., 2024; Garat and Wonsever, 2022). In India, the most populous country in the world, such a turn to automation is essential due to the overwhelming volume of court decisions (Kalamkar et al., 2022). In Switzerland, automation has been introduced to assist court officials and legal experts in the anonymization process (Niklaus et al., 2023). These NER-based methods report Precision, Recall, and F1-scores of 96.43%, 95.86%, 96.14% for Arabic (Moussaoui et al., 2023), 92.26%, 92.57%, 92.40% for German, French, and Italian texts (Switzerland), 89.92%, 90.50%, 91.90% for Spanish (Uruguay), 92.00%, 90.20%, 91.10% for Indian texts, and 85.00%, 92.46%, 88.60% for Italian texts (Italy).

Having a substantial post-processing approach is critical in de-identifying court judgments. For instance, over-anonymization or unprincipled anonymization may undermine the readability of rulings when publicly disclosed (Judicial Policy Research Institute of Korea, 2023). The majority of previous studies (Oksanen et al., 2022; Niklaus et al., 2023; Salierno et al., 2024) focus on how to detect personal identifiers in court judgments using NER, and less attention has been paid to discussing how the identified entities should be handled in the post-processing stage. Although the Uruguay study briefly addresses this issue, broader discussion and systematic approaches remain limited.

3 Methods

There are three challenges unique to constructing datasets for the de-identification of court judgments in South Korea. First, since de-identification of court judgments prior to publication is a legal obligation of judicial institutions [2], and only the fully anonymized judgments are available for external use, we need a method to generate datasets using anonymized and unannotated data.

Second, there are legal rules to define categories of personal identifiers to be anonymized[3]. However, they are not detailed enough to cover various attributes related to the persons involved in proceedings. They merely provide a direct identifier category and a broad quasi-identifier category that includes any other information that can identify the individual.

Finally, since the South Korean judiciary heavily relies on manual de-identification, which is time-consuming (National Assembly of Korea, 2019; National Court Administration of Korea, 2025), a large volume of court rulings that can immediately be used as a legal corpus for training is not available.

3.1 Data Collection

We initially compile 6,700 anonymized court decisions from a dataset provided by Korean Ministry of Government Legislation[4], AI-hub[5] and Hwang et al. (2022)[6]. After removing duplicates across different sources, the final dataset comprises 3,000 civil, 3,000 criminal, and 700 administrative cases.

Our dataset encompasses a wide range of civil, criminal, and administrative scenarios, as summarized in Table 1. By doing this, our dataset is bet-

[2]Korean Criminal Procedure Act, Art. 59-3; Korean Civil Procedure Act, Art. 163-2

[3]Korean Supreme Court Regulation No. 2809 and Judicial Rule No. 1778

[4]https://www.moleg.go.kr/

[5]https://www.aihub.or.kr/

[6]The dataset is released under the CC BY-NC 4.0 license.

ter suited for identifying various types of domain-specific personal identifiers in court judgements.

We focus on collecting judgments rendered by courts of first instance. A significant portion of these judgments in Korea is dedicated to examining and clarifying facts, which is different from the approach taken in common law countries. At this level, the courts prioritize fact-finding and resolving disputed facts based on the investigations and evidence presented in court. Consequently, the collected judgments contain numerous direct and quasi-identifiers related to multiple individuals involved in the proceedings.

3.2 Annotation Scheme

We need a systematic annotation scheme for the annonymized court judgments to ensure that data labeling is consistent, reliable, and useful for our DNN-based de-identification process. The labeling process following the annotation scheme should be consistent across annotators and reproducible. The scheme should also speed up training for new annotators and helps maintain quality over large datasets.

Without legal rules defining all relevant categories of personal identifiers, we develop an annotation scheme in four phases. First, human annotators identify placeholders (i.e., the anonymized sections in the judgment) in the provided text and label them using a set of entity categories we initially prepared based on an analysis of existing laws and practices. Second, while reviewing the labeling results for consistency among different annotators, we establish a new annotation scheme for PII with a *three-tiered hierarchical structure* that classifies a range of entity types. Third, annotators make adjustments and corrections according to the annotation scheme. Finally, we resolve any issues where annotators may disagree or have doubts.

3.3 Placeholder Detection and Labeling

We have seventeen annotators who are fluent in Korean and possess a good understanding of NLP. They have completed an initial training session that provided guidelines on two main aspects: the key features of the task, which include a multi-stage process we designed for this project, and the rules regarding the scope and method of anonymization as applied in court practice.

Korean Judicial Rule No. 1778 establishes principles to guide court officials in using various de-identification methods. Depending on the type of identifiers involved, individuals can be represented with English letters (e.g., A and B) or combinations of letters (e.g., ABB, AAB). The complete removal of certain direct identifiers, such as resident registration numbers, is mandatory. For example, the text "... 피고인 홍길동 (561231-1234567) ... " ("... defendant Hong Gildong (561231-1234567) ...") would be anonymized to "... 피고인 A (주민등록번호 1) ..." ("... defendant A (resident registration number 1) ..."), where 561231-1234567 is a specific resident registration number. In this case, "피고인 A (주민등록번호 1)" represents the anonymized information obtained from the judgment within the collected corpus and is not labeled or annotated. The resident registration number is designated as 1 to differentiate between multiple individuals present in the judgment.

Annotators manually identify the placeholders A and 1, labeling them to indicate the specific types of entities they represent as follows: "... ≪내국인이름≫A≪/내국인이름≫(≪주민등록번호≫B≪/주민등록번호≫) ...," where "내국인이름" refers to Korean names, and "주민등록번호" refers to a resident registration number. ≪내국인이름≫ and ≪/내국인이름≫ are markers and they point the beginning and end of the entity mention, respectively. "내국인이름" is a label to represent the category of the entity mention in our PII scheme (see Appendix B).

In an adjudication setting, locational information, such as the residential addresses of the parties involved in a case and the address of the crime scene, is essential for confirming the court's jurisdiction. It is standard practice to provide the exact address; however, under Korean Judicial Rule No. 1778, specific lower-level details of the address, like districts and streets, must be masked. At first glance, the address of a location or the name of a place may not seem like identifying information. However, their direct association with specific criminal activities can help identify the individuals involved in the case. Therefore, in accordance with existing laws and practices, lower-level address components and the names of all incident-related places must be de-identified. Similarly, contextual attributes such as the date of an event may also be considered quasi-identifiers and should be masked. For more examples of masking and labeling, please see Appendix C.

Annotators identified and labeled a total of 48,306 named entities across 6,700 court judgments. Table 1 shows the number of documents

and identified entities for the crime categories in the collected judgments.

3.4 PII Categories

As discussed earlier, existing law broadly defines the scope of de-identification. Aside from clear direct identifiers, quasi-identifiers often require more than just a textual assessment of the relevant attributes that can make an individual identifiable. The scope can be as extensive as "any other information identifying the persons involved in the case and third parties"[7]. Since it is nearly impossible to list all privacy-sensitive identifiers in writing, court officials are instructed to use discretion and analyze the specific context and its connection to the individuals involved in the case.

During the initial review of labeling, we found that many of the identified entities, specifically, the information anonymized in the collected judgments, do not consistently fit within the predefined categories of identifiers. There are challenging cases where the same type of entity may be evaluated differently across multiple judgments.

For example, as a general rule, names of government institutions and public authorities (such as the Seoul Police Agency and the Seoul Correctional Institution) are not subject to de-identification. However, if these organizations are associated with the location where a crime was committed, exceptions may apply. This contextual interpretation of the case can lead to varying outcomes.

For another example, consider the following de-identified judgment text:"... 피고인 F와 피해자 G는 H 교도소 I 팀 소속의 교정공무원으로 ..." ("... defendant F and victim G were prison officers at team I of H correctional institution ..."). In actual de-identification practice, the name of the correctional institution ("교도소") is anonymized because it identifies the workplace where both the defendant and victim were colleagues. If this information is not anonymized in public disclosures, it could increase the chances of identifying the two individuals due to its direct connection to the circumstances surrounding the crime committed.

A different challenge in annotation arises when there are many individuals involved, and the specific roles each person plays in the case are not clearly defined during the anonymization process. This is particularly evident in cases of fraud, where a large group of victims is often targeted by illegal organizations, each member responsible for different aspects of the criminal activities. Additionally, in trials involving accomplices to a crime, it is crucial to anonymize identifiable information about various third parties, such as witnesses, appraisers, and forensic experts, to mitigate the risk of retaliation.

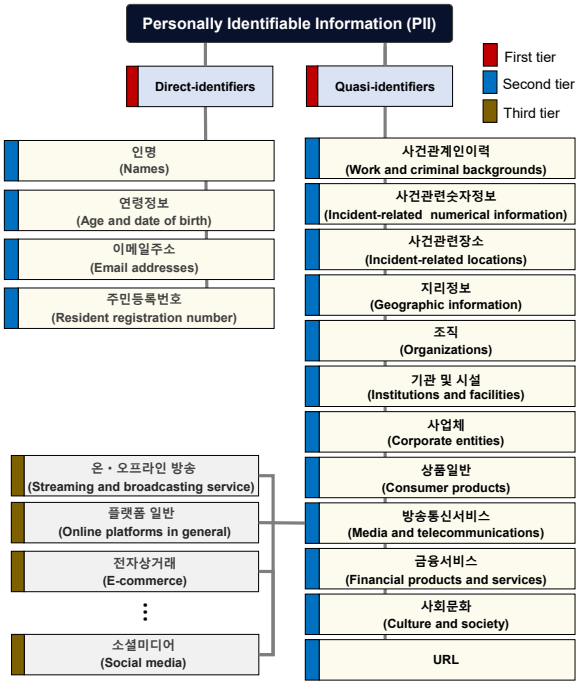

Figure 2: The three-tiered categorization scheme for PII in the domain of law and adjudication.

While the annotators made adjustments and corrections in accordance with the annotation scheme, we resolved any issues where the annotators disagreed or had uncertainties.

After reviewing all the named entities in the judgments, we developed our own PII annotation scheme that classifies various entity types into two main categories: direct identifiers and quasi-identifiers. This scheme includes 16 subcategories and 80 granular categories. Figure 2 illustrates the hierarchy of the categories. Each of the third-tier categories is associated with labels for annotation. Using this scheme, we annotated the identified named entities with a total of 729 labels. To the best of our knowledge, this is the first PII annotation scheme specifically designed for the de-identification of court judgments in Korea. Further details on the annotation scheme and its categories are provided in Appendix D.

[7]Korean Judicial Rule No. 1778, Art. 4

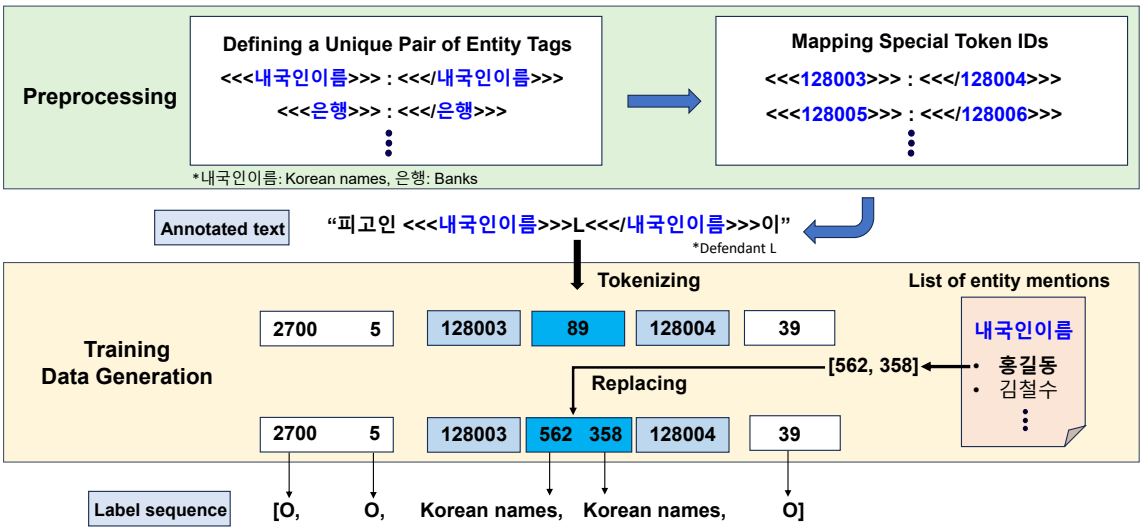

Figure 3: Tokenization and training data generation.

3.5 Replacement Lists

To improve the size and diversity of our training data, we create an extensive list of entity mentions using two different methods: manual curation and rule-based generation.

Manual curation. We selectively choose reliable and verified information (entity mentions) sourced from the Korean government's licensing databases[8] and public data portals[9]. We generate entity mentions for the majority of labels—691 out of 729. Our goal is to compile an average of at least 100 items for each label.

We also conduct searches on domain-specific websites to collect entity names related to specialized locations. For example, we gather lists of exhibition halls and conventions from the Coex Center, obtain names of ships and vessels from the Korea Seafarer's Welfare & Employment Center, and collect names of junk dealers and recycling companies from the Korea Waste Recycling Institute. Additionally, we perform general web searches to supplement these results, ensuring a broad and diverse set of entity mentions that accurately reflect real-world usage.

Rule-based generation. The second strategy uses rule-based generation to create entity mentions involving personal identifiers in standardized and structured formats. Simple rules are employed to generate entities such as Korean names, addresses in Korea, and numerical identifiers, which include resident registration numbers, phone numbers, and bank account numbers.

[8] https://www.localdata.go.kr/main.do
[9] https://www.data.go.kr/

3.6 Training Data Generation

When we train our language model, we generate training data from the annotated dataset. In this process, we replace the labels in the dataset with actual entity mentions in the replacement list. Each labeled court judgment is augmented multiple times (N times) through entity mention replacements to maximize the amount of training data.

For model training, documents are converted into tokenized input sequences (referred to as X) and corresponding label sequences (referred to as Y). Our tokenizer has been extended to include 1,458 special tokens that represent 729 different entities (labels). This extension is prioritized to ensure that proper nouns do not merge with other particles. Each judgment document is transformed into a token sequence, where subsequences marked with the start token ⋘, placeholder tokens, and the end token ⋙ (e.g., "⋘name⋙A⋘/name⋙" after tokenization) are replaced with actual entity mention token sequences ("홍길동" after tokenization). Tokens within these subsequences are assigned the relevant label in Y (e.g., "name") for de-identification, while tokens in other subsequences receive an "O" (outside) label in Y to indicate that they do not require de-identification.

3.7 Tokenization

We develop a custom tokenizer trained on a subset of one million sentences sampled from our corpus to effectively segment sensitive entities, such as names and organizations. Our tokenizer integrates a dictionary-based morphological analyzer, Mecab-ko[10], with Byte Pair Encoding (BPE) (Sennrich

[10] https://github.com/hephaex/mecab-ko

et al., 2016).

We choose Mecab-ko due to its ability to handle the Korean language's agglutinative morphology. It segments text into morphemes using a predefined dictionary, accurately distinguishing between nouns, particles, affixes, and adjectives. Studies have demonstrated Mecab-ko's effectiveness for recognizing domain-specific terms and proper nouns in Korean NLP tasks (Park et al., 2020; Cho et al., 2021; Jeon et al., 2023).

Unlike English, where proper nouns like "홍길동" remain unsegmented, Korean attaches nominative particles, such as "-이" and "-을," to nouns (e.g., "홍길동이"). Mecab-ko's dictionary-based segmentation separates "홍길동이" into "홍길동" and "-이", ensuring that only the target entity ("홍길동") is de-identified while the particles remain intact. This approach helps the de-identified text flow smoothly and naturally. In addition, such precision is essential, given that the original (i.e., unanonymized and unannotated) court decisions lack clear boundaries for all entities.

While using a morphological analyzer like Mecab-ko is powerful, its fixed dictionary may not be able to capture rare legal terms or proper nouns, leading to out-of-vocabulary (OOV) issues. To overcome this limitation, we chose BPE, which builds a vocabulary through frequent character pair merges and represents unseen terms as subword units.

Tokenization algorithm. The tokenizer recognizes special tokens and assigns unique token IDs to the beginning and end marker tokens of an entity mention. For instance, consider Figure 3. We assign 128003 to ⋘내국인이름⋙ and 128004 to ⋘/내국인이름⋙. Given an input text from the annotated dataset, such as "피고인 ⋘내국인이름⋙L⋘/내국인이름⋙이...", the text is tokenized into a sequence: [2700, 5, 128003, 82, 128004, 39], where L (token ID 82) serves as a placeholder for a labeled entity. Here, 내국인이름 refers to Korean names, and "피고인 L이" refers to "Defendant L".

Next, the token sequence is scanned to identify start marker tokens (e.g., 128003) and their corresponding end marker tokens (e.g., 128004), thus detecting the range of tokens between them. This range includes the placeholder (e.g., [128003, 82, 128004]). The placeholder within this range is then replaced with one of the entity mentions selected from the pre-defined replacement list. For example, in the sequence [2700, 5, 128003, 82, 128004, 39], the segment [128003, 82, 128004] is replaced by a token sequence [562, 358], which represents a name "홍길동" in the replacement list. This results in the updated sequence: [2700, 5, 562, 358, 39].

Subsequently, a corresponding label sequence is generated based on the indices of the replaced tokens, ensuring that the position and type of the labeled entity are retained (i.e., marking "홍길동" as a Korean names). For instance, the token sequence [2700, 5, 562, 358, 39] generates the label sequence [O, O, Korean names, Korean names, O], where "O" represents "Outside". This label sequence serves as the ground truth for supervised learning. Finally, the modified token sequence and its associated label sequence form a training data instance in the dataset (Figure 3).

3.8 Data Augmentation

Due to the limited availability of publicly accessible court judgments, there will inevitably be instances where new entity types arise that the existing PII labels cannot represent. To address this limitation, we prepare a set of additional labels using LLM-assisted augmentation.

We begin by selecting specific granular categories that have significantly fewer labels compared to others. Next, we employ a large language model (LLM), such as ChatGPT (OpenAI, 2022), to generate additional labels and create corresponding lists of entity mentions. For instance, "socio-cultural event" is one of the granular categories under "Culture and Society" in the proposed PII scheme (Figure 2). If, during the annotation process, we identify only a few labels within the "socio-cultural event" granular category, we can instruct the LLM to generate more labels for this category. Subsequently, we manually create several entity mentions for each additional label generated by the LLM.

4 Experiments

This section evaluates Thunder-DeID and the experimental methodology.

4.1 Experimental Setup

Training datasets. Besides our annotation dataset, we collect a bilingual corpus of approximately 76.7GB, comprising Korean and English texts from publicly available Web sources. This corpus is used for tokenizer training and pre-training

Model	#Params	Single Replacement (Binary Token-Level)			Per-Epoch Replacement (Binary Token-Level)		
		Precision	Recall	F1	Precision	Recall	Micro F1
Polyglot-ko	1.3B	0.9774	0.9570	0.9669	0.9710	0.9695	0.9701
Exaone	2.4B	0.9774	0.9542	0.9656	0.9688	0.9666	0.9677
Thunder-DeID-360M	360M	0.9767	0.9264	0.9509	0.9628	0.9679	0.9654
Thunder-DeID-800M	800M	0.9786	0.9767	0.9776	0.9757	0.9826	0.9791
Thunder-DeID-1.5B	1.5B	0.9855	0.9683	0.9769	0.9755	0.9862	**0.9808**

(a) **Binary token-level** (Precision, Recall, and F1)

Model	#Params	Single Replacement (Token-Level)			Per-Epoch Replacement (Token-Level)		
		Precision	Recall	F1	Precision	Recall	Micro F1
Polyglot-ko	1.3B	0.8816	0.8631	0.8723	0.8772	0.8758	0.8765
Exaone	2.4B	0.8785	0.8576	0.8679	0.8762	0.8742	0.8752
Thunder-DeID-360M	360M	0.8895	0.8438	0.8660	0.8848	0.8895	0.8871
Thunder-DeID-800M	800M	0.9099	0.9082	0.9090	0.9073	0.9137	**0.9105**
Thunder-DeID-1.5B	1.5B	0.9091	0.8933	0.9011	0.9021	0.9120	0.9071

(b) **Token-level** (Precision, Recall, and Micro F1)

Table 2: Performance comparison under different data generation settings. Each sub-table reports **Precision, Recall, and F1** on the test set for the indicated evaluation granularity (Binary token-level vs Token-level). Values are averaged over three random seeds (1200, 1203, 1205). The best performance results are highlighted in bold.

for our language model. We also generate a dataset for NER-based de-identification using the method described in subsection 3.6. The dataset is divided into 80% training (2,400, 2,401, and 560 documents), 10% validation (300, 298, and 70 documents), and 10% test (300, 301, and 70 documents), for civil, criminal, and administrative cases, respectively.

Language models used. We train DeBERTa-v3-based models (He et al., 2023), Thunder-DeID, with 370M, 800M, and 1.5B parameters for the de-identification of Korean court judgments through token classification. These models are compared against Korean-specialized language model baselines, namely Polyglot-Ko (Ko et al., 2023) and EXAONE-3.5 (An et al., 2024), to assess their performance on our proposed dataset. For detailed information on the architectures and training configurations, please refer to Table E.1 in Appendix E.

Pre-training the models. Thunder-DeID models are pre-trained from scratch using subsets of our bilingual corpus, which includes both English and Korean, containing 60 billion tokens for the 1.5 billion parameter model, 30 billion tokens for the 800 million parameter model, and 14 billion tokens for the 370 million parameter model. Training begins with a sequence length of 512 tokens, which is later extended to 2048 tokens to accommodate longer contexts. Unlike the original DeBERTa-v3, which uses post-LayerNorm, we adopt pre-LayerNorm (Xiong et al., 2020) because post-LayerNorm failed to converge for larger models, whereas pre-LayerNorm converged reliably under the same settings. For more details, please see Table E.1 in Appendix E.

Fine-tuning the models. Thunder-DeID models and the baseline models were fine-tuned on our dataset, which consists of 5,361 training documents (2,400, 2,401 and 560), 668 validation documents (300, 298, 70) and 671 test documents (300, 301, 70) for civil, criminal, and administrative cases, respectively. We employ both Per-Epoch and Single Entity Replacement methods to assess the effects of data variation. The training use a sequence length of 2,048 tokens over the course of 30 epochs. For detailed training information, please refer to Table E.1 in Appendix E, and for the results, see Table 2.

Evaluation metrics. We use three metrics — precision, recall, and F1-score — to assess the performance of our model on the de-identification task. Each metric is evaluated under two settings: binary token-level (Dernoncourt et al., 2016; Yue and Zhou, 2020; Salierno et al., 2024; Kim et al., 2024) and token-level (Dernoncourt et al., 2016; Yue and Zhou, 2020; Kim et al., 2024). The binary token-level setting measures the model's ability to cor-

rectly classify tokens that require de-identification and those that do not, without considering the type of entity. For the details of the two settings and metric definitions, please see Appendix F.

4.2 Experimental result

Table 2 shows the performance of our models compared to two Korean-specialized Decoder models, Polyglot-ko (1.3B) and Exaone (2.4B), under two data generation settings: Single Replacement and Per-Epoch Entity Replacement. Thunder-DeID models consistently outperform the baselines in both binary token-level and token-level micro F1 scores. Our largest model Thunder-DeID-1.5B achieves a binary token-level F1 of 0.9808 and 800M model achieves a token-level F1 of 0.9105 under the Per-Epoch Entity Replacement setting, establishing a state-of-the-art (SOTA) benchmark for NER-based de-identification of Korean court judgments. Notably, even our smallest model Thunder-DeID-370M (0.8871) outperforms both Polyglot-ko (0.8765) and Exaone (0.8752) in the token-level micro F1 metric. For a detailed breakdown of performance by case type, please refer to Appendix I.

The high binary token-level F1 score for Thunder-DeID under Per-Epoch Entity Replacement demonstrates that the model is proficient in identifying which tokens need to be de-identified. Additionally, the high token-level micro F1 score indicates that Thunder-DeID effectively classifies the entity types of these de-identifiable tokens. Given that the model is required to classify as many as 729 distinct labels, achieving a token-level F1 score exceeding 0.91 is a strong indicator of its robust multi-class classification performance.

The Per-Epoch Entity Replacement technique significantly outperforms Single Replacement in all models, including Polyglot-ko and Exaone. This consistent improvement highlights the quality of our dataset, its annotation scheme, and the corresponding list of entity mentions for realistic value generation. Frequent entity replacements enhance data diversity while maintaining high-quality augmentation and effective generalization.

The 800M model demonstrates a slightly higher token-level micro F1 score under the per-epoch setting compared to the 1.5B model. In our data-limited scenario, the 800M model may be better suited to the dataset size, allowing it to generalize slightly better. In contrast, the 1.5B model may overfit to rare labels. However, the difference between the two models is minimal and could diminish with additional data for rare labels or the application of stronger regularization.

Thunder-DeID demonstrates weaknesses in identifying low-frequency labels that seldom appear in the training corpus. For example, it frequently misclassifies "뷔페 (buffet restaurant)"—which should fall under "외식업 (eating and drinking places)"—as "기계설비회사 (machinery and equipment company)" within the "제조업 (manufacturing)" category. As our annotators reviewed the fully anonymized court judgments, we noted some exceptional cases where it was challenging to accurately determine the exact type of entity, despite careful contextual analysis. These instances also resulted in misclassifications, such as labeling "불특정제품명 (unspecified product name)" under "상품 일반 (general products)" as "불특정회사명 (unspecified company)" under "기업 일반 (companies and businesses in general)."

Thunder-DeID significantly outperforms the rule-based system currently used by the Korean National Court Administration, which reportedly achieves an overall accuracy of 8 to 15% (National Assembly of Korea, 2019; National Court Administration of Korea, 2025). These results position Thunder-DeID as a new and effective framework for Named Entity Recognition (NER)-based de-identification of court judgments.

5 Conclusion

In this paper, we propose a DNN-based solution, referred to as Thunder-DeID, for NER aimed at improving the efficiency and consistency of de-identifying court judgments. We address the complex challenges currently faced in the de-identification process within the Korean judiciary. Our work includes the development of the first Korean legal dataset, which contains 6,700 judgments from civil, criminal, and administrative cases, encompassing a total of 48,306 labeled named entities. We also introduce a three-tiered annotation scheme for PII, which systematically categorizes a wide variety of personal identifiers. Furthermore, we provide a comprehensive list of entity mentions that can be used to replace the 729 token-level labels found in the training dataset. In addition, we outline a tokenization method for the training data generated from these replacements. Our experimental results show that Thunder-DeID achieves state-of-the-art performance in the de-identification of court judgments.

Limitations

Our study has some limitations. First, original (unanonymized) court judgments are not accessible due to legal restrictions. As mentioned earlier, we only have access to fully anonymized judgments that have been processed and reviewed by court officials before being made public. This limitation prevents us from evaluating our model's performance in real-world settings. To address this issue and make our model more applicable to actual de-identification practices within the Korean judiciary, we plan to develop a more strategic method of data augmentation for future research. This includes creating synthetic data that closely resembles court judgments. By pursuing this direction, we aim to increase the size and diversity of our training data, allowing for more robust testing of our model.

Second, our model was specifically trained using judgments from the field of civil, criminal, and administrative law and procedure. De-identification in the legal domain is highly context-sensitive, which means the model's performance may decrease when applied to court decisions involving different types of legal disputes. However, we anticipate that our model will still perform reasonably well, as there are shared characteristics regarding direct identifiers across various types of court judgments. Additionally, our dataset encompasses a wide range of entity types. Thus, our system has important implications even for court judgments in entirely different areas of law. Further research is necessary to evaluate the model's performance in these other areas and to explore how the proposed method can be adapted and enhanced for effective de-identification tasks across diverse legal contexts.

Ethics Statement

All court judgments used in this study were obtained from publicly available anonymized datasets, including those released by the Korean Ministry of Government, AI-hub[11] and published by Hwang et al. (2022), none of which contain any PII. To support data reconstruction and model training, replacement lists were compiled exclusively from open-access sources, including government licensing databases[12], public data portals[13], and official institutional websites[14]. No private or sensitive information was used at any stage of this research.

Although the dataset is fully anonymized and all sources are publicly available, we ensured that our data processing procedures—including the creation of replacement lists—adhered to the principles of the Korean Personal Information Protection Act (PIPA).

[11] https://www.aihub.or.kr/

[12] https://www.localdata.go.kr/main.do

[13] https://www.data.go.kr/

[14] https://data.seoul.go.kr/

Acknowledgements

We thank the anonymous reviewers and the meta-reviewer for their valuable feedback on this paper. We also sincerely thank Gyeongje Cho, Hyeonggeun Jeon, Sungmok Jung, Dayeon Kang, Jia Kang, Minsu Kim, Sangho Kim, Jongmin Kim, Dongyoung Lee, Joonhak Lee, Changjin Lee, Jongyeon Park, Yoonhee Park, Seho Pyo, Jiheon Seok, Yeonkyoung So, and Youngjun Son for their dedicated work as annotators.

This work was supported in part by the National Research Foundation of Korea (NRF) grant (No. RS-2023-00222663, Center for Optimizing Hyperscale AI Models and Platforms), by the Institute for Information and Communications Technology Promotion (IITP) grant (No. 2018-0-00581, CUDA Programming Environment for FPGA Clusters), by the BK21 Plus programs for BK21 FOUR Intelligence Computing (Dept. of Computer Science and Engineering, SNU, No. 4199990214639), all funded by the Ministry of Science and ICT (MSIT) of Korea. This work was also supported in part by the Artificial intelligence industrial convergence cluster development project funded by the Ministry of Science and ICT (MSIT, Korea) and Gwangju Metropolitan City. ICT at Seoul National University provided research facilities for this study.

References

Bayan Altalla', Sameera Abdalla, Ahmad Altamimi, Layla Bitar, Amal Al Omari, Ramiz Kardan, and Iyad Sultan. 2025. Evaluating gpt models for clinical note de-identification. *Scientific Reports*, 15(1):3852.

Jiyong An, Jiyun Kim, Leonard Sunwoo, Hyunyoung Baek, Sooyoung Yoo, and Seunggeun Lee. 2025. De-identification of clinical notes with pseudo-labeling using regular expression rules and pre-trained bert. *BMC Medical Informatics and Decision Making*, 25(1):82.

Soyoung An, Kyunghoon Bae, Eunbi Choi, Kibong Choi, Stanley Jungkyu Choi, Seokhee Hong, Junwon Hwang, Hyojin Jeon, Gerrard Jeongwon Jo, Hyunjik Jo, et al. 2024. Exaone 3.5: Series of large language

models for real-world use cases. *arXiv e-prints*, pages arXiv–2412.

Hanna Berg, Aron Henriksson, and Hercules Dalianis. 2020. The impact of de-identification on downstream named entity recognition in clinical text. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, pages 1–11.

Danbi Cho, Hyunyoung Lee, and Seungshik Kang. 2021. An empirical study of korean sentence representation with various tokenizations. *Electronics*, 10(7).

Franck Dernoncourt, Ji Young Lee, Ozlem Uzuner, and Peter Szolovits. 2016. De-identification of patient notes with recurrent neural networks.

Yaroslav Emelyanov. 2021. Towards task-agnostic privacy-and utility-preserving models. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 394–401.

European Parliament and Council. 2016. General Data Protection Regulation (EU) 2016/679, Recital 35. https://eur-lex.europa.eu/eli/reg/2016/679/oj. Official Journal of the European Union, L119, pp. 1–88.

Diego Garat and Dina Wonsever. 2022. Automatic curation of court documents: Anonymizing personal data. *Information*, 13(1):27.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing.

Wonseok Hwang, Dongjun Lee, Kyoungyeon Cho, Hanuhl Lee, and Minjoon Seo. 2022. A multi-task benchmark for korean legal language understanding and judgement prediction. *Advances in Neural Information Processing Systems*, 35:32537–32551.

Taehee Jeon, Bongseok Yang, Changhwan Kim, and Yoonseob Lim. 2023. Improving korean nlp tasks with linguistically informed subword tokenization and sub-character decomposition. *arXiv preprint arXiv:2311.03928*.

Judicial Policy Research Institute of Korea. 2021. A study on personal data protection in criminal trial procedures. Technical report, Judicial Policy Research Institute.

Judicial Policy Research Institute of Korea. 2023. Contemporary meanings and limitations of the principle of open justice.

Prathamesh Kalamkar, Astha Agarwal, Aman Tiwari, Smita Gupta, Saurabh Karn, and Vivek Raghavan. 2022. Named entity recognition in indian court judgments. *arXiv preprint arXiv:2211.03442*.

Woojin Kim, Sungeun Hahm, and Jaejin Lee. 2024. Generalizing clinical de-identification models by privacy-safe data augmentation using gpt-4. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21204–21218.

Hyunwoong Ko, Kichang Yang, Minho Ryu, Taekyoon Choi, Seungmu Yang, Jiwung Hyun, Sungho Park, and Kyubyong Park. 2023. A technical report for polyglot-ko: Open-source large-scale korean language models.

Zengjian Liu, Buzhou Tang, Xiaolong Wang, and Qingcai Chen. 2017. De-identification of clinical notes via recurrent neural network and conditional random field. *Journal of biomedical informatics*, 75:S34–S42.

Zhengliang Liu, Xiaowei Yu, Lu Zhang, Zihao Wu, Chao Cao, Haixing Dai, Lin Zhao, Wei Liu, Dinggang Shen, Quanzheng Li, et al. 2023. Deid-gpt: Zero-shot medical text de-identification by gpt-4. *arXiv preprint arXiv:2303.11032*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

Stephane M Meystre, F Jeffrey Friedlin, Brett R South, Shuying Shen, and Matthew H Samore. 2010. Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC medical research methodology*, 10:1–16.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al. 2017. Mixed precision training. *arXiv preprint arXiv:1710.03740*.

Taoufiq El Moussaoui, Loqman Chakir, and Jaouad Boumhidi. 2023. Preserving privacy in arabic judgments: Ai-powered anonymization for enhanced legal data privacy. *IEEE Access*, 11:117851–117864.

National Assembly of Korea. 2019. Debate for promoting open publication of court judgments. Online. PDF file checked, p.39.

National Court Administration of Korea. 2025. National court administration of korea, final report: Information strategy plan (isp) for development of ai models to support trials (2025). Technical report, National Court Administration of Korea.

National Intelligence Service. 2023. Security guidelines for using generative ai such as chatgpt. https://www.ncsc.go.kr:4018/main/cop/bbs/selectBoardArticle.do?bbsId=InstructionGuide_main&nttId=54340&pageIndex=1.

Joel Niklaus, Robin Mamié, Matthias Stürmer, Daniel Brunner, and Marcel Gygli. 2023. Automatic anonymization of swiss federal supreme court rulings. *arXiv preprint arXiv:2310.04632*.

Arttu Oksanen, Eero Hyvönen, Minna Tamper, Jouni Tuominen, Henna Ylimaa, Katja Löytynoja, Matti Kokkonen, and Aki Hietanen. 2022. An anonymization tool for open data publication of legal documents. In *International Workshop on Artificial Intelligence Technologies for Legal Documents/International Workshop on Knowledge Graph Summarization*, pages 12–21. CEUR-WS. org.

OpenAI. 2022. Introducing chatgpt. https://openai.com/blog/chatgpt. Accessed: 2025-01-02.

Kyubyong Park, Joohong Lee, Seongbo Jang, and Dawoon Jung. 2020. An empirical study of tokenization strategies for various Korean NLP tasks. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 133–142, Suzhou, China. Association for Computational Linguistics.

Giulio Salierno, Rosamaria Bertè, Luca Attias, Carla Morrone, Dario Pettazzoni, and Daniela Battisti. 2024. Giusberto: A legal language model for personal data de-identification in italian court of auditors decisions. *arXiv preprint arXiv:2406.15032*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

U.S. Department of Health and Human Services. 1996. HIPAA for Professionals: Laws & Regulations. https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html. Accessed: 2025-05-19.

Özlem Uzuner, Yuan Luo, and Peter Szolovits. 2007. Evaluating the state-of-the-art in automatic de-identification. *Journal of the American Medical Informatics Association*, 14(5):550–563.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu. 2020. On layer normalization in the transformer architecture.

Xiang Yue and Shuang Zhou. 2020. PHICON: Improving generalization of clinical text de-identification models via data augmentation. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 209–214, Online. Association for Computational Linguistics.

Appendix

A Issues in Prompt-based De-identification

We identify the following five categories of problems frequently appearing in the GPT-assisted de-identification dicussed in Section 1. These cases represent the ways in which prompting-based anonymization can lead to compromise textual integrity of public records and undermine legal precision required for settling disputes effectively.

- First, rewriting and paraphrasing frequently occurred. For example, the verb "입금하였다 (deposited)" was changed to "송금하였다 (wire transferred)." While both can describe sending money to someone, the forms and implications of these behaviors are differently conceived in legal and financial contexts.
- Second, we also found cases of partial omission when GPT removed, for instance, the phrase "제때 (on time)" from the original text. The original phrase "그 대금을 제때 변제하여" ("by repaying the amount on time") was shortened to "대금을 변제하여" ("by repaying the amount") in GPT-4's output. The omission of "제때" ("on time") removes an important indication of timely payment, which is often critical in determining whether the legal obligation was properly met.
- Third, (unsolicited) summarization of the original text resulted in the loss of detailed facts and strategies concerning the crimes committed. Unlike the original text, it merely provides a brief summary of the factual backgrounds of the case. For instance, after going through GPT-assisted de-identification, three sentences containing important details about defendant's intention and plan to defraud victim and the amount of damage caused were vaguely summarized and reduced to a single sentence, "피고인은 이를 개인 용도로 사용하였다 (The defendant used it for personal purposes)".
- Fourth, in the cases where multiple individuals and institutions are involved in the litigation, we often identified entity collapse: a number of different entities were anonymized with the same letter (e.g., 광주은행 (Gwangju Bank), 우정사업본부 (Korea Post), 부산은행 (Busan Bank) → A, A, A).
- Lastly, distortion of facts occurred. For example, specific numbers in the judgment were altered during de-identification "총 3명 (a total of three people)" was altered to "총 명수1 (a total of one person)".

Moreover, due to privacy and information security concerns, the use of API-based LLM services such as ChatGPT is restricted in Korean government institutions. Domestic regulations (issued by the National Intelligence Service and the Ministry of the Interior and Safety) require public officials across government departments to refrain from putting in any sensitive internal data and personal information while using such services.

B Data Samples

Since it is a legal obligation of the courts to anonymize judgments prior to public disclosures, there is no way to access unannoymized judgments which could have served as ground truth for our research. After collecting fully anonymized judgments, we manually annotate the whole corpus based on the three-tiered categorization scheme classifying a range of personal identifiers. (See Section 3.3)

To give our readers the gist of the collection and annotation process, Appendix C presents one of the examples of the court judgment initially compiled for data construction as in Figure B.1. Next to this anonymized court judgment, the annotated version of that same judgment appears.

C Masking and Labeling Examples

Appendix B illustrates the example discussed in Section 3.3 in more detail.

Example with a functional descriptor. Locational information in the sentence "… 나리 식당에서 근무하는 피고인 홍길동은 …" can be de-identified as "… A 식당에 근무하는 피고인 B는 …". According to Korean Judicial Rule No. 1778, "나리 식당" qualifies as identifying information due to its contextual specificity. While the generic term "식당" (diner) remains intact as a non-identifying functional descriptor, the unique component "나리" is replaced with the placeholder "A". Similarly, the name "홍길동" is replaced with "B". The resulting sentence is labeled as:

"… ≪식당≫A≪/식당≫ 식당에 근무하는 피고인 ≪내국인이름≫B≪/내국인이름≫는

An example of anonymized court judgment initially collected for data construction

이 총책인 중국 산둥성 칭다오시에 있는 금융기관 사칭 보이스피싱 조직(이하 '이 사건 보이스피싱 조직'이라 한다)의 구성원은 (가명 AA), (가명 AB)이 중간관리자급 팀장으로, 피고인(가명 AC) 및 (가명 AD), (가명 AE 또는 AF), (가명 AG), (가명 AH), (가명 AI), (가명 AJ), (가명 AK), (가명 AL), (가명 AM), (가명 AN) 등이 보이스피싱 콜센터 상담원으로 있었다.
피고인과 이 사건 보이스피싱 조직의 조직원들은 중국 산둥성 칭다오시에 있는 아파트, 아파트 등지에 보이스피싱 콜센터 사무실과 조직원들의 숙소를 차려두고 불특정 다수의 대한민국 사람들을 상대로 전화하여 AO, AP, AQ의 직원이라고 말하면서 금융기관을 사칭하여 '기존의 대출금을 상환하면 저금리 대환대출을 해주겠다'고 속이고 금원을 대포 계좌로 교부받기로 공모하였다.
이에 따라 AR은 총책으로서 중국 산둥성칭다오시에 콜센터 사무실과 콜센터 상담원 숙소를 임차하여 사무실에 컴퓨터, 전화기, 책상 등을 구비하는 한편, 보이스피싱 대상자에게 연락하기 위한 DB(데이터베이스) 자료, 피해자들로부터 피해금원을 송금받을 대포통장을 구해오고 피해금원이 대포통장을 통해 보이스피싱 조직의 수익금 계좌로 입금되면 조직원들의 맡은 업무와 실적에 따라 팀장인 AA을 통해 보이스피싱 조직원들에게 수익금을 분배하는 역할, AB은 콜센터 팀장으로서 상담원들의 업무와 숙식을 관리·감독하면서 콜센터 1차 상담원 업무를 담당할 신규 조직원을 섭외하여 위 조직에 가담하게 하고 총책 로부터 받은 보이스피싱 대상자들의 DB 자료를 콜센터 상담원들에게 나누어준 다음 콜센터 상담원들에게 보이스피싱 대상자에게 연락하도록 독려하고 1차 상담원들이 보이스피싱 대상자들을 속여 위 콜센터 2차 상담 전화번호로 전화가 걸려오면 '기존 대출금을 상환해야 한다'고 말하면서 대포통장 계좌를 불러주어 대포통장 계좌로 대출금 명목으로 피해금을 입금하게 하고 로부터 정산받은 보이스피싱 수익금을 콜센터 상담원들에게 분배하는 역할, 피고인 및 AC 등은 콜센터 상담원으로서 평일 아침 콜센터 사무실로 출근하여 등으로부터 받은 DB 자료를 토대로 보이스피싱 대상자들에게 전화하여 AS AT 등 금융기관 직원을 사칭하면서 '기존의 대출금을 상환하면 저금리 대환대출을 해주겠다'고 거짓말하고 2차 상담원에게 연결해주거나 '신용보증기금 보증서 및 신용 등급 향상을 위한 조회 건수 삭제 비용이 든다'고 말하여 이에 속은 피해자로 하여금 대포통장 계좌로 금원을 이체하도록 하는 등의 역할을 각각 분담하였다.
위와 같은 공모 내용과 역할 분담에 따라 이 사건 보이스피싱 조직원인 는 2017. 12. 27. 오전경 중국 산둥성 칭다오 이하 불상지에 있는 보이스피싱 콜센터 사무실에서, 발신번호 AU 번호로 피해자 에게 전화하여 'AV대리'를 사칭하면서 '기존에 있던 대출금을 상환하면 저금리로 대출해주겠다'고 거짓말을 하였다.
그러나 피고인 및 이 사건 보이스피싱 조직원들은 직원이 아니었고 피해자로부터 금원을 송금받으면 이를 일정한 비율에 따라 서로 나누어가질 생각이었다. 그럼에도 불구하고 피고인들 및 이 사건 보이스피싱 조직원들은 피해자로부터 2017. 12. 29.경 대포통장 계좌인 AW명의 우체국 계좌(계좌번호 :계좌번호 1 생략)로 5,530,000원, 2018. 1. 3.경 대포통장 계좌인 명의 조합 계좌(계좌번호 : 계좌번호 2 생략)로 1,000,000원씩 3회에 걸쳐 3,000,000원을 무통장송금 또는 계좌이체 송금받는 등 8,530,000원을 송금받은 것을 비롯하여 그 때부터 2018. 1. 31.경까지 별지 범죄일람표 기재와 같이 총 16명으로부터 합계 95,311,339원을 송금받았다.
이로써 피고인은 이 사건 보이스피싱 조직원들과 공모하여 피해자들을 기망하여 재물을 교부받았다.

The judgment data annotated pursuant to the three-tiered categorization scheme of PII

이 총책인 중국 산둥성 칭다오시에 있는 금융기관 사칭 보이스피싱 조직(이하 '이 사건 보이스피싱 조직'이라 한다)의 구성원은 (가명 <<<내국인이름>>>AA<<</내국인이름>>>), (가명 <<<내국인이름>>>AB<<</내국인이름>>>)이 중간관리자급 팀장으로, 피고인(가명 <<<내국인이름>>>AC<<</내국인이름>>>) 및 (가명 <<<내국인이름>>>AD<<</내국인이름>>>), (가명 <<<내국인이름>>>AE<<</내국인이름>>> 또는 <<<내국인이름>>>AF<<</내국인이름>>>), (가명 <<<내국인이름>>>AG<<</내국인이름>>>), (가명 <<<내국인이름>>>AH<<</내국인이름>>>), (가명 <<<내국인이름>>>AI<<</내국인이름>>>), (가명 <<<내국인이름>>>AJ<<</내국인이름>>>), (가명 <<<내국인이름>>>AK<<</내국인이름>>>), (가명 <<<내국인이름>>>AL<<</내국인이름>>>), (가명 <<<내국인이름>>>AM<<</내국인이름>>>), (가명 <<<내국인이름>>>AN<<</내국인이름>>>) 등이 보이스피싱 콜센터 상담원으로 있었다.
피고인과 이 사건 보이스피싱 조직의 조직원들은 중국 산둥성 칭다오시에 있는 아파트, 아파트 등지에 보이스피싱 콜센터 사무실과 조직원들의 숙소를 차려두고 불특정 다수의 대한민국 사람들을 상대로 전화하여 <<<은행>>>AO<<</은행>>>, <<<은행>>>AP<<</은행>>>, <<<은행>>>AQ<<</은행>>>의 직원이라고 말하면서 금융기관을 사칭하여 '기존의 대출금을 상환하면 저금리 대환대출을 해주겠다'고 속이고 금원을 대포 계좌로 교부받기로 공모하였다.
이에 따라 <<<내국인이름>>>AR<<</내국인이름>>>은 총책으로서 중국 산둥성칭다오시에 콜센터 사무실과 콜센터 상담원 숙소를 임차하여 사무실에 컴퓨터, 전화기, 책상 등을 구비하는 한편, 보이스피싱 대상자에게 연락하기 위한 DB(데이터베이스) 자료, 피해자들로부터 피해금원을 송금받을 대포통장을 구해오고 피해금원이 대포통장을 통해 보이스피싱 조직의 수익금 계좌로 입금되면 조직원들의 맡은 업무와 실적에 따라 팀장인 <<<내국인이름>>>AA<<</내국인이름>>>을 통해 보이스피싱 조직원들에게 수익금을 분배하는 역할, <<<내국인이름>>>AB<<</내국인이름>>>은 콜센터 팀장으로서 상담원들의 업무와 숙식을 관리·감독하면서 콜센터 1차 상담원 업무를 담당할 신규 조직원을 섭외하여 위 조직에 가담하게 하고 총책 로부터 받은 보이스피싱 대상자들의 DB 자료를 콜센터 상담원들에게 나누어준 다음 콜센터 상담원들에게 보이스피싱 대상자에게 연락하도록 독려하고 1차 상담원들이 보이스피싱 대상자들을 속여 위 콜센터 2차 상담 전화번호로 전화가 걸려오면 '기존 대출금을 상환해야 한다'고 말하면서 대포통장 계좌를 불러주어 대포통장 계좌로 대출금 명목으로 피해금을 입금하게 하고 로부터 정산받은 보이스피싱 수익금을 콜센터 상담원들에게 분배하는 역할, 피고인 및 <<<내국인이름>>>AC<<</내국인이름>>> 등은 콜센터 상담원으로서 평일 아침 콜센터 사무실로 출근하여 등으로부터 받은 DB 자료를 토대로 보이스피싱 대상자들에게 전화하여 <<<은행>>>AS<<</은행>>> <<<은행>>>AT<<</은행>>> 등 금융기관 직원을 사칭하면서 '기존의 대출금을 상환하면 저금리 대환대출을 해주겠다'고 거짓말하고 2차 상담원에게 연결해주거나 '신용보증기금 보증서 및 신용 등급 향상을 위한 조회 건수 삭제 비용이 든다'고 말하여 이에 속은 피해자로 하여금 대포통장 계좌로 금원을 이체하도록 하는 등의 역할을 각각 분담하였다.
위와 같은 공모 내용과 역할 분담에 따라 이 사건 보이스피싱 조직원인 는 2017. 12. 27. 오전경 중국 산둥성 칭다오 이하 불상지에 있는 보이스피싱 콜센터 사무실에서, 발신번호 <<<전화번호>>>AU<<</전화번호>>> 번호로 피해자 에게 전화하여 '<<<내국인이름>>>AV<<</내국인이름>>>대리'를 사칭하면서 '기존에 있던 대출금을 상환하면 저금리로 대출해주겠다'고 거짓말을 하였다.
그러나 피고인 및 이 사건 보이스피싱 조직원들은 직원이 아니었고 피해자로부터 금원을 송금받으면 이를 일정한 비율에 따라 서로 나누어가질 생각이었다. 그럼에도 불구하고 피고인들 및 이 사건 보이스피싱 조직원들은 피해자로부터 2017. 12. 29.경 대포통장 계좌인 <<<내국인이름>>>AW<<</내국인이름>>>명의 <<<은행>>>우체국<<</은행>>> 계좌(계좌번호 :<<<계좌번호>>>계좌번호 1 생략 <<</계좌번호>>>)로 5,530,000원, 2018. 1. 3.경 대포통장 계좌인 명의 조합 계좌(계좌번호 : <<<계좌번호>>>계좌번호 2 생략<<</계좌번호>>>)로 1,000,000원씩 3회에 걸쳐 3,000,000원을 무통장송금 또는 계좌이체 송금받는 등 8,530,000원을 송금받은 것을 비롯하여 그 때부터 2018. 1. 31.경까지 별지 범죄일람표 기재와 같이 총 16명으로부터 합계 95,311,339원을 송금받았다.
이로써 피고인은 이 사건 보이스피싱 조직원들과 공모하여 피해자들을 기망하여 재물을 교부받았다.

Figure B.1: Examples of court judgment data before and after annotation.

…", where the label 식당 refers to a place for eating and drinking.

Example without a functional descriptor. In constrast, some place names do not contain an explicit functional descriptor. For example, the sentence "... 피해자 김철수를 기다리며 맥도날드에서 음식을 주문하고 ..." ("... ordered food at McDonald's while waiting for the victim Kim Chulsoo to arrive ...") can be anonymized to "... 피해자 D를 기다리며 E에서 음식을 주문하고 ..." ("... ordered food at E while waiting for the victim D to arrive ..."). In this case, "맥도날드" ("McDonald's") does not have a functional descriptor, so court officials are instructed to replace the entire word with a placeholder, E. Therefore, "... 피해자 D를 기다리며 E에서 음식을 주문하고 ..." in the de-identified judgment will be labeled by the annotators as follows: "... 피해자 ≪내국인 이름≫D≪/내국인 이름≫를 기다리며 ≪식당≫E≪식당≫에서 음식을 주문하고 ...."

D Personally Identifiable Information (PII) Categorization

Appendix B provides a complete overview of the three-tiered categorization scheme classifying personal identifiers in the domain of law and adjudication, as detailed in Section 3.4. Under two main categories, 16 subcategories, and 80 granular categories, we present a total of 729 labels alphabetically ordered in Korean along with the English translation of each label.

D.1 사건관계인 특정 정보 (Direct identifiers)

D.1.1 인명 (Names)

내국인이름 (Korean names):

외국인이름 (Non-Korean names): 몽골인이름 (Mongolian names), 베트남이름 (Vietnamese names), 세례명 (baptismal names), 영어이름 (English names), 일본인이름 (Japanese names), 중국인이름 (Chinese names), 캄보디아이름 (Cambodian names), 태국인이름 (Thai names), 필리핀이름 (Filipino names), 러시아권이름 (Russian names), 법명(Dharma names)

아이디•닉네임 (IDs and Nicknames): 가수 (aliases), 닉네임 (nicknames), 대화명 (usernames), 별명 (nicknames), 블로그 (blogs), 아이디 (IDs), 법호(Dharma nickname)

D.1.2 연령정보 (Age and Date of Birth)

나이(age), 출생연도(year of birth), 생년월일(date of birth)

D.1.3 이메일주소 (Email Address)

이메일주소 (email address)

D.1.4 주민등록번호 (Resident Registration Number)

D.2 기타 (사건관계인이나 제3자를 특정할 수 있는) 정보 (Quasi-identifiers)

D.2.1 사건관계인이력 (Work and Criminal backgrounds of the persons involved in the case)

범죄경력 (Criminal records): 죄 (crime)

D.2.2 사건 관련 숫자 정보 (Incident-related numerical information)

고유번호 (Various Numbers Uniquely Identifying Specific Individuals and Objects): 계좌번호 (bank account number), 관리번호 (management number), 금괴일련번호 (gold bar serial number), 사건번호 (case number), 선박번호 (IMO ship number), 비트코인개인지갑 (bitcoin wallet), 수표번호 (check number), 카드번호 (card number), 어선 (fishing vessel number), 어음번호 (bill number), 범죄경력등조회회보서 (criminal record certificate), 차량번호 (vehicle registration number), 특허번호 (patent number), 휴대폰번호 (mobile phone number), 군번 (military service number), 면허번호 (license number), 훈장번호 (decoration number), 전화번호 (phone number), 내선번호 (extension number), 수험번호 (examination number), 보훈번호 (veterans registration number), 보증번호 (guarantee number), 고시번호 (official notice number), 비밀번호 (password / PIN), 등기번호 (registration number), 사업자등록번호 (business registration number), 접수번호 (receipt number), 민원번호 (civil complaint number), 경매번호 (auction number), 채권번호 (bond number), 일련번호 (serial number), 법인등록번호 (corporate registration number)

장소 관련 번호 (Numbers Assigned to Specific Places): 골프장코스 (golf course), 구역 (zone), 라인 (line), 지하철칸 (subway compartment), 항공편 (flight number) 번(number), 호선(line number), 호실(room number), 호(unit number), 출구번호 (exit number), 동(building number), 층(floor), 노선번호(route number), 레일(rail number), 승강장번호(platform number), 열차번호(train number), 탑승장번호(boarding platform number), 번호(num-

ber), 광역버스(express bus number), 단지 (housing complex), 로트 (lot), 블록 (block), 실(room), 번지(lot address number)

기타 사건 관련 숫자 (Other Incident-related Numbers): 기수 (class number), 명수(number of people), 연도(year), 날짜(date)

D.2.3 사건 관련 장소 (Incident-related sites and locations)

시설 내부 공간 (Interior Spaces): 건물내장소 (a place in the building), 공공기관내장소 (a place in the public institution), 공원내장소 (a place in the park), 광장(square), 소분류장(small classification yard), 사무실(office), 교도소내장소 (a place in the correction facility), 구치소내장소 (a place in the detention center), 대학교내장소 (a place in the university), 문 (gate),물류센터레일 (rails at logistics center), 법원내장소 (a place in the courthouse), 병원내장소 (a place in the hospital), 생활관 (residential hall), 아파트내장소 (a place in the apartment), 기숙사 (dormitory), 군부대내장소 (a place in the military facility)

교통 (Transport Infrastructure): 버스공영차고지 (bus garage) 버스정류장 (bus stop), 요금소 (tollgate)

건설 (Construction Sites): 공사장 (construction yard), 현장 (site), 야적장 (storage yard) 공사현장 (construction site)

산림•하천 (Forest and Water): 둘레길 (perimeter trail), 등산로 (hiking path), 산책로 (walking trail), 약수터 (mineral spring)

해양 (Places related to Maritime Activities): 선박명 (ship name), 여객선 (passenger ship name), 군함명 (warship name)

D.2.4 지리정보 (Geographic information)

주소 (Address): 도아래주소 (address under province) 구아래주소 (address under district/Gu) 군아래주소 (address under county/Gun) 읍아래주소 (address under town/Eup) 동아래주소 (address under neighborhood/Dong) 시아래주소 (address under city/Si) 주소 (address) 임야 (forest land) 토지 (land) 필지 (parcel/lot) 국외주소 (overseas address)

지역명 (Geographic units): 마을 (village), 산 (mountain), 선거구 (coinstituency), 선거단위 (electoral district), 외국도시 (foreign city), 지구 (district), 해안지역명 (coastal area name), 해수욕장 (bathing beach), 국외하천 (overseas river/stream), 행정구 (Gu: district-level administrative unit), 행정군 (Gun: county-level administrative unit), 행정동 (Dong: neighborhood-level administrative unit), 행정리 (Ri: village-level administrative unit), 행정면 (Myeon: township-level administrative unit), 행정시 (Si: city-level administrative unit), 행정읍 (Eup: town-level administrative unit), 행정도 (Do: province-level administrative unit), 베트남전관련지명 (Vietnam War related place names), 지사및지청명 (branch and local office names), 특정지역범위명 (specific regional boundary names), 고지 (highland/hill), 국가명 (country name), 고개 (mountain pass), 섬이름 (island name), 전투지역(combat zone), 해변 (beach), 호수(lake), 하천 (river/stream),

도로명 (Roads and Streets): 골목 (alley), 교차로 (intersection), 길 (street), 도로 (road), 인터체인지 (interchange), 로터리 (rotary)

구간 (Sections) 도로구간 (road section), 철도구간 (railway section)

D.2.5 조직 (Organizations)

친목•문화 (Community Gatherings): 단체명 (uncategorized gatherings), 독서토론모임 (book club), 동호회 (uncategorized clubs), 모임 (social gatherings), 봉사단체 (volunteer group), 산악회 (hikers club), 연합회 (uncategorized coalitions), 체육회 (sports club)

사회•종교 단체(Social and Religious Groups): 노회 (presbytery), 사회복지법인 (social welfare organization), 종교단체 (religious organization), 종중 (class association)

정치•경제 단체 및 협의체 (Various Associations of Like-minded People in Politics, Commerce and Labor): 공제조합 (mutual aid association), 노동조합 (labor union), 선거캠프 (election camp), 재개발정비조합 (redevelopment partnership), 재건축정비조합 (reconstruction partnership), 정당 (political party), 조합 (uncategorized partnerships), 지역주택조합 (local housing association), 협동조합(cooperative association), 협의회 (uncategorized councils), 협회 (uncategorized associations), 상인회 (merchant association), 사단법인 (non-profit corporation), 의료법인 (medical corporation), 위원회 (committee), 재단법인 (foundation), 학교법인 (educational foundation), 의료재단 (medical foundation), 어촌계 (fishermen's association), 총회 (general assembly)

국방•치안 (Specific Units in Military and Law Enforcement Agencies): 국정원비밀조직 (secret agency under the National Intelligence Service), 대대 (battalion), 헌병대 (military police), 사단 (division), 여단 (brigade), 소대 (platoon), 연대 (regiment), 중대 (company), 사령부 (headquarters), 해군전단 (naval squadron), 본부 (headquarters), 해군함대 (naval fleet)

조직 내 세부부서 (Specific Units and Departments in the Organizations): 단과대학 (college), 반 (kindergarten class), 부서 (departments), 지회 (branches), 팀 (teams), 학과 (college majors), 교통공사내부서 (department within transportation corporation)

조직 내 업무•권한 등(Job levels and duties within organizations) 직급 (job level), 회원등급 (membership level), 군계급 (military rank), 직무 (job duty), 보직 (official post)

불법 단체 (Illegal Organizations): 범죄조직 (criminal organization)

D.2.6 기관 및 시설 (Institutions and Facilities):

정부기관 및 지방자치단체 (Public Administrative Bodies and Local Municipalities): 공사 및공단 (public institution), 시청 (city hall), 우체국 (post office), 행정복지센터 (community service center), 중앙행정기관 (central administrative agency), 해양및산림등관리기관 (maritime and forestry management agency), 교육청 (office of education), 등기소 (registry office), 세무서 (tax office)

군사 (Military Bases): 군부대 (military camp), 미군부대 (US Army), 훈련소 (military training center), 군정비및관리시설 (military maintenance and management facility), 군소속교육기관 (military-affiliated educational institution)

치안 및 교정 (Policing and Correctional Facilities): 경찰서 (police office), 경찰청 (national police agency), 구치소 (detention center), 지구대 (police substation), 치안센터 (community police center), 파출소 (police substation)

소방 및 재난 (Agencies for Fire Safety and Disaster Response): 소방서 (fire station), 안전센터 (safety center)

사회기반시설 (Public Infrastructure): 공항 (airport), 발전소 (power plant), 버스터미널 (bus terminal), 선착장 (dock), 육교 (pedestrian overpass), 저수지 (resorvoir), 지하차도 (underpass), 지하철역 (subway station), 태양광발전소 (solar power plant), 터널 (tunnel), 항구 (port), 교량 (bridge), 비행장 (airfield), 검사및검문소 (inspection and checkpoint), 상하수도시설 (water supply and sewage facilities), 변전소 (substation), 원자력본부 (nuclear headquarters), 부두 (pier/wharf), 기차역 (train station)

사회복지시설 (Social Security and Welfare Facilities): 복지시설 (welfare facility), 요양원 (nursing home), 육아원 (child care center), 장애인이용시설 (facility for person with disabilities), 재가장기요양기관 (home-based long-term care institution)

주민편의시설 (Residential Convenience Facilities): 공원 (park), 농어촌근린시설 (rural community facility), 마을회관 (community center), 주민쉼터 (community rest area), 경로당 (senior center), 유원지 (recreational area), 놀이터 (playground), 마당 (yard)

스포츠시설 (Sports Facilities) 경기장 (stadium), 야구장 (baseball stadium)

주거시설 (Residential Buildings): 고급주택 (luxury residence), 맨션 (low-rise apartment), 빌라 (multiplex housing), 아파트 (apartment), 오피스텔 (studio apartment), 주택 (single-family home), 타운하우스 (townhouse)

의료기관 (Healthcare Institutions): 내과 (internal medicine clinic), 병원 (hospitals), 산부인과의원 (OB-GYN clinic), 성형외과 (plastic surgery), 신경외과 (neurosurgery), 안과 (ophthalmalmic clinic), 요양병원 (nursing hospital), 의원 (local clinic), 정신병원 (mental hospital), 정형외과 (orthopedics clinic), 치과 (dentistry), 치과의원 (dental clinic), 의료원 (medical center), 보건소 (public health center), 재활원 (rehabilitation center), 이비인후과 (ENT clinic), 피부과 (dermatology clinic) 한방병원 (Korean medicine hospital), 한의원 (oriental medicine clinic)

교육기관 (Educational Institutions): 고등학교 (highschool), 대학교 (university), 어린이집 (daycare center), 연수원 (training center), 유치원 (kindergarten), 중학교 (middle school), 직업능력개발훈련시설 (vocational training center), 초등학교 (elementary school), 국외중고등학교 (overseas middle and high school), 국외대학교 (overseas university), 사관학교 (military academy), 전

문학교(vocational school), 중고등학교 (secondary school)

문화•예술 (Art and Cultural Facilities): 도서관 (library), 문화시설 (culture center), 미술관 (art museum), 전시장 (exhibition hall), 청소년수련관 (youth training center), 박물관(museum)

종교시설 (Place of Worship): 교회 (church), 사찰 (temple)

상업시설 (Commercial Buildings and Facilities): 빌딩 (building), 상가 (shopping plaza), 시장 (market), 아울렛 (outlet), 장례식장 (funeral home), 전기차충전소 (EV charging station), 중고차매매단지 (used car sales complex), 지하상가 (underground shopping center), 휴게소 (rest area), 예식장 (wedding hall), 매표소 (ticket booth), 놀이시설 (amusement facility), 건물 (building), 모델하우스 (show house), 자동차매매단지 (car sales complex)

연구개발기관 (Research and Development Institutions): 연구소 (research institute), 주행시험장 (driving test center)

산업•물류 (Industrial Development and Logistic Complex): 공단 (public corporation), 물류단지 (logistics complex), 산업단지 (industrial complex)

복합단지 및 개발지구 (Industrial Development and Logistic Complex): 친환경복합단지 (eco-friendly complex)

금융관련공공기관 (financial regulators): 금융기관 (financial services agency) , 은행 (bank), 저축은행 (savings bank)

D.2.7 사업체 (Corporate entities)

외식업 (Eating and Drinking Places): 가요주점 (karaoke pub), 고깃집 (Korean BBQ restaurant), 노래주점 (singing bar), 다방 (traditional Korean cafe), 동남아음식점 (Southeast Asian restaurant), 라이브카페 (live music cafe), 레스토랑 (restaurant), 바 (bar), 분식점 (snack bar), 뷔페 (buffet restaurant), 애견카페 (pet cafe), 일식당 (Japanese restaurant), 주점 (pub), 중식당 (Chinese restaurant), 치킨집 (fried chicken restaurant), 푸드트럭 (food truck), 한식당 (Korean restaurant), 해외식당 (international restaurant), 횟집 (sashimi restaurant), 키즈카페 (kids cafe), 카페 (cafe)

도•소매 및 유통 (Wholesale and Retail Trade): 가게 (shop), 가구매장 (furniture store), 가스업체 (gas supply company), 가전제품판매업 (home appliance store), 고물업체 (scrap metal business), 골프용품판매점 (golf equipment store), 과일가게 (fruit shop), 귀금속점 (jewelry store), 꽃가게 (flower shop), 농산물판매업 (agricultural product sales), 대리점 (distributor), 떡집 (rice cake shop), 마트 (grocery store), 문구점 (stationery store), 반찬가게 (side dish shop), 백화점 (department store), 빵집 (bakery), 상품권판매업체 (gift certificate vendor), 생활용품매장 (household goods store), 석유대체연료판매업체 (alternative fuel retailer), 슈퍼마켓 (supermarket), 스포츠용품점 (sporting goods store), 스포츠의류 (sportswear), 식품유통업 (food distribution business), 신발판매업체 (shoe store), 악기회사 (musical instrument company), 안경점 (optical shop), 반려동물분양업체 (pet shop), 약국 (pharmacy), 오디오샵 (audio equipment store), 옷가게 (clothing store), 원단공급업체 (fabric supplier), 유압벨브판매업체 (hydraulic valve vendor), 유통업 (distribution business), 의류매장 (apparel store), 자동차대리점 (car dealership), 자동차백화점 (auto megastore), 자동차판매점 (car sales shop), 전자제품매장 (electronics store), 정육점 (butcher shop), 제과점 (pastry shop), 주유소 (gas station), 중고도서매매업체 (used book store), 중고차매매업체 (used car dealership) 카드단말기판매업체 (credit card terminal distributor), 캠핑업체 (camping service provider), 컴퓨터판매업체 (computer retailer), 타이어판매업체 (tire shop), 페인트판매업 (paint supplier), 편의점 (convenience store), 화원 (flower shop), 화학약품판매업 (chemical supplier), 휴대전화판매업체 (mobile phone store), 휴대폰케이스매장 (mobile accessories shop), 보청기판매점 (hearing aid store), 자전거판매업 (bicycle shop), 기계도소매업 (machinery wholesale and retail business) 수산물유통업 (seafood distribution business), 매장 (store), 매점 (shop)

금융•세무 (Financial Institutions, Insurance and Other Financial Intermediaries): , 금융회사 (financial company), 대부업 (loan business), 보험사 (insuarance company), 신탁회사 (trust company), 전당포 (pawnshop), 증권사 (securities company), 카드회사 (credit card company), 투자회사 (investment frim), 해외은행 (foreign bank), 해외증권사 (foreign securities company), 세무법인 (tax corporation), 회계법인 (accounting corporation), 감정평가법인 (appraisal corporation), 집합투자기구 (collective investment scheme), 감정평가사사무소 (appraisal office), 세무사사무소 (tax accountant office)

법무 (Law Practice): 법률사무소 (law office), 법무법인 (law firm), 노무법인 (labor law firm),

법무사사무소 (judicial scrivener office)

부동산 중개 및 임대 매매 (Real Estate Business): 공인중개사 (real estate agent), 부동산매매임대회사 (real estate sales and rental company), 부동산분양사무실 (real estate sales office), 분양대행사 (real estate marketing agency), 중개법인 (brokerage corporation)

정보통신업 (Information and Communications): 방송국 (broadcasting station), 신문 (newspaper), 언론사 (media company), 출판사 (publishing company), 통신사 (telecommunications company), 전화국 (telephone office), 방송(broadcasting),

건설 (Construction): 건설업체 (construction company), 부동산개발업 (real estate development business), 토공사 (civil engineering company), 토목업 (civil engineering business), 재개발업체 (redevelopment company), 조경업체 (landscaping company)

운수업 (Transportation): 택배및운송회사 (transportation company), 택시회사 (taxi company), 여객운송회사 (passenger transport company), 이삿짐센터 (moving company)

물류 (Logistics and Distribution): 물류센터 (logistics center), 물류창고 (logistics warehouse), 물류회사 (logistics company)

제조업 (Manufacturing): 가구공장 (furniture factory), 건설자재회사 (building materials company), 공장 (factory), 금속제조업 (metal manufacturing), 기계설비회사 (machinery and equipment company), 목공소 (woodworking shop), 미용기기업체 (beauty equipment company), 보일러회사 (boiler manufacturer), 복합기업체 (multifunction printer manufacturer), 봉제업체 (sewing company), 비료회사 (fertilizer company), 석재가공업체 (stone processing company), 선박제조업 (shipbuilding company), 식품가공업 (food processing company), 식품업체(food company), 식품회사 (food company), 육류업체 (meat processing company), 음료회사 (beverage company), 의료기기회사 (medical device company), 의류브랜드 (clothing brand), 이동식주택 (mobile home manufacturer), 자동차부품생산업체 (auto parts manufacturer), 자동차회사 (automobile company), 전기배터리업체 (battery manufacturer), 전자전기제조업 (electronic component manufacturing), 조선회사 (shipbuilding company), 주류회사 (alcoholic beverage company), 질소발생기제조업체 (nitrogen generator manufacturer), 철골제조업(steel frame manufacturing), 철판제조업 (steel plate manufacturing), 철판가공업 (sheet metal processing), 플라스틱가공업 (plastic processing company), 화장품회사 (cosmetics company), 중공업회사 (heavy industry company), 세라믹제조업 (ceramics manufacturing), 침장제조업 (bedding manufacturing company), 화학공업사 (chemical industry company), 정미소 (rice mill), 제약회사 (pharmaceutical company), 제철소 (steel mill)

농축물수산업 및 임업 (Agriculture, Fisheries, and Forestry): 농장 (farm), 축산농장 (livestock farm), 어업회사 (fishery company), 과수원 (orchard)

광업 및 각종 자원 채굴•채취 (Mining and Quarrying): 금광채굴업 (gold mining), 광업소 (mining company)

숙박업 (Lodging and Accommodation): 고시원 (gosiwon: a small single-room accommodation), 리조트 (resort), 모텔 (motel), 무인텔 (unmanned motel), 산장 (mountain lodge), 여관 (inn), 콘도 (condominium resort), 펜션 (pension), 호텔 (hotel), 게스트하우스 (guest house), 숙박시설 (lodging facility)

오락 및 스포츠 (Recreation, Leisure and Sports): 극단 (theater troupe), PC방 (internet café), 게임장 (arcade), 골프연습장 (golf practice range), 골프장 (golf club), 낚시터 (fishing spot), 노래방 (karaoke room), 당구장 (billiard hall), 볼링장 (bowling alley), 수영장 (swimming pool), 승마장 (equestrian center), 실내낚시터 (indoor fishing cafe), 영화관 (movie theater), 오락실 (arcade), 온천 (hot spring), 워터파크 (water park), 캠핑장 (campground), 풀장 (pool), 헬스장 (fitness center), 기원 (baduk club), 당구장 (billiard hall), 수족관 (aquarium), 야구연습장 (batting cage), 스키장 (ski resort), 수상레저업(water leisure business)

미용•욕탕•신체관리 서비스 (Beauty and Body Care): 마사지 (massage shop), 목욕탕 (public bathhouse), 미용실 (hair salon), 사우나 (sauna), 안마시술소 (massage parlor), 안마원 (therapeutic massage clinic), 왁싱샵 (waxing shop), 이발소 (barbershop), 찜질방 (Korean spa), 피어싱 (piercing studio), 반려동물미용샵 (pet grooming shop), 네일샵 (nail salon)

유흥업 (Adult entertainment): 나이트클럽 (nightclub), 노래빠 (karaoke bar), 단란주점 (karaoke lounge with host services), 룸살롱 (high-end adult entertainment venue with private rooms),

업소 (adult entertainment venue), 카지노 (casino), 클럽 (club), 호스트바 (host bar), 무도장 (dance hall)

서비스 일반 (Other Service Sectors): 건물임대관리회사 (building management company), 광고회사 (advertising company), 대리운전회사 (designated driver service company), 동물병원 (veterinary clinic), 방역회사 (pest control company), 배달대행업체 (delivery agency), 상담소 (counseling center), 상조회사 (funeral service agency), 선박임대판매업 (ship rental and sales company), 세차장 (car wash), 세탁소 (laundry), 소개소 (labor dispatch agency), 스튜디오 (studio), 여행사 (travel agency), 오토바이수리점 (motorcycle repair shop), 요가학원 (yoga studio), 용역회사 (outsourcing service company), 운전면허학원 (driving school), 유학알선업체 (study abroad agency), 인터넷설치업체 (internet installation service), 인테리어 (interior design service), 자동차임대업체 (car rental company), 재직정보제공업체 (employment verification company), 전자제품렌탈업 (electronics rental business), 정비공업사 (auto repair shop), 주차장 (parking lot), 주차장관리회사 (parking lot management company), 철거업체 (demolition company), 철학관 (fortune telling house), 청소대행업체 (cleaning service company), 컨설팅 (consulting firm), 태권도장 (Taekwondo gym), 택시면허매매중개업 (taxi license brokerage), 파티룸 (party room), 학원 (private academy), 기업행사대행업체 (event management company), 화실 (art studio), 점집 (fortune telling house), 운전면허시험장 (driver's license test center), 보안경비회사 (security management company), 폐기물처리업체 (waste disposal company), 금속분석업 (metal analysis business), 산후조리원 (postnatal care center), 결혼준비대행업 (wedding planning agency), 건축사사무소 (architectural office), 사진관 (photo studio), 음원서비스 (music streaming service)

기업 일반 (Companies and Businesses in General): IT회사 (IT company), 불특정회사명 (unspecified company), 무역회사 (trading company), 유한공사 (limited company), 유한회사(limited liability company), 주식회사 (corporation), 지점 (branch office), 지주회사 (holding company), 국외기업 (foreign company), 상사회사 (trading company), 합자회사 (limited partnership company), 합명회사 (general partnership company), 농업회사법인 (agricultural corporation), 영농조합법인 (agricultural cooperative corporation)

D.2.8 상품 일반 (Consumer Products)

식•의약품 (Foods and Medical Products): 식품(food), 음료 (beverage), 의약품 (pharmaceutical product), 피자 (pizza)

공산품 (Industrial Products): 가전제품 (home appliance), 공작기계 (machine tool), 마스크팩 (sheet mask), 발전기 (generator), 악기모델명 (instrument model name), 임플란트제품명 (dental implant product), 작업용차량 (work vehicle), 차량종류 (vehicle type), 철강제품 (steel product), 의료기기 (medical device), 교구 (teaching aid), 미용제품 (cosmetic product), 불특정제품명 (unspecified product name), 농약 (pesticide), 비료 (fertilizer), 항공기 (aircraft)

출판물 (Publications) 서적 (books)

정보통신 상품 (Computer Equipment and Software): 소프트웨어 (software)

D.2.9 방송통신서비스 (Media and Telecommunications)

온•오프라인 방송 (Streaming and Broadcasting service): 방송마일리지 (streaming donation points), 방송프로그램 (broadcasting program), 방송플랫폼 (streaming platform)

플랫폼 일반 (Online platforms in General): 구인사이트 (job search site), 번역사이트 (translation site), 사이트 (uncategorized websites), 어플 (application), 포털 (portal site), 보이스피싱어플리케이션 (voice phising application)

전자상거래 (E-commerce): 미술품경매사이트 (art auction site), 배달어플리케이션 (delivery application), 쇼핑몰 (online shopping mall), 중고거래사이트 (secondhand marketplace website)

소셜미디어 (Social Media): SNS (social networking service), 밴드 (group communication application), 소개팅어플리케이션 (dating application), 인터넷동성사이트 (LGBT dating app), 채팅어플리케이션 (chatting application), 커뮤니티사이트 (online community site), 국외메신저 (foreign messaging application), 온라인게시판명 (name of online bulletin board), 온라인게시글명 (title of online post), 온라인대화방명 (name of online chat room)

게임 (Online Games): 게임마일리지 (game mileage), 게임서버 (game server), 게임아이템 (in-game item), 게임아이템거래카페 (item trading forum), 모바일게임 (mobile game), 온라인게임 (online game), 인터넷도박 (online gambling)

D.2.10 금융서비스 (Financial Products and Services)

투자•보험•대출 서비스 (Investment, Insurance and Personal Loan Services): 골프보험 (golf insurance), 금융투자상품 (financial investment product), 대출상품 (loan product), 보험상품 (insurance plan)

가상자산 (Virtual Assets): 가상화폐 (cryptocurrency), 가상화폐거래프로그램 (crypto trading platform), 가상화폐거래소 (cryptocurrency exchange)

D.2.11 사회•문화 (Culture and Society)

국가유산 (National Heritage and Other Cultural Features): 중요무형문화재 (intangible cultural heritage)

예술 (Fine Arts, Visual Arts, Performing Arts): 공연 (performance), 영화 (film)

교육 및 학술 (Education Programs and Academic Curriculum): 교과목 (curriculum)

각종 행사 (Socio-cultural Events): 공청회 (public hearing), 낚시대회 (fishing competition), 등산행사 (hiking event), 임플란트세미나 (implant seminar), 사회공헌•자선행사 (charity event), 축제 (festival), 행사 (uncategorized events)

스포츠 (Sports) 운동종목 (sports category)

각종 과업 (Various Projects) 공사 (construction work), 사업 (project), 용역 (service contract)

D.2.12 URL

URL: URL

E Model and Training

This section provides additional details on our model architecture, and training procedures introduced in the main paper (Section 4). We first describe Thunder-DeID model family developed for Korean court judgment de-identification. We then outline the pre-training and fine-tuning strategies applied to both our models and the baselines.

E.1 Model configuration

Model. We introduce Thunder-DeID, a family of models based on the DeBERTa-v3 architecture, designed for de-identification through token classification. Thunder-DeID family includes three models: 370M, 800M and 1.5B models. The 370M model has 370 million parameters, a hidden dimension of 1024, 24 transformer layers, 16 attention heads, and a vocabulary size of 32,000. The 800M model has 800 million parameters, a hidden dimension of 1280, 36 transformer layers, 20 attention heads, and a vocabulary size of 32,000. The 1.5B model has 1.5 billion parameters, a hidden dimension of 2048, 24 transformer layers, 32 attention heads, and a vocabulary size of 128,000. The smaller vocabulary size for the 370M and 800M models prevents the embedding matrix from becoming disproportionately large relative to the transformer layers to ensure balanced model architecture.

E.2 Training

Pre-training. Thunder-DeID models are pre-trained from scratch on the bilingual corpus from Section 4.1, which yields 60 billion tokens (22 billion Korean, 38 billion English) when tokenized with our custom tokenizer. The 370M model is pre-trained on a 14 billion token subset (7 billion Korean, 7 billion English) sampled from the corpus, conducted over 2 hours using 32 NVIDIA H100 80GB GPUs. The 800M model is pre-trained on a 30 billion token subset (15 billion Korean, 15 billion English) sampled from the corpus, conducted over 9 hours using 32 NVIDIA H100 80GB GPUs. The 1.5B model is pre-trained on the full 60 billion tokens (22 billion Korean, 38 billion English), conducted over 19 hours using 32 NVIDIA H100 80GB GPUs. For the 370M model, initial pre-training uses a global batch size of 2048, a peak learning rate of 7.5e-5, a masked language modeling (MLM) probability of 0.15, and a maximum sequence length of 512, with the DeepSpeed framework under ZeRO Stage 0 (DDP). For the 800M and 1.5B models, the same configuration is used but with a peak learning rate of 5e-5. All models are optimized using AdamW (Loshchilov and Hutter, 2019) optimizer with $\beta = (0.9, 0.999)$. A learning rate schedule with a warm-up phase for the first 10% of training steps and cosine decay for the remainder is applied across all models. To handle longer inputs, each model undergoes additional training on 2 million tokens with a maximum sequence length of 2048, using the same learning rate schedule. All models use FP16 mixed precision (Micikevicius et al., 2017) training.

Fine-Tuning. We fine-tune Thunder-DeID models and the baseline models on the token classifica-

tion task using the dataset (Section 4.1). We employ two data augmentation settings: Per-Epoch Entity Replacement, where entity mentions in each document are replaced with new samples from a predefined list at every epoch to increase data diversity, and Single Replacement, where entity mentions are replaced once and remain fixed throughout training. At each epoch under Per-Epoch Entity Replacement, the model sees a different variant of every document, and the full training completes over 30 epochs to cover the entire augmented set. The validation set remains unchanged to ensure consistent evaluation. For the 370M model, we set the global batch size to 32, the peak learning rate to 5e-5. For the 800M model, 1.5B model, Polyglot-Ko and Exaone-3.5, the same configuration is used but with a peak learning rate of 2e-5. All models are trained with a maximum input length capped at 2048 tokens (the model limit). Inputs longer than this limit are truncated from the end (head-only, tail truncation), so very long court rulings—especially civil and administrative cases—may not be fully covered by the model input. We apply FP16 mixed precision across all models and optimize these models using AdamW optimizer with $\beta = (0.9, 0.999)$.

F Evaluation Metrics

This section details the evaluation metrics used to assess model performance in the de-identification of Korean court judgments, as discussed in the main paper (Section 4). We describe Binary Token-Level F1 and Token-Level Micro F1, including their mathematical definitions and significance for result analysis.

Backgrounds. In token classification for de-identification, model performance is measured with true positives (TP), false positives (FP), and false negatives (FN). TP is the number of tokens correctly predicted as a target label. FP is the number of tokens incorrectly predicted as a target label when they belong to another label. FN is the number of tokens belonging to a target label but incorrectly predicted as another label. Precision is the proportion of correctly predicted tokens among all tokens predicted as the target label, and recall is the proportion of correctly predicted tokens among all tokens truly belonging to the target label. These are defined as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$
$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Binary Token-Level F1. Binary Token-Level F1 evaluates the model's ability to classify tokens requiring de-identification from those that do not regardless of entity type. High scores ensure accurate detection of all tokens requiring de-identification like "홍길동" (Hong Gildong) while excluding others like "이" (i). This metric is critical because missing even one token that requires de-identification can immediately lead to increase identifiability of the person and thus compromise privacy. By treating all entity types as a single class, it provides a simple yet robust baseline widely adopted in de-identification research (Dernoncourt et al., 2016; Yue and Zhou, 2020; Salierno et al., 2024; Kim et al., 2024). The binary token-level F1 score in our experiment is calculated as follows:

$$\text{Binary Precision} = \frac{\text{TP}_{\text{bin}}}{\text{TP}_{\text{bin}} + \text{FP}_{\text{bin}}}$$
$$\text{Binary Recall} = \frac{\text{TP}_{\text{bin}}}{\text{TP}_{\text{bin}} + \text{FN}_{\text{bin}}}$$

Aspect	Thunder-DeID-370M	Thunder-DeID-800M	Thunder-DeID-1.5B	Polyglot-Ko	EXAONE-3.5
Parameters	370M	800M	1.5B	1.3B	2.4B
Hidden Dimension	1024	1280	2048	2048	2560
Transformer Layers	24	36	24	24	30
Attention Heads	16	20	32	16	32
Vocabulary Size	32,000	32,000	128,000	30,080	102,400
Pre-train Corpus	14B (7B Ko / 7B En)	30B (15B Ko / 15B En)	60B (22B Ko / 38B En)	-	-
Pre-train Hardware	32× NVIDIA H100 80GB	32× NVIDIA H100 80GB	32× NVIDIA H100 80GB	-	-
Pre-train Duration	2 hours	9 hours	19 hours	-	-
Pre-train Learning Rate	7.5e-5	7.5e-5	7.5e-5	-	-
Pre-train Batch Size	2048	2048	2048	-	-
Pre-train Seq Length	512 → 2048	512 → 2048	512 → 2048	-	-
Pre-train AdamW Betas	$\beta = (0.9, 0.98)$	$\beta = (0.9, 0.98)$	$\beta = (0.9, 0.98)$	-	-
Pre-train AdamW Weight Decay	0.01	0.01	0.01	-	-
Fine-tuning Hardware	8× NVIDIA H100 80GB	8× NVIDIA H100 80GB	8× NVIDIA H100 80GB	8× NVIDIA H100 80GB	8× NVIDIA H100 80GB
Fine-tuning Learning Rate	5e-5	2e-5	2e-5	2e-5	2e-5
Fine-tuning Batch Size	32	32	32	32	32
Fine-tuning Seq Length	2048	2048	2048	2048	2048
Fine-tuning AdamW Betas	$\beta = (0.9, 0.98)$	$\beta = (0.9, 0.98)$	$\beta = (0.9, 0.98)$	$\beta = (0.9, 0.98)$	$\beta = (0.9, 0.98)$
Fine-tuning AdamW Weight Decay	0.01	0.01	0.01	0.01	0.01

Table E.1: Comparison of Thunder-DeID models and baseline Korean models, Exaone and Polyglot-ko.

$$\text{Binary Token-Level F1} = 2 \cdot \frac{\text{Binary Precision} \cdot \text{Binary Recall}}{\text{Binary Precision} + \text{Binary Recall}}$$

$$\text{Token-Level Micro F1} = 2 \cdot \frac{\text{Micro Precision} \cdot \text{Micro Recall}}{\text{Micro Precision} + \text{Micro Recall}}$$

where the positive class is any non-"Outside" label (e.g., name, phone number). Here, TP_{bin} is the number of tokens truly non-"Outside" and correctly predicted as non-"Outside", FP_{bin} is the number of tokens actually "Outside" but incorrectly predicted as non-"Outside", and FN_{bin} is the number of tokens truly non-"Outside" but incorrectly predicted as "Outside".

Token-Level Micro F1. Token-Level Micro F1 measures how well the model classifies tokens into specific entity types such as name of the person and phone numbers. It excludes the "Outside" label and calculates performance using aggregated precision and recall for each entity type. High scores indicate correct identification and labeling of tokens requiring de-identification, such as classifying "홍길동" (Hong Gildong) as a name of the person rather than a corporate entity.

Accurate classification of entity types is essential for proper de-identification of court judgments. This gets importance in the post-processing stage because without precise entity type prediction, the identified parts containing sensitive information cannot be properly replaced with contextually congruent phrases. Inaccurate classification can result in awkward or incorrect replacements in post-processing and ultimately lead to undermine the readability of the anonymized text.

For example, account numbers must be accurately identified and replaced with phrases like "계좌번호 1 생략" (Account number 1 omitted) during post-processing. If classified as different entity type such as a phone number, the misclassified account number might be incorrectly replaced with a phrase like "전화번호 1 생략" (Phone number 1 omitted). If the same case is classified as a business entity, it might be replaced with "A", and the post-processing result is not compatible with the current law and practice concerning the methods of anonymization (Judicial Rule No. 1778).

The token-level F1 score in our experiment is calculated as follows:

$$\text{Micro Precision} = \frac{\sum_{c \in C} \text{TP}_c}{\sum_{c \in C} (\text{TP}_c + \text{FP}_c)}$$

$$\text{Micro Recall} = \frac{\sum_{c \in C} \text{TP}_c}{\sum_{c \in C} (\text{TP}_c + \text{FN}_c)}$$

where C is the set of entity types (labels excluding the "Outside"), and for each entity type $c \in C$, TP_c, FP_c, and FN_c are the true positives, false positives, and false negatives respectively.

G Annotators

The authors participated in the annotation process for 20 hours per week over a period of 4 weeks. Seventeen external annotators contributed to the task for 12 hours per week over 4 weeks. These annotators were compensated at a rate of 10,000 KRW per hour, amounting to a total payment of 480,000 KRW per person. We consider this compensation appropriate given the local standards of living and the scope of the work.

H Issues in Prompt-based De-identification

We identify the following five categories of problems frequently appearing in the GPT-assisted de-identification dicussed in Section 1. These cases represent the ways in which prompting-based anonymization can lead to compromise textual integrity of public records and undermine legal precision required for settling disputes effectively.

- First, rewriting and paraphrasing frequently occurred. For example, the verb "입금하였다 (deposited)" was changed to "송금하였다 (wire transferred)." While both can describe sending money to someone, the forms and implications of these behaviors are differently conceived in legal and financial contexts.

- Second, we also found cases of partial omission when GPT removed, for instance, the phrase "제때 (on time)" from the original text. The original phrase "그 대금을 제때 변제하여" ("by repaying the amount on time") was shortened to "대금을 변제하여" ("by repaying the amount") in GPT-4's output. The omission of "제때" ("on time") removes an important indication of timely payment, which is often critical in determining whether the legal obligation was properly met.

- Third, (unsolicited) summarization of the original text resulted in the loss of detailed facts

and strategies concerning the crimes committed. Unlike the original text, it merely provides a brief summary of the factual backgrounds of the case. For instance, after going through GPT-assisted de-identification, three sentences containing important details about defendant's intention and plan to defraud victim and the amount of damage caused were vaguely summarized and reduced to a single sentence, "피고인은 이를 개인 용도로 사용하였다 (The defendant used it for personal purposes)".

- Fourth, in the cases where multiple individuals and institutions are involved in the litigation, we often identified entity collapse: a number of different entities were anonymized with the same letter (e.g., 광주은행 (Gwangju Bank), 우정사업본부 (Korea Post), 부산은행 (Busan Bank) → A, A, A).

- Lastly, distortion of facts occurred. For example, specific numbers in the judgment were altered during de-identification "총 3명 (a total of three people)" was altered to "총 명수1 (a total of one person)".

Moreover, due to privacy and information security concerns, the use of API-based LLM services such as ChatGPT is restricted in Korean government institutions. Domestic regulations (issued by the National Intelligence Service and the Ministry of the Interior and Safety) require public officials across government departments to refrain from putting in any sensitive internal data and personal information while using such services.

I Performance by Case Type

We report case-type precision, recall, binary token-level F1, and token-level micro F1 under two data regimes: Single Replacement and Per-Epoch Entity Replacement discussed in 4.2.

See the tables below for detailed results: binary token-level in Table I.1 (Single) and Table I.2 (Per-Epoch), and token-level in Table I.3 (Single) and Table I.4 (Per-Epoch). All tables report Precision and Recall; F1 is binary for binary token-level and micro-averaged for token-level. All values are averaged over three runs (seeds 1200, 1203, 1205) for each case type.

Domain	Case type	Model	Single Replacement (Binary Token-Level)		
			P	R	F1
Civil	Compensation for damage	Polyglot-ko (1.3B)	0.9843	0.9589	0.9714
		Exaone (2.4B)	0.9818	0.9462	0.9637
		Thunder-DeID-360M	0.9718	0.9303	0.9506
		Thunder-DeID-800M	0.9870	0.9774	0.9822
		Thunder-DeID-1.5B	0.9954	0.9663	0.9806
	Eviction	Polyglot-ko (1.3B)	0.9700	0.9336	0.9514
		Exaone (2.4B)	0.9681	0.9529	0.9604
		Thunder-DeID-360M	0.9672	0.9070	0.9361
		Thunder-DeID-800M	0.9763	0.9506	0.9632
		Thunder-DeID-1.5B	0.9709	0.9632	0.9671
	Purchase-price of a sale	Polyglot-ko (1.3B)	0.9727	0.9520	0.9623
		Exaone (2.4B)	0.9812	0.9594	0.9701
		Thunder-DeID-360M	0.9713	0.9148	0.9421
		Thunder-DeID-800M	0.9851	0.9628	0.9738
		Thunder-DeID-1.5B	0.9854	0.9600	0.9725
	Security deposit disputes	Polyglot-ko (1.3B)	0.9865	0.9726	0.9795
		Exaone (2.4B)	0.9826	0.9614	0.9719
		Thunder-DeID-360M	0.9816	0.9317	0.9559
		Thunder-DeID-800M	0.9865	0.9684	0.9773
		Thunder-DeID-1.5B	0.9881	0.9695	0.9787
Criminal	Bodily injury	Polyglot-ko (1.3B)	0.9826	0.9588	0.9706
		Exaone (2.4B)	0.9724	0.9718	0.9720
		Thunder-DeID-360M	0.9886	0.9623	0.9752
		Thunder-DeID-800M	0.9905	0.9800	0.9852
		Thunder-DeID-1.5B	0.9884	0.9806	0.9845
	Drunk driving	Polyglot-ko (1.3B)	0.9728	0.9508	0.9616
		Exaone (2.4B)	0.9831	0.9473	0.9649
		Thunder-DeID-360M	0.9714	0.9164	0.9430
		Thunder-DeID-800M	0.9733	0.9488	0.9608
		Thunder-DeID-1.5B	0.9817	0.9508	0.9660
	Fraud	Polyglot-ko (1.3B)	0.9775	0.9659	0.9717
		Exaone (2.4B)	0.9707	0.9601	0.9654
		Thunder-DeID-360M	0.9845	0.9418	0.9627
		Thunder-DeID-800M	0.9718	0.9806	0.9762
		Thunder-DeID-1.5B	0.9911	0.9766	0.9838
	Sexual misconduct	Polyglot-ko (1.3B)	0.9837	0.9690	0.9763
		Exaone (2.4B)	0.9837	0.9561	0.9697
		Thunder-DeID-360M	0.9803	0.9260	0.9524
		Thunder-DeID-800M	0.9872	0.9705	0.9788
		Thunder-DeID-1.5B	0.9881	0.9650	0.9764
	Violence	Polyglot-ko (1.3B)	0.9758	0.9644	0.9701
		Exaone (2.4B)	0.9702	0.9701	0.9701
		Thunder-DeID-360M	0.9664	0.9316	0.9486
		Thunder-DeID-800M	0.9749	0.9778	0.9763
		Thunder-DeID-1.5B	0.9740	0.9809	0.9774
Administrative	Administrative litigation	Polyglot-ko (1.3B)	0.9641	0.9321	0.9478
		Exaone (2.4B)	0.9743	0.9383	0.9559
		Thunder-DeID-360M	0.9814	0.9254	0.9526
		Thunder-DeID-800M	0.9877	0.9555	0.9713
		Thunder-DeID-1.5B	0.9842	0.9811	0.9827

Table I.1: Binary token-level metrics (Precision, Recall, and F1) for the **Single Replacement** setting, reported by case type and model (parameters shown in parentheses).

Domain	Case type	Model	Per-Epoch Replacement (Binary Token-Level)		
			P	R	F1
Civil	Compensation for damage	Polyglot-ko (1.3B)	0.9779	0.9687	0.9732
		Exaone (2.4B)	0.9770	0.9591	0.9679
		Thunder-DeID-360M	0.9611	0.9763	0.9686
		Thunder-DeID-800M	0.9796	0.9889	0.9842
		Thunder-DeID-1.5B	0.9796	0.9891	0.9843
	Eviction	Polyglot-ko (1.3B)	0.9644	0.9639	0.9641
		Exaone (2.4B)	0.9635	0.9597	0.9616
		Thunder-DeID-360M	0.9482	0.9566	0.9524
		Thunder-DeID-800M	0.9615	0.9711	0.9663
		Thunder-DeID-1.5B	0.9569	0.9803	0.9685
	Payment of purchase price	Polyglot-ko (1.3B)	0.9630	0.9650	0.9640
		Exaone (2.4B)	0.9679	0.9714	0.9696
		Thunder-DeID-360M	0.9463	0.9667	0.9564
		Thunder-DeID-800M	0.9712	0.9822	0.9766
		Thunder-DeID-1.5B	0.9748	0.9851	0.9799
	Security deposit disputes	Polyglot-ko (1.3B)	0.9770	0.9732	0.9751
		Exaone (2.4B)	0.9732	0.9736	0.9734
		Thunder-DeID-360M	0.9714	0.9661	0.9687
		Thunder-DeID-800M	0.9795	0.9864	0.9829
		Thunder-DeID-1.5B	0.9807	0.9878	0.9842
Criminal	Bodily injury	Polyglot-ko (1.3B)	0.9777	0.9746	0.9761
		Exaone (2.4B)	0.9697	0.9803	0.9749
		Thunder-DeID-360M	0.9811	0.9820	0.9815
		Thunder-DeID-800M	0.9875	0.9870	0.9872
		Thunder-DeID-1.5B	0.9868	0.9898	0.9883
	Drunk driving	Polyglot-ko (1.3B)	0.9667	0.9645	0.9656
		Exaone (2.4B)	0.9726	0.9685	0.9705
		Thunder-DeID-360M	0.9612	0.9572	0.9592
		Thunder-DeID-800M	0.9592	0.9795	0.9692
		Thunder-DeID-1.5B	0.9660	0.9739	0.9699
	Fraud	Polyglot-ko (1.3B)	0.9754	0.9776	0.9765
		Exaone (2.4B)	0.9651	0.9702	0.9676
		Thunder-DeID-360M	0.9739	0.9767	0.9753
		Thunder-DeID-800M	0.9843	0.9840	0.9841
		Thunder-DeID-1.5B	0.9850	0.9895	0.9873
	Sexual misconduct	Polyglot-ko (1.3B)	0.9788	0.9744	0.9766
		Exaone (2.4B)	0.9770	0.9638	0.9705
		Thunder-DeID-360M	0.9667	0.9698	0.9682
		Thunder-DeID-800M	0.9814	0.9840	0.9827
		Thunder-DeID-1.5B	0.9786	0.9851	0.9818
	Violence	Polyglot-ko (1.3B)	0.9679	0.9736	0.9707
		Exaone (2.4B)	0.9610	0.9754	0.9681
		Thunder-DeID-360M	0.9599	0.9745	0.9672
		Thunder-DeID-800M	0.9706	0.9874	0.9789
		Thunder-DeID-1.5B	0.9724	0.9942	0.9831
Administrative	Administrative litigation	Polyglot-ko (1.3B)	0.9603	0.9564	0.9583
		Exaone (2.4B)	0.9589	0.9605	0.9597
		Thunder-DeID-360M	0.9666	0.9623	0.9644
		Thunder-DeID-800M	0.9802	0.9814	0.9808
		Thunder-DeID-1.5B	0.9739	0.9898	0.9818

Table I.2: Binary token-level metrics (Precision, Recall, and F1) for the **Per-Epoch Replacement** setting, reported by case type and model (parameters shown in parentheses).

Domain	Case type	Model	Single Replacement (Token-Level)		
			P	R	Micro F1
Civil	Compensation for damage	Polyglot-ko (1.3B)	0.8793	0.8566	0.8677
		Exaone (2.4B)	0.8285	0.7988	0.8134
		Thunder-DeID-360M	0.7518	0.7195	0.7352
		Thunder-DeID-800M	0.7949	0.7872	0.7910
		Thunder-DeID-1.5B	0.8280	0.8037	0.8156
	Eviction	Polyglot-ko (1.3B)	0.8936	0.8602	0.8766
		Exaone (2.4B)	0.9108	0.8965	0.9036
		Thunder-DeID-360M	0.8963	0.8405	0.8675
		Thunder-DeID-800M	0.9234	0.8989	0.9109
		Thunder-DeID-1.5B	0.8985	0.8913	0.8949
	Payment of purchase price	Polyglot-ko (1.3B)	0.8386	0.8207	0.8296
		Exaone (2.4B)	0.8189	0.8000	0.8092
		Thunder-DeID-360M	0.8619	0.8118	0.8361
		Thunder-DeID-800M	0.9057	0.8854	0.8954
		Thunder-DeID-1.5B	0.9094	0.8859	0.8975
	Security deposit disputes	Polyglot-ko (1.3B)	0.8991	0.8864	0.8927
		Exaone (2.4B)	0.8959	0.8766	0.8861
		Thunder-DeID-360M	0.9312	0.8839	0.9069
		Thunder-DeID-800M	0.9411	0.9239	0.9324
		Thunder-DeID-1.5B	0.9440	0.9261	0.9349
Criminal	Bodily injury	Polyglot-ko (1.3B)	0.8852	0.8639	0.8744
		Exaone (2.4B)	0.8962	0.8956	0.8958
		Thunder-DeID-360M	0.9344	0.9096	0.9218
		Thunder-DeID-800M	0.9479	0.9378	0.9428
		Thunder-DeID-1.5B	0.9433	0.9360	0.9396
	Drunk driving	Polyglot-ko (1.3B)	0.8644	0.8448	0.8545
		Exaone (2.4B)	0.9047	0.8718	0.8879
		Thunder-DeID-360M	0.8784	0.8286	0.8527
		Thunder-DeID-800M	0.9097	0.8867	0.8980
		Thunder-DeID-1.5B	0.9078	0.8790	0.8931
	Fraud	Polyglot-ko (1.3B)	0.9040	0.8933	0.8987
		Exaone (2.4B)	0.9039	0.8940	0.8989
		Thunder-DeID-360M	0.9385	0.8978	0.9177
		Thunder-DeID-800M	0.9456	0.9286	0.9370
		Thunder-DeID-1.5B	0.9322	0.9186	0.9253
	Sexual misconduct	Polyglot-ko (1.3B)	0.8900	0.8767	0.8833
		Exaone (2.4B)	0.8505	0.8265	0.8383
		Thunder-DeID-360M	0.8879	0.8387	0.8626
		Thunder-DeID-800M	0.8958	0.8807	0.8882
		Thunder-DeID-1.5B	0.8941	0.8731	0.8834
	Violence	Polyglot-ko (1.3B)	0.8704	0.8602	0.8653
		Exaone (2.4B)	0.8828	0.8826	0.8827
		Thunder-DeID-360M	0.9036	0.8711	0.8871
		Thunder-DeID-800M	0.9203	0.9231	0.9217
		Thunder-DeID-1.5B	0.9138	0.9205	0.9171
Administrative	Administrative litigation	Polyglot-ko (1.3B)	0.8661	0.8373	0.8515
		Exaone (2.4B)	0.8999	0.8666	0.8829
		Thunder-DeID-360M	0.9246	0.8718	0.8974
		Thunder-DeID-800M	0.9481	0.9172	0.9324
		Thunder-DeID-1.5B	0.9363	0.9334	0.9349

Table I.3: Token-level metrics (Precision, Recall, and Micro F1) for the **Single Replacement** setting, reported by case type and model (parameters shown in parentheses).

Domain	Case type	Model	Per-Epoch Replacement (Token-Level)		
			P	R	Micro F1
Civil	Compensation for damage	Polyglot-ko (1.3B)	0.8774	0.8688	0.8730
		Exaone (2.4B)	0.8525	0.8372	0.8448
		Thunder-DeID-360M	0.7435	0.7553	0.7493
		Thunder-DeID-800M	0.8121	0.8197	0.8159
		Thunder-DeID-1.5B	0.8141	0.8220	0.8179
	Eviction	Polyglot-ko (1.3B)	0.8878	0.8874	0.8875
		Exaone (2.4B)	0.9007	0.8971	0.8989
		Thunder-DeID-360M	0.8939	0.9019	0.8979
		Thunder-DeID-800M	0.9035	0.9125	0.9080
		Thunder-DeID-1.5B	0.8967	0.9186	0.9075
	Payment of purchase price	Polyglot-ko (1.3B)	0.8322	0.8339	0.8330
		Exaone (2.4B)	0.8460	0.8490	0.8474
		Thunder-DeID-360M	0.8646	0.8833	0.8738
		Thunder-DeID-800M	0.8902	0.9002	0.8952
		Thunder-DeID-1.5B	0.8933	0.9029	0.8981
	Security deposit disputes	Polyglot-ko (1.3B)	0.8958	0.8923	0.8940
		Exaone (2.4B)	0.8833	0.8838	0.8835
		Thunder-DeID-360M	0.9268	0.9218	0.9243
		Thunder-DeID-800M	0.9430	0.9497	0.9463
		Thunder-DeID-1.5B	0.9469	0.9538	0.9503
Criminal	Bodily injury	Polyglot-ko (1.3B)	0.8792	0.8765	0.8778
		Exaone (2.4B)	0.8857	0.8953	0.8904
		Thunder-DeID-360M	0.9306	0.9314	0.9310
		Thunder-DeID-800M	0.9519	0.9515	0.9517
		Thunder-DeID-1.5B	0.9518	0.9546	0.9532
	Drunk driving	Polyglot-ko (1.3B)	0.8697	0.8678	0.8688
		Exaone (2.4B)	0.8932	0.8894	0.8913
		Thunder-DeID-360M	0.8869	0.8832	0.8851
		Thunder-DeID-800M	0.9045	0.9238	0.9140
		Thunder-DeID-1.5B	0.9158	0.9231	0.9194
	Fraud	Polyglot-ko (1.3B)	0.9097	0.9117	0.9107
		Exaone (2.4B)	0.8964	0.9010	0.8987
		Thunder-DeID-360M	0.9212	0.9238	0.9225
		Thunder-DeID-800M	0.9399	0.9396	0.9397
		Thunder-DeID-1.5B	0.9204	0.9246	0.9225
	Sexual misconduct	Polyglot-ko (1.3B)	0.8799	0.8759	0.877
		Exaone (2.4B)	0.8556	0.8441	0.8498
		Thunder-DeID-360M	0.8888	0.8916	0.8902
		Thunder-DeID-800M	0.9021	0.9045	0.9033
		Thunder-DeID-1.5B	0.8768	0.8827	0.8797
	Violence	Polyglot-ko (1.3B)	0.8617	0.8669	0.8643
		Exaone (2.4B)	0.8679	0.8810	0.8744
		Thunder-DeID-360M	0.8981	0.9118	0.9049
		Thunder-DeID-800M	0.9043	0.9200	0.9120
		Thunder-DeID-1.5B	0.9209	0.9416	0.9311
Administrative	Administrative litigation	Polyglot-ko (1.3B)	0.8691	0.8654	0.8672
		Exaone (2.4B)	0.8913	0.8928	0.8920
		Thunder-DeID-360M	0.9138	0.9097	0.9118
		Thunder-DeID-800M	0.9372	0.9384	0.9377
		Thunder-DeID-1.5B	0.9297	0.9448	0.9372

Table I.4: Token-level metrics (Precision, Recall, and Micro F1) for the **Per-Epoch Replacement** setting, reported by case type and model (parameters shown in parentheses).