A Appendix

A.1 Related Works

In this section, we compare LBox Open to previous studies. In general, LBox Open differs to other works in that it exclusively treats Korean legal cases especially from the lower courts.

Chalkidis et al. (2019) introduces the ECtHR dataset that consists of 11k cases from the European Court of Human Rights. In this task, a model needs to predict a set of related articles for the given facts. The Statute dataset also consists of the facts and corresponding statues yet it handles criminal cases from the lower court with more diverse legal domain (5th column in Table 2).

Niklaus et al. (2021) releases the Swiss-Judgements-Prediction dataset that consists of 85k multilingual cases–German, French, and Italian–from the Federal Supreme Court of Switzerland. In this task, a model needs to predict "approval" or "dismissal" regrading the validity of the plaintiff's request for the given facts of the cases. Similarly, in LJP-civil task, a model predicts the claim acceptance degree for the given facts and the gist of claim. However, the dataset differs in that the acceptance degree is quantitatively estimated based on the ratio between the claimed money from the plaintiffs and the approved money by the judges. Also, LJP-civil handles cases from the 1st trials from the district courts of Korea.

Xiao et al. (2018) introduces the CAIL dataset which consists of 2.7m Chinese criminal cases. Given the facts, a model needs to predict corresponding law articles, charges, and prison terms. The dataset is similar to Statute (law articles), Case name (∼charge), and LJP-criminal (∼prison terms) datasets with following differences; (1) both Statute, and Case name include the civil cases; (2) in LJP-criminal task, a model needs to predict three types of punishments, fine, imprisonment with labor, and imprisonment without labor. Additionally, LJP-criminal includes "the reason for sentencing" section in which the judges describe various aspects that can affect the final results such as upper and lower bounds of punishment ranges described in the related statutes, ages and attitude of defendants, victim's opinions etc.

Ma et al. (2021) studies legal judgement prediction over low level data by using plaintiff's claims and court debate as an input. On the contrary, LJP-civil and LJP-criminal relies on the factual description written by judges in precedents. The court debates are not publicly available in Korea.

Chalkidis et al. (2022a) introduces a benchmark dataset for legal NLU in English focusing on classification tasks over various legal documents; cases, legislation, contracts, Terms of Service, and holdings in cases. On the other hand, LBox Open present one legal corpus, two classification tasks, two legal judgement tasks (may be considered as classification or regression task in a broader sense), and one summarization task using processed (by ourselves) Korean precedents exclusively.

Chalkidis et al. (2022b) investigate legal fairness over four legal judgement datasets with additional attributes such as race, gender, region, language, age etc. As LBox Open handles anonymized Korean cases exclusively, there is almost no data heterogeneity is terms of race, language, and regions and thus the fairness is not investigated as a main topic in this study. On the other hand, we find in LJP-criminal dataset, certain case category such as "indecent act by compulsion (강제추행)" shows gender bias. The possible risk and the limitation of LBox Open is discussed in Ethical Considerations.

A.2 Precedent redaction rule

Data subjected to anonymization are as follows[18].

- Name and the equivalents: Name, nickname, pen name, ID, and corresponding nouns that point to a specific person are replaced with upper case alphbets A, B, C, etc., without redundancy.
- Contact information: Phone number, e-mail address, residential address and corresponding contact data are deleted in the meta information(당사자 단락) and replaced with upper case alphabets in the reasoning(이유 단락)

[18] `https://www.scourt.go.kr/portal/information/finalruling/anony/index.html`, `https://glaw.scourt.go.kr/wsjo/gchick/sjo330.do?contId=3202812#1660965241959`

- Financial information: Account number, credit card number, check number and corresponding financial data are replaced with upper case alphabets without redundancy.
- Other personally identifible information: Social security number is deleted. car registration number, address of real estate holdings and corresponding personal information are replaced with upper case alphabets without redundancy.

In the case of felonies such as serial killing, there are some exceptions where the Korean government can decide not to anonymize for the social benefits[19].

A.3 Precedent disclosure status

Opening the precedents public meets the constitutional need to guarantee the right to know and right to a public trial, while allowing the general public to see that the justice system is functioning properly and treating defendants fairly. On the other hand, non-disclosure of certain documents and anonymization are to compromise the stated benefits with one's privacy and honor.

Since the Civil Procedure Act was revised in the December of 2020[20], all civil court decisions including those not settled and appealed to higher courts will be open to general public from 01.01.2023. In the same context, expanding the scope of disclosure while minimizing the side effects is being discussed in the National Assembly of Korea.

A.4 The example of Korean precedent

- Meta information
  - Plaintiffs: A 주식회사
  - Defendants: 주식회사 B (전동 킥보드 제조 회사)
  - Case name: 구상금
- Gist of claim: [청구취지] 피고는 원고에게 95,569,454원 및 이에 대하여 2020. 1. 15.부터이 사건 소장 부본 송달일까지는 연 5%, 그 다음날부터 각 다 갚는 날까지는 연 12%의 각 비율로 계산한 돈을 지급하라.
- Facts: [사실관계] ○ 원고는 손해보험업을 영위하는 보험회사로서, 주식회사 C와 생산물배상책임보험(이하 '이 사건 보험계약'이라 한다)을 체결한 보험자이고, 피고는 D 전동킥보드를 제조하여 주식회사 C에 납품하는 업체이다.
  ○ E은 (중략) 전동킥보드를 구입하였고, (중략) 개조를 하였다.
  ○ (중략) 이 사건 전동킥보드를 충전하던 중 화재가 발생하였고 (중략) ...
- Claim of plaintiffs: [원고의 주장] 피고가 제조한 이 사건 전동킥보드 (중략) 전기적 결함으로 인하여 이 사건 화재가 발생한 것이므로, 피고는 (중략) 원고에게 그 보험금 상당액을 지급할 의무가 있다.
- Reasoning [판사의 판단] (중략)제조물 결함으로 인한 배상책임이 인정되기 위하여는, (중략) 해당 제조물의 결함 없이는 (중략) 발생하지 아니한다는 사실이 (중략) 증명되어야 한다. 그런데 (중략) 전동킥보드를 개조함으로써 그 개조 과정의 하자 등으로 인하여 발생한 것일 가능성이 있다는 점에서 이 사건 전동킥보드 배터리 등 자체의 결함으로 인하여 발생한 것이라고 단정하기 어렵고 달리 이를 인정할 만한 충분한 증거가 없다 (중략)
- Ruling: [판사의 결론] 원고의 청구를 기각한다.

A.5 LBox Open examples

A.5.1 Precedent corpus

- Ruling: 주문 피고인을 징역 6개월에 처한다.다만, 이 판결 확정일로부터 1년간 위 형의 집행을 유예한다.

[19] for example, see `https://en.wikipedia.org/wiki/Yoo_Young-chul`, `https://www.law.go.kr/판례/(2004고합972)`

[20] `https://law.go.kr/lsInfoP.do?lsiSeq=223439&lsId=&efYd=20230101&chrClsCd=010202&urlMode=lsEfInfoR&viewCls=lsRvsDocInfoR&ancYnChk=0#`

- Reason: 이유 범 죄 사 실 1. 사기 피고인은 2020. 12. 15. 16:00경 경북 칠곡군 B에 있는 피해자 C이 운영하는 'D'에서, 마치 정상적으로 대금을 지급할 것처럼 행세하면서 피해자에게 술을 주문하였다. 그러나 사실 피고인은 수중에 충분한 현금이나 신용카드 등 결제 수단을 가지고 있지 않아 정상적으로 대금을 지급할 의사나 능력이 없었다. 그럼에도 피고인은 위와 같이 피해자를 기망하여 이에 속은 피해자로부터 즉석에서 합계 8,000원 상당의 술을 교부받았다. 2. 공무집행방해 피고인은 제1항 기재 일시 · 장소에서, '손님이 술값을 지불하지 않고 있다'는 내용의 112신고를 접수하고 현장에 출동한 칠곡경찰서 E지구대 소속 경찰관 F로부터 술값을 지불하고 귀가할 것을 권유받자, "징역가고 싶은데 무전취식했으니 유치장에 넣어 달라"고 말하면서 순찰차에 타려고 하였다. 이에 경찰관들이 수회 귀가 할 것을 재차 종용하였으나, 피고인은 경찰관들을 향해 "내가 돌로 순찰차를 찍으면 징역갑니까?, 내여경 엉덩이 발로 차면 들어갈 수 있나?"라고 말하고, 이를 제지하는 F의 가슴을 팔꿈치로 수회 밀쳐 폭행하였다. 이로써 피고인은 경찰관의 112신고사건 처리에 관한 정당한 직무집행을 방해하였다. 증거의 요지 1. 피고인의 판시 제1의 사실에 부합하는 법정진술 1. 증인 G, F에 대한 각 증인신문조서 1. 영수증 1. 현장사진 법령의 적용 1. 범죄사실에 대한 해당법조 및 형의 선택 형법 제347조 제1항, 제136조 제1항, 각 징역형 선택 1. 경합범가중 형법 제37조 전단, 제38조 제1항 제2호, 제50조 1. 집행유예 형법 제62조 제1항 양형의 이유 1. 법률상 처단형의 범위: 징역 1월∼15년 2. 양형기준에 따른 권고형의 범위 가. 제1범죄(사기) [유형의 결정] 사기범죄 > 01. 일반사기 > [제1유형] 1억 원 미만 [특별양형인자] - 감경요소: 미필적 고의로 기망행위를 저지른 경우 또는 기망행위의 정도가 약한 경우, 처벌불원 [권고영역 및 권고형의 범위] 특별감경영역, 징역 1월∼1년 [일반양형인자] 없음 나. 제2범죄(공무집행방해) [유형의 결정] 공무집행방해범죄 > 01. 공무집행방해 > [제1유형] 공무집행방해/직무강요 [특별양형인자] - 감경요소: 폭행 · 협박 · 위계의 정도가 경미한 경우 [권고영역 및 권고형의 범위] 감경영역, 징역 1월∼8월 [일반양형인자] - 감경요소: 심신미약(본인 책임 있음) 다. 다수범죄 처리기준에 따른 권고형의 범위: 징역 1월∼1년4월(제1범죄 상한 + 제2범죄 상한의 1/2) 3. 선고형의 결정: 징역 6월에 집행유예 1년 만취상태에서 식당에서 소란을 피웠고, 112신고로 출동한 경찰관이 여러 차례 귀가를 종용하였음에도 이를 거부하고 경찰관의 가슴을 밀친 점 등을 종합하면 죄책을 가볍게 볼 수 없으므로 징역형을 선택하되, 평소 주량보다 훨씬 많은 술을 마신 탓에 제정신을 가누지 못해 저지른 범행으로 보이고 폭행 정도가 매우 경미한 점, 피고인이 술이 깬 후 자신의 경솔한 언동을 깊이 반성하면서 재범하지 않기 위해 정신건강의학과의 치료 및 상담을 받고 있는 점, 식당 업주에게 피해를 변상하여 용서를 받은 점, 피고인의 나이와 가족관계 등의 사정을 참작하여 형의 집행을 유예하고, 범행 경위와 범행 후 피고인의 태도 등에 비추어 볼 때 재범의 위험성은 그다지 우려하지 않아도 될 것으로 보여 보호관찰 등 부수처분은 부과하지 않음. 이상의 이유로 주문과 같이 판결한다.

A.5.2 CASE NAME

- Facts (input): 질병관리청장, 시 · 도지사 또는 시장 · 군수 · 구청장은 제1급 감염병이 발생한 경우 감염병의 전파방지 및 예방을 위하여 감염병의심자를 적당한 장소에 일정한 기간 격리시키는 조치를 하여야 하고, 그 격리조치를 받은 사람은 이를 위반하여서는 아니 된다. 피고인은 해외에서 국내로 입국하였음을 이유로 2021. 4. 21.경 감염병의심자로 분류되었고, 같은 날 창녕군수로부터 '2021. 4. 21.부터 2021. 5. 5. 12:00경까지 피고인의 주거지인 경남 창녕군 B에서 격리해야 한다'는 내용의 자가격리 통지서를 수령하였다. 1. 2021. 4. 27.자 범행 그럼에도 불구하고 피고인은 2021. 4. 27. 11:20경에서 같은 날 11:59경까지 사이에 위 격리장소를 무단으로 이탈하여 자신의 승용차를 이용하여 경남 창녕군 C에 있는 'D' 식당에 다녀오는 등 자가격리 조치를 위반하였다. 2. 2021. 5. 3.자 범행 피고인은 2021. 5. 3. 10:00경에서 같은 날 11:35경까지 사이에 위 격리장소를 무단으로 이탈하여 자신의 승용차를 이용하여 불상의 장소를 다녀오는 등 자가격리 조치를 위반하였다.
- Case name (ground truth): 감염병의예방및관리에관한법률위반

A.5.3 STATUTE

- Facts (input): 1. 사문서위조 피고인은 2014. 5. 10.경 서울 송파구 또는 하남시 이하 알 수 없는 장소에서 영수증문구용지에 검정색 볼펜을 사용하여 수신인란에 'A', 일금란에 '오천오백육십만원정', 내역 란에 '2010가합7485사건의 합의금 및 피해 보상금 완결조', 발행일란에 '2014년 5월 10일'이라고 기재한 뒤, 발행인 옆에 피고인이 임의로 만들었던 B의 도장을 찍었다. 이로써 피고인은 행사할 목적으로 사실증명에 관한 사문서인 B 명

의의 영수증 1장을 위조하였다. 2. 위조사문서행사 피고인은 2014. 10. 16.경 하남시 이하 알 수 없는 장소에서 피고인이 B에 대한 채무를 모두 변제하였기 때문에 B가 C회사에 채권을 양도한 것을 인정할 수 없다는 취지의 내용증명원과 함께 위와 같이 위조한 영수증 사본을 마치 진정하게 성립한 문서인 것처럼 B에게 우편으로 보냈다. 이로써 피고인은 위조한 사문서를 행사하였다.

- Statutes (ground truth): 형법 제231조, 형법 제234조

A.5.4 LJP-CRIMINAL

- Facts (input): 피고인은 2016.4. 6.경 지인으로부터 피해자 B을 소개받아 알게 된 사이이다. 피고인은 2016. 4. 8.경 장소불상지에서 피해자에게 전화하여 '일수 빚을 갚아야 하니 돈을 빌려주면 반드시 변제하겠다'고 거짓말하였다. 그러나 피고인은 특별한 재산이 없고 1,400만 원 이상 다액의 일수 및 대출금 채무가 있었던 상황으로 피해자로부터 금원을 빌리더라도 이를 변제할 의사나 능력이 없었다. 그럼에도 불구하고 피고인은 위와 같이 피해자를 기망하여 이에 속은 피해자로부터 2016. 4. 8. 경 부산 부산진구 C에 있는 D은행 가야동지점 인근에서 현금 200만 원을 교부받은 것을 비롯하여, 그 무렵부터 2016. 10. 11. 경까지 별지 범죄일람표 기재와 같이 피해자로부터 총 17회에 걸쳐 차용금 명목으로 2,435만 원을 교부받았다.
- Punishment (ground truth): 징역 3월
- Punishment (quantized ground truth): 벌금0, 징역1, 금고0
- Reason for sentencing (optional input): 양형의 이유. 1. 양형기준에 따른 권고형의 범위 [유형의 결정] 사기범죄 > 01. 일반사기 > [제1유형] 1억 원 미만 [특별양형인자] 감경요소: 처벌불원 또는 상당부분 피해 회복된 경우 [권고영역 및 권고형의 범위] 감경영역, 징역 1월 1년 2. 선고형의 결정: 피고인이 진지하게 반성하는 점, 피해회복 후 피해자와 원만히 합의한 점, 별다른 처벌전력이 없는 점, 그밖에 피고인의 연령, 성행, 환경, 범죄 후의 정황 등을 고려해서 주문과 같이 형을 정한다.
- Ruling: 피고인을 징역 3월에 처한다. 다만 이 판결 확정일로부터 1년간 위 형의 집행을 유예한다.

A.5.5 LJP-CIVIL

- Facts (input 1): 가. 원고는 (차량번호 1 생략) 포터 화물차(이하 '원고 차량'이라 한다)에 관하여 그 소유자인 B와 자동차종합보험계약을 체결한 보험자고, 피고는 사고 장소인 도로의 관리주체다. 나. B는 2019. 12. 13. 15:40경 원고 차량을 운전하여 경북 성주군 금수면 명천리 도로 부근에서 전방 및 좌우를 살펴 조향 및 제동장치를 정확하게 조작하며 안전하게 운전하여야 할 주의의무를 위반하여 원고 차량 기준 우측에서 좌측으로 진행하는 오토바 이를 충격하게 되었다(이하 '이 사건 사고'라 한다). 다. 이 사건 사고로 인하여 오토바이 운전자와 동승자가(이하 '망인들'이라 한다) 사망하고, 오토바이가 파손되었으며, 원고는 오토바이 망인들의 사망보험금 및 오토바이 수리비로 합계 333,931,050원을 지급하였다.
- Gist of claim (input 2) 청구취지. 피고는 원고에게 66,786,210원과 이에 대하여 2020. 3. 19.부터 이 사건 소장 부본 송달일까지는 연 5%의, 그 다음날부터 다 갚는 날까지는 연 12%의 각 비율로 계산한 돈을 지급하라.
- The degrees of claim acceptance level (ground truth): 0.5 ($= 33393105/66786210$)
- The degrees of claim acceptance level (quantized ground truth): 부분 인정
- Ruling: 1. 피고는 원고에게 33,393,105원과 이에 대하여 2020. 3. 19.부터 2022. 1. 13. 까지는 연 5%의, 그 다음날부터 다 갚는 날까지는 연 12%의 각 비율로 계산한 돈을 지급하라. 2. 원고의 나머지 청구를 기각한다. 3. 소송비용은 50%는 원고가, 나머지는 피고가 각 부담한다. 4. 제1항은 가집행할 수 있다.

A.6 Data engineering pipeline

Here we provide the additional details of the precedent engineering pipeline. Layout-classifier was prepared by training custom ResNet using 300k training examples consisting of document image and label pairs. Layout-parser is based on Mask-R-CNN and was trained on 162k examples consisting of document image and coordinates of the figures/tables pairs using detectron2 library (Wu et al., 2019).

The initial checkpoint was prepared by loading pre-trained model on PubLayNet dataset (Zhong et al., 2019). For Post-OCR Text corrector, we use character level transformers trained on 137 k examples that consist of OCR outputs and manually corrected text pairs. The initial checkpoint was loaded from KorCharElectra[21]. The transformer generates tags indicating whether to "keep", "delete", "replace", or "insert" for individual characters. When the generated tags are either "replace" or "insert", the transformer generates an additional new character. All training sets were prepared via our own data labeling platform LWorks (https://lworks.kr/work) where we employ ∼100 part-time annotators.

The precision and recalls of individual ML modules on the static test sets are as follows; Layout-classifiers, P 99.94%, R 82.0%; Layout-parser, P 98.6%, R 97.1% with IOU threshold 0.5; Post OCR Corrector, P 91.1%, R 74.6%.

The overall automation efficiency on the deployed system is ∼80% for PDF-type precedents and ∼53% for non-readable type precedents. The remaining precedents are manually monitored and corrected via the LWorks platform. We are currently preparing a separate technical report explaining the further details of the method.

A.7 Data statistics

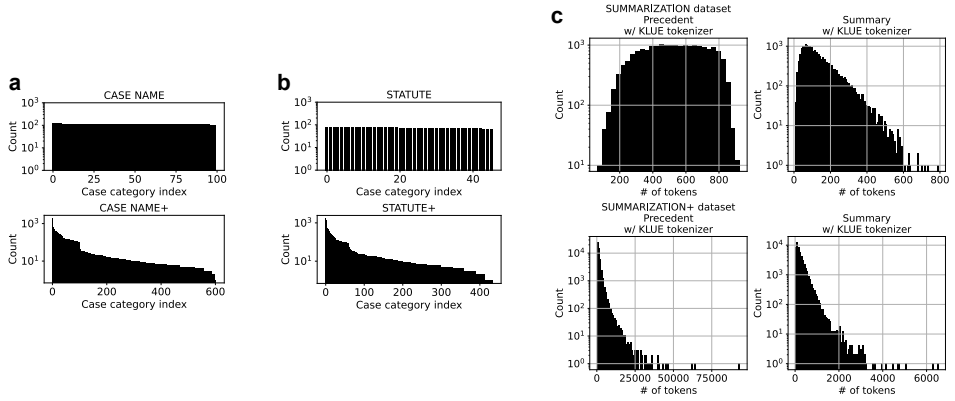

Figure 2: Data distribution histograms.

A.7.1 LJP-CRIMINAL statistics

Table 6: Statistics of LJP-CRIMINAL dataset

Case category	fine	imprisonment w/ labor	imprisonment w/o labor
"indecent act by compulsion" (강제추행)	857	756	0
"obstruction of performance of official duties" (공무집행방해)	421	1,210	0
"bodily injuries from traffic accident" (교통사고처리특례법위반(치상))	656	2	955
"drunk driving" (도로교통법위반(음주운전))	228	1,414	0
"fraud" (사기)	300	1,311	0
"inflicting bodily injuries" (상해)	792	811	0
"violence" (폭행)	1,286	276	0
total	4,540	5,780	955

A.8 The domain quantization scheme of two LJP tasks

A.9 Pretraining

A.10 Author's Statement

We guarantee that LBOX OPEN and LCUBE are released under Attribution-NonCommercial-NoDerivatives 4.0 lincese and publicly available for non-commercial use. We also use only the officially anonymyzed precedents issued by the Korean governments during the construction of the datasets to avoid possible confidentiality issues. LBOX OPEN, LCUBE, and the code to reproduce baseline results are available from `https://github.com/lbox-kr/lbox-open`.

[21] `https://github.com/monologg/KoCharELECTRA`

Table 7: Class labels for legal judgement prediction tasks.

Tasks	# of classes	subtask	unit	ranges	label	number
LJP-CRIMINAL-Lv0	2	fine	KRW 1000	0 (null label)	0	6,888
				$0 < \cdot$	1	4,540
LJP-CRIMINAL-Lv0	2	imprisonment with labor,	month	0 (null label)	0	16,121
	2	imprisonment without labor		$0 < \cdot$	1	6,735
LJP-CRIMINAL-Lv1	3	fine	KRW 1000	0	0	6,888
				$0 < \cdot \leq 2000$	1	2,281
				$2000 < \cdot$	2	2,259
LJP-CRIMINAL-Lv1	3	imprisonment with labor,	month	0	0	16,121
	3	imprisonment without labor		$0 < \cdot \leq 6$	1	3,219
				$6 < \cdot$	2	3,516
LJP-CRIMINAL-Lv2	5	fine	KRW 1000	0	0	6,888
				$0 < \cdot < 1000$	1	982
				$1000 \leq \cdot < 3000$	2	1,344
				$3000 \leq \cdot < 10000$	3	2,046
				$10,000 \leq \cdot$	4	168
LJP-CRIMINAL-Lv2	6	imprisonment with labor,	month	0	0	16,121
	6	imprisonment without labor		$0 < \cdot < 6$	1	942
				$6 \leq \cdot < 12$	2	3,769
				$12 \leq \cdot < 18$	3	1,615
				$18 \leq \cdot < 60$	4	408
				$60 \leq \cdot$	5	1
LJP-CRIMINAL-Lv3	∞	fine	KRW 1000	-	-	
LJP-CRIMINAL-Lv3	∞	imprisonment with labor, imprisonment without labor	month	-	-	
LJP-CIVIL-Lv1	3	the degrees of claim acceptance (=approved money / claimed money)	-	0	0	2,862
				$0 < \cdot < 1$	1	2,132
				1	2	87
LJP-CIVIL-Lv2	13	the degrees of claim acceptance (=approved money / claimed money)	-	$0 \leq \cdot < 0.05$	0	2,935
				$0.05 \leq \cdot < 0.15$	0.1	183
				$0.15 \leq \cdot < 0.25$	0.2	227
				$0.25 \leq \cdot < 0.35$	0.3	250
				$0.35 \leq \cdot < 0.45$	0.4	245
				$0.45 \leq \cdot < 0.55$	0.5	294
				$0.55 \leq \cdot < 0.65$	0.6	213
				$0.65 \leq \cdot < 0.75$	0.7	229
				$0.75 \leq \cdot < 0.85$	0.8	194
				$0.85 \leq \cdot < 0.95$	0.9	160
				$0.95 \leq \cdot < 1.05$ †	1	143
				$1.05 \leq \cdot < 1.15$ †	1.1	7
				$1.15 \leq \cdot < 1.25$ †	1.2	1

†: The ratio of approved money / claimed money can be larger than one when the interest of money is involved.

Table 8: Statistics of the pretraining corpora

Domain	Pretraining Corpus	# Tokens	Size (GB)
Wiki	Wikipedia	0.166B	1.49
New, Book	Modu	4.330B	37.18
Legal	LBox-Open	0.263B	2.1

Table 9: The model configuration and the pre-training parameters of LCUBE

Name	n_{params}	n_{layers}	d_{model}	n_{head}	d_{head}	Batch Size	Learning Rate
LCUBE-base	124M	12	768	12	64	512	6.0×10^{-4}
LCUBE-medium	354M	24	1024	16	64	512	6.0×10^{-4}

A.11 Datasheet for the dataset

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The datasets were created to stimulate research on natural legal language understanding tasks and to develop the real world machine learning applications in legal domain. The number of avail-

Table 10: Comparison of various models on LJP-CRIMINAL task under regression setting.

Name	size	regression (F_1)		
		fine	imp. w/ labor	imp. w/o labor
KoGPT-2-base	125M	19.2	51.6	36.6
LCUBE-base-zero	124M	21.6	51.5	36.0
mt5-small	300M	22.3	51.8	39.7
mt5-large (512 g)	1.2B	22.1	52.1	36.5
KoGPT-2-base + d.a.	125M	23.7	51.5	36.6
LCUBE-base	124M	20.0	49.9	39.1
LCUBE-base + d.a.	124M	25.8	52.0	40.0
LCUBE-medium	354M	21.8	52.1	36.0
LCUBE-medium + d.a.	354M	24.7	53.5	34.3
mt5-small + d.a.	300M	25.4	53.3	36.6
Most frequent label	-	12.7	27.9	5.4
Null ratio	-	0.605	0.486	0.916
Top non-null label ratio	-	0.068	0.162	0.028
Top non-null label ratio (w/o null)	-	0.172	0.314	0.333

† The reason for the sentencing.

able datasets in legal domain is small especially for non-English language. Thus, we release the first large-scale benchmark of Korean legal AI datasets.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The dataset is created by the authors of this papers from LBox Co. Ltd. and KAIST.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

LBox Co. Ltd.

Any other comments? None.

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The instances represent individual Korean precedents structured via our custom data engineering pipeline (Fig. 1.

How many instances are there in total (of each type, if appropriate)?

The dataset statistics are shown in Table 1 and 2.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The datasets are the sub set of entire Korean precedents whose their numbers are about few thousands million[22]. As social phenomena and related legal aspects are so diverge, it is difficult to construct the representative sub set. Thus, in CASE NAME, STATUTE, LJP-CRIMINAL, and LJP-CIVIL datasets, we used only the precedents (1) from 1st trials and (2) from limited case categories to construct the datasets. Under these constraints, we randomly sampled the precedents and the resulting datasets may approximately represent the parent distribution.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

The instances of LBOX OPEN consist of texts extracted from the Korean precedents except the labels of LJP-CRIMINAL and LJP-CIVIL. The labels are generated by parsing and quantizing the numbers from the raw text that include either amount of claimed money, fine amount, or imprisonment period.

Is there a label or target associated with each instance? If so, please provide a description.

[22] https://www.scourt.go.kr/portal/justicesta/JusticestaListAction.work?gubun=10

Yes. See previous question. Also, see Table 2 and Appendix A.5.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

No.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

No.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

Yes. See Table 2.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

To make LJP-CRIMINAL, we parsed the ruling to extract the fine amount and the imprisonment period using the custom language model. Although we filtered out the results with the confidence score from the language model and with a set of heuristically rules, the resulting parses could include imprecise numbers. Similarly, in LJP-CIVIL, the parsed monies from the gist of claim and the ruling could include noises. For the quality assurance, we manually checked around 40 examples from each dataset but found no errors.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The datasets are self-contained.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description.

No. Confidential information is anonymized by Korean government except some special cases where the governments decide to reveal for the social benefit.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

The instances from the criminal cases may include the detailed description of crimes. Please see "Ethical Consideration" section in the main paper.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Yes.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

The genders and ages may be implicitly identifiable from the facts that are included in the instances. But as they are written in free-form text, they are not easily identified without developing the dedicated parser.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

No. Such information is anonymized by Korean government except some special cases where the government decide to reveal for the social benefit.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social se-

curity numbers; criminal history)? If so, please provide a description.

The instances from the criminal cases may include the detailed description of crimes, which can include sensitive data, and criminal history but the individuals are anonymized. Please see "Ethical Consideration" section in the main paper.

Any other comments? None.

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

All data are directly observable except the claim acceptance level in LJP-CIVIL and the fine and imprisonment levels in LJP-CRIMINAL. For these cases, we include corresponding raw data, the gist of claim and the ruling.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

The detailed procedure about how the dataset was constructed is described in Section 3.1 and Fig. 1.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

We selected precedents from most frequent case categories. The detailed information is described in Section 3.2.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

The employees of LBox worked together.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The dataset was created by buildling dedicated data engineering pipeline (Fig. 1) and took around an year. Without considering the pipeline building time, the process took around six monthes.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

The precedents are issued from the Korean courts and used to fill the empty space of the laws. Also, as they are mostly anonymized except some special cases, we did not take additional ethical review processes.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Yes.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

We used the precedents issued by the Korean governments.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

No (See previous question).

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

No (see previous question).

If consent was obtained, were the consenting individuals provided with a mech-

anism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

N/A.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A.

Any other comments?

None.

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

Yes. Please refer to Section 3.2.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

When the raw data is required to reconstruct the processed data, we include them (for instance, in LJP-CIVIL and LJP-CRIMINAL, we include the raw text from which the labels were generated).

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

No.

Any other comments?

None.

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

No.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

No.

What (other) tasks could the dataset be used for?

The datasets can be used for (1) to train legal language models, (2) to train and benchmark machine learning models on natural legal language understanding tasks.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

The facts in the datasets can include personal information such as genders, ages, and regions. This information can bias the machine learning model trained using the datasets Chalkidis et al. (2022b). The users should be aware of this general properties of any data-driven approach and be cautious in interpreting the prediction of the model especially in the legal judgement prediction tasks.

Are there tasks for which the dataset should not be used? If so, please provide a description.

The datasets only partially cover various social phenomena and their related legal aspects. Because of this limitation and the reason explained in previous question, the trained models can generate imprecise and biased outputs. Thus the datasets should be used for academic purpose only.

Any other comments? None.

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf

of which the dataset was created? If so, please provide a description.

Yes, the datasets are publicly available.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub) Does the dataset have a digital object identifier (DOI)?

The datasets LBox Open and pre-trained legal language model LCube can be downloaded from HuggingFace hub (LBox Open: `https://huggingface.co/datasets/lbox/lbox_open`, LCube: `https://huggingface.co/lbox/lcube-base/tree/main`).

When will the dataset be distributed?

The datasets are currently available.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The dataset are distributed under `Attribution-NonCommercial-NoDerivatives 4.0` license.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No.

Any other comments?

None.

Maintenance

Who will be supporting/hosting/maintaining the dataset?

Wonseok Hwang, Dongjun Lee, and Kyoungyeon Cho will be maintaining/supporting the datasets.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

Via emails.

Is there an erratum? If so, please provide a link or other access point.

None found yet.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

Yes. The possible labeling errors, additional examples, and new tasks will be updated and announced via the LBox Open GitHub repository. The update interval is expected to be few months.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

No.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

Yes. To maintain older version, we use HuggingFace hub (`https://huggingface.co`).

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

Yes. They can open the issue in the LBox Open GitHub repository.

Any other comments?

None.