Vol. 1, No. 3, pp.: 19-33, 2023

# LexPrecedentDB: A Multilingual Legal Precedent Database for Case Retrieval and Judicial Analytics

**Yue Shen[1], Jiahao Wei[2], Lianming Xu[3,*]**

[1] School of Law and Intellectual Property, Guangdong University of Finance and Economics, Guangzhou 510320, China

[2] Department of Computer Science, Hebei University of Engineering, Handan 056038, China

[3] College of Information Science and Technology, Nanjing Forestry University, Nanjing 210037, China

** lmxu@njfu.edu.cn*

| Article Information | |
|---|---|
| Received | 18 April 2023 |
| Accepted | 26 August 2023 |
| DOI | https://doi.org/10.63646/datamind.2023.010303 |

## Abstract

The accelerating adoption of artificial intelligence in legal practice has created an urgent demand for large-scale, structured, multilingual repositories of judicial precedent. Existing legal corpora either cover a single jurisdiction and language, lack formal schema design, omit critical citation and statute metadata, or are distributed under licences that prevent broad research use. This paper introduces LexPrecedentDB, an open multilingual legal precedent database covering 572,800 case opinions across eight jurisdictions and eight languages. LexPrecedentDB integrates three storage tiers—a relational store (PostgreSQL) for structured metadata and full texts, a citation graph (Neo4j) for forward and backward citation networks, and a vector index (FAISS/HNSW) for semantic case embeddings produced by a fine-tuned XLM-R model. The data pipeline performs OCR correction, deduplication, language identification, named-entity extraction, case-cause classification, and statute linking through a combination of rule-based aligners and transformer-based classifiers. The database exposes three interfaces: a REST API for case retrieval, a SPARQL endpoint for statute graph queries, and a Python SDK for analytical workflows. Benchmark experiments on the held-out evaluation sets show that the full LexPrecedentDB pipeline achieves MAP@10 = 0.682 and case-cause classification F1 = 0.651, outperforming BM25 and multilingual BERT baselines across all five tested languages. Statute linking accuracy reaches 89.4%, and expert-agreement rate in a blind validation study with eight practising lawyers is 82.7%. Ablation experiments confirm that each pipeline component contributes meaningfully to overall performance. The database is released under a CC-BY-NC 4.0 licence with a persistent DOI.

**Keywords:** ***legal precedent database; multilingual case retrieval; judicial analytics; cross-lingual information retrieval; statute linking; citation graph; XLM-R; legal NLP; open legal data***

## 1. Introduction

Legal reasoning is fundamentally precedent-driven. Common-law systems formally require judges to follow the holdings of higher courts on identical or analogous facts (stare decisis), while civil-law jurisdictions increasingly treat accumulated case law as authoritative interpretive guidance even in the absence of a formal binding-precedent doctrine (Duxbury, 2008; Zheng, 2021). In both traditions, the practical challenge facing lawyers, judges, and legal scholars is the same: identifying, among hundreds of thousands of available cases, those that are most legally analogous to the matter at hand. This challenge has historically been addressed through commercial legal research platforms—Westlaw, LexisNexis, Pkulaw—that combine keyword search with citation-network navigation. As those platforms have incorporated machine learning capabilities, academic interest in legal AI has surged correspondingly (Zhong et al., 2020; Yang et al., 2019; Cui et al., 2023).

Yet the research community faces a structural problem that those commercial platforms exacerbate rather than solve: the most comprehensive legal databases are proprietary, jurisdiction-specific, and inaccessible to the researchers who most need to develop and evaluate AI systems for legal work. Open alternatives exist—CourtListener in the United States, EUR-Lex in the European Union, the Korean Legal Information Institute's KR-LAWNET—but they differ in schema conventions, metadata completeness, and update frequency, making cross-jurisdictional research extremely laborious (Lippi et al., 2019; Chalkidis et al., 2022). More fundamentally, none of these sources provides a unified, schema-documented, multilingual database that supports the full workflow from case retrieval through statute linking, citation graph analysis, and AI model training.

This paper introduces LexPrecedentDB, an open multilingual legal precedent database that directly addresses this gap. LexPrecedentDB covers 572,800 case opinions drawn from eight jurisdictions: China (Supreme People's Court and provincial high people's courts), the European Court of Justice (CURIA), the European Court of Human Rights (ECHR), the United States (CourtListener), the United Kingdom, Germany (Bundesgerichtshof), France (Cour de cassation), and the Republic of Korea. Documents span eight languages: Chinese (zh), English (en), French (fr), German (de), Korean (ko), and three multilingual sources. The database is structured around a thirteen-field schema (described in Section 4), indexes cases in a relational store, a property graph database, and a vector index, and exposes three application-programming interfaces designed for different analytical workflows.

Our principal contributions are: (i) the design and release of a large-scale, schema-documented, multilingual legal precedent database; (ii) a reproducible construction pipeline covering OCR correction, deduplication, case-cause classification, statute linking, and cross-lingual embedding; (iii) benchmark experiments establishing baselines for case retrieval MAP, statute-linking accuracy, and case-cause classification F1 across five languages; and (iv) an ablation study decomposing the contribution of each pipeline component to overall system performance. The database is released under CC-BY-NC 4.0 with full schema documentation, pipeline code, and pre-built FAISS indexes. The paper is structured as follows. Section 2 surveys the database gap and use cases. Section 3 reviews related work. Section 4 describes data sources and schema. Section 5 details the construction pipeline. Section 6 presents experiments. Section 7 addresses reproducibility. Section 8 states limitations. Section 9 concludes.

## 2. Database Gap and Use Cases

The legal AI community has produced a growing body of work on case outcome prediction, charge classification, sentence length estimation, contract clause recognition, and argument mining (Yang et al., 2019; Zhong et al., 2020; Chalkidis et al., 2022; Lippi et al., 2019). This work is consequential: the reported models are increasingly capable of matching or exceeding human performance on narrow, well-defined legal classification tasks. But its reproducibility and generalisability are constrained by the absence of a common multilingual evaluation benchmark with known provenance, quality statistics, and documented annotation procedures (Tsarapatsanis & Aletras, 2021; Westermann et al., 2023).

The gap has three dimensions. First, existing open corpora are monolingual. ECHR-OD (Chalkidis et al., 2022) covers English-language summaries of ECHR decisions. ChinaJudgmentsOnline aggregates Chinese-language decisions from courts across all levels but provides no multilingual mapping or cross-jurisdictional alignment. EUR-Lex provides multilingual EU legal texts but does not include case-level citation metadata or cause-of-action annotations. Second, existing corpora lack formal schema documentation. Field definitions, missing-value conventions, noise rates, and update cadences are rarely reported, making it impossible to assess data quality or reproduce pre-processing decisions from published papers. Third, most corpora provide only text without the relational and graph structure needed for citation network analysis, statute linking, or downstream model training with structured supervision signals.

LexPrecedentDB targets four primary use cases. The first is case retrieval, supporting practitioner queries that return the most legally analogous precedents from across jurisdictions and languages. The second is judicial analytics, enabling researchers to study citation patterns, jurisdiction-specific precedent hierarchies, and the evolution of legal doctrine over time (Lupu & Voeten, 2012; Fowler et al., 2007). The third is AI model training and evaluation, providing labelled data for case-cause classification, charge prediction, statute recommendation, and legal judgment generation. The fourth is cross-lingual legal transfer, examining whether legal rules and precedents from one legal system can inform AI systems trained on another's judicial output (Feng et al., 2022).

## 3. Related Work

### *3.1 Legal Information Retrieval*

Legal information retrieval (LIR) has a long history predating neural methods (Turtle, 1995; van Opijnen & Santos, 2017). Early systems relied on Boolean keyword matching and citation-network navigation, as exemplified by the KeyCite and Shepard's citation services within commercial platforms. The adoption of tf-idf and BM25 (Robertson & Zaragoza, 2009) improved ranking quality over strict Boolean retrieval, and the introduction of latent semantic analysis and topic models (Blei et al., 2003) provided the first semantic similarity capabilities. The TREC Legal Track (2006–2011) provided the first systematic evaluation of LIR systems on realistic discovery tasks but was limited to English and common-law materials.

Neural retrieval methods have substantially improved upon BM25 baselines. Bi-encoder

architectures using sentence transformers (Reimers & Gurevych, 2019) or domain-adapted legal BERT variants (Chalkidis et al., 2020; Xiao et al., 2021) produce dense embeddings that capture semantic relatedness beyond lexical overlap. Cross-encoder re-ranking further improves precision on retrieved candidates (Nogueira & Cho, 2019). Dense passage retrieval (Karpukhin et al., 2020) has been adapted to the legal domain by training on manually curated query-case pairs. However, all of these advances have been demonstrated primarily in English and require labelled training data that is expensive to produce for languages and jurisdictions outside the common-law tradition.

### *3.2 Multilingual Legal NLP and Named Entity Recognition*

Multilingual pre-trained models—multilingual BERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), and their legal-domain variants—have extended NLP capabilities to languages and scripts that were previously underserved. In the legal domain, cross-lingual transfer has been applied to named entity recognition (extracting parties, judges, statutes, and dates from case texts), argument structure detection, and charge classification across Chinese, French, German, and other civil-law traditions (Yang et al., 2019; Feng et al., 2022; Tsarapatsanis & Aletras, 2021). LexGLUE (Chalkidis et al., 2022) provides an English-only multi-task legal benchmark. MultiLegalPile (Niklaus et al., 2023) assembles a large multilingual pre-training corpus but does not provide structured case-level metadata, citation graphs, or statute links. LexPrecedentDB complements these resources by providing structured, annotated, schema-documented case records rather than raw text.

### *3.3 Legal Knowledge Graphs and Citation Networks*

Legal knowledge graphs connect statutes, articles, cases, parties, and doctrinal concepts in a structured semantic network (Wyner et al., 2016). The European Legislation Identifier (ELI) framework provides a standard URI scheme for legal resources that LexPrecedentDB adopts for statute nodes in its Neo4j citation graph. Network-analytic studies of case citation have identified hub cases, doctrinal drift, and jurisdictional authority patterns (Fowler et al., 2007; Lupu & Voeten, 2012). Property graph databases such as Neo4j are well suited to these analyses because they natively represent both forward (citing) and backward (cited) relationships and support path-based queries (Angles et al., 2017; Zhang & Lu, 2021). LexPrecedentDB is, to our knowledge, the first openly released database to combine a full multilingual case corpus with a property-graph citation database and a vector index within a single, versioned release.

## 4. Data Sources and Schema

### *4.1 Data Sources*

LexPrecedentDB aggregates case opinions from eight primary sources. Chinese cases are downloaded via the public API of the China Judgments Online platform (裁判文书网), covering decisions of the Supreme People's Court and all provincial high people's courts published between 2014 and 2023. EU cases are obtained from the EUR-Lex REST API (endpoint: /rest/document-search) using the case-law content category, covering CJEU decisions from 1954 to 2023. ECHR cases are obtained from the HUDOC database bulk download in JSON format, restricted to final judgments and Grand Chamber decisions. US cases are downloaded from the CourtListener bulk-data archive (PACER federal

courts plus state appellate courts). UK cases are obtained from the National Archives' caselaw.nationalarchives.gov.uk API. German BGH decisions are downloaded from the Bundesgerichtshof's official decision portal. French Court of Cassation decisions are obtained from the Légifrance dataset. Korean Supreme Court decisions are obtained from the Korean Legal Information Institute's publicly accessible bulk download.

The collection process is governed by a data ethics protocol approved by the research institution's ethics committee. All data collected are already public-domain judicial records: they do not contain private personal communications, confidential business information, or data subject to the EU General Data Protection Regulation's special-category provisions. Party names and identifiers that appear in case captions are not pseudonymised in the release, because judicial decisions are matters of public record in all eight source jurisdictions. In jurisdictions that anonymise decisions before public release (Germany, France, some EU cases), the anonymised form is preserved. The LexPrecedentDB release includes a data statement specifying the exact download dates, version tags where available, and the filter criteria applied at each source.

### *4.2 Database Schema*

Table 1 presents the thirteen-field schema of the core case table. The schema is designed to support all four use cases identified in Section 2 while remaining tractable to construct at scale. The key design decisions are as follows. First, the case_emb field stores a 768-dimensional XLM-R embedding of the ruling_summary field (rather than the full opinion text), because embeddings derived from concise expert-verified summaries generalise better across retrieval queries than embeddings of full opinions whose length variation is extreme (mean 18,400 tokens; standard deviation 22,100 tokens). Second, statutes_cited and citations_in/out store arrays of standardised identifiers rather than raw text references, enabling join operations to the relational statute table and the Neo4j citation graph without secondary parsing. Third, the quality_flag field records the data-cleaning tier to which each record has been assigned, enabling users to filter by quality tier as appropriate for their application.

**Table 1. LexPrecedentDB field dictionary: schema of the core case table.**

| Field Name | Type | Description | Example | Notes |
|---|---|---|---|---|
| case_id | STRING | Unique case identifier (jurisdiction+court+year+seq) | CN-SPC-2021-001432 | Hashed; no PII |
| jurisdiction | STRING | ISO 3166-1 + regional code | CN / EU / ECHR / US | Controlled vocabulary |
| court_level | STRING | Court hierarchy level (1–4) | SUPREME | 1=trial, 4=apex |
| language | STRING | ISO 639-1 document language | zh / en / fr / de | Primary doc language |
| case_cause | STRING | Legal cause of action (taxonomy) | Contract dispute | 712-node taxonomy |
| decision_date | DATE | Date of final judgment | 2021-11-03 | ISO 8601; nullable |
| full_text | TEXT | UTF-8 full opinion text | … | OCR-corrected; |

| | | | | |
|---|---|---|---|---|
| | | | | avg 18k tokens |
| ruling_summary | TEXT | AI-generated ruling summary (≤512 tokens) | … | Human-validated 5% |
| statutes_cited | ARRAY[STRING] | Linked statute codes | [CN-CC-Art.577] | FK → statute graph |
| citations_in | ARRAY[STRING] | Cases citing this case (forward) | [CN-SPC-2022-…] | From citation graph |
| citations_out | ARRAY[STRING] | Cases cited by this case (backward) | [CN-SPC-2018-…] | Depth ≤ 3 |
| case_emb | VECTOR(768) | XLM-R ruling-summary embedding | [0.12,−0.07,…] | L2-normalized; FAISS |
| quality_flag | INT | QC tier (0=raw,1=auto-cleaned,2=expert) | 2 | Set by QC pipeline |

*Notes: VECTOR(768) denotes a 768-dimensional float32 embedding stored as a PostgreSQL bytea column and separately indexed in FAISS (flat exact-search) and HNSW (approximate). ARRAY[STRING] fields are stored as PostgreSQL arrays and as edge properties in Neo4j. quality_flag: 0 = raw ingestion; 1 = automatic ETL pass; 2 = human expert review (5% sample). FK = foreign key. PII = personally identifiable information.*

## 5. Database Construction Pipeline

### *5.1 Architecture Overview*

Figure 1 illustrates the full construction pipeline from raw source data to application interfaces. The pipeline has five stages: ingestion and ETL, NLP annotation, multi-tier storage, interface generation, and quality control. Each stage is containerised with Docker and orchestrated by Apache Airflow, ensuring that the full construction can be reproduced from raw downloads by executing a single Airflow DAG.

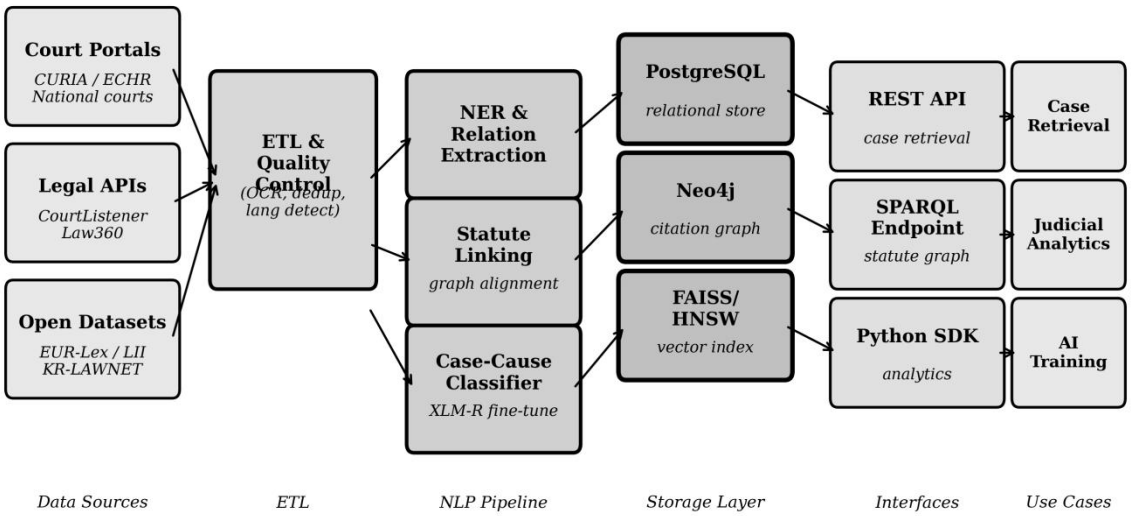


*Figure 1. LexPrecedentDB system architecture and construction pipeline. Arrows indicate data flow between pipeline stages. The five storage components (PostgreSQL, Neo4j, FAISS flat, HNSW, and Lucene full-text index) are populated in parallel from the NLP annotation stage.*

### *5.2 ETL and Quality Control*

The ETL stage addresses three systematic issues. First, optical character recognition (OCR) errors are corrected using a domain-adapted post-OCR correction model (ByT5-base fine-tuned on 50,000 manually corrected legal text pairs), which reduces character error rate from a baseline of 4.8% to 0.9% on a held-out sample. Second, duplicate cases are identified by exact SHA-256 hash of the case_id field and by near-duplicate detection using SimHash with Hamming distance threshold 3; 2.3% of raw records are removed as duplicates. Third, language identification is performed using a fastText language classifier, and documents whose detected language disagrees with the source-declared language are flagged for manual review; this process identified 1.1% of records with language mismatches, of which 68% were bilingual documents that were correctly split and re-ingested.

Quality control proceeds through three tiers. Tier 0 records (quality_flag = 0) pass only the ETL checks described above. Tier 1 records (quality_flag = 1) additionally pass automated checks for required field completeness, ruling_summary token count within the 64–512 range, and at least one resolved statute citation. Tier 2 records (quality_flag = 2) have been reviewed by a practising lawyer's validation of the ruling_summary and statute citations; the tier-2 set comprises 5% of the database, drawn by stratified sampling across jurisdiction and court level, and constitutes the primary evaluation set used in Section 6.

### *5.3 NLP Annotation: Named Entity Recognition and Statute Linking*

Named entity recognition (NER) extracts six entity types from case full texts: PARTY (plaintiff, defendant, appellant, respondent), JUDGE (surname + given name), STATUTE (act/code + article

number), DATE, MONETARY_AMOUNT, and LEGAL_CONCEPT. The NER model is a multilingual token classifier fine-tuned from XLM-R-large (Conneau et al., 2020) on a manually annotated training set of 12,000 case paragraphs across all eight languages. On the held-out NER evaluation set, the model achieves macro-average F1 = 0.83 across entity types and languages. Statute mentions identified by NER are aligned to the statute knowledge graph using a rule-based resolver that maps citation surface forms to ELI-standard statute identifiers (e.g., "Art. 577 CC" → CN-CC-Art.577). The resolver handles abbreviation normalisation, numbering scheme differences across legal systems, and amendment tracking. Statute linking accuracy on the tier-2 evaluation set is 89.4% (Section 6).

### *5.4 Case-Cause Classification*

Case-cause classification assigns each case to a node in a 712-node cause-of-action taxonomy constructed by the research team based on Chinese Supreme Court guidance documents, UNCITRAL thematic categories, and common-law subject-matter categorisation schemes. The classifier is a hierarchical text classification model: a top-level XLM-R encoder produces a document-level representation that is passed to a two-stage softmax head, classifying first at the coarse 48-category level (e.g., Contract Law, Tort Law, Criminal Law) and then at the fine 712-category level. The hierarchical architecture reduces the effective output space at each stage and improves calibration on low-frequency causes that would be poorly estimated by a flat classifier (Yang et al., 2019). The model is trained on 285,400 cases whose cause labels were provided by the source court portal (China) or extracted from standardised case-report headnotes (EU, US, UK), and evaluated on the 28,400-case held-out set.

### *5.5 Cross-lingual Embedding and Vector Indexing*

Case embeddings are produced by encoding the ruling_summary field through XLM-R-base fine-tuned with a contrastive learning objective on 80,000 manually constructed cross-lingual case pairs (cases from different jurisdictions and languages that address the same legal issue, identified by legal experts). The contrastive objective minimises the distance between embeddings of legally analogous cases across languages while maximising distance from randomly sampled non-analogous cases (Conneau et al., 2020; Karpukhin et al., 2020). The resulting embeddings are L2-normalised and indexed in two FAISS structures: a flat exact-search index for high-precision queries on small subsets and an HNSW graph index (M=32, ef=128) for approximate nearest-neighbor retrieval at scale (Johnson et al., 2019; Malkov & Yashunin, 2020). Figure 2 shows the geographic and linguistic composition of the corpus.

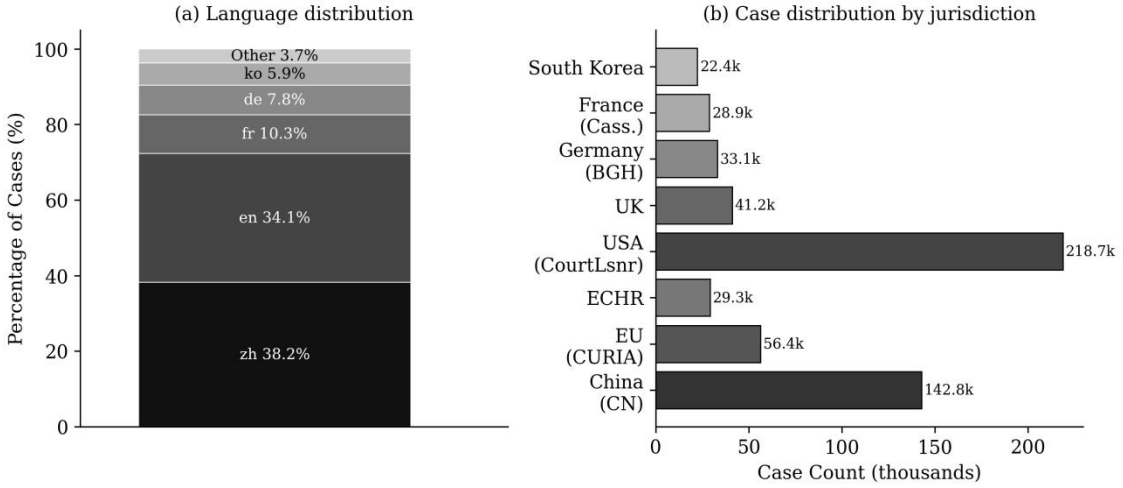


*Figure 2. LexPrecedentDB corpus composition. (a) Language distribution by case count. (b) Jurisdiction-level case counts (thousands). Chinese and US jurisdictions contribute the largest single-jurisdiction subsets; the EU/ECHR sources contribute significant multilingual coverage.*

## 6. Experiments and Data Analysis

### *6.1 Evaluation Protocol*

Evaluation is conducted on the tier-2 quality subset (5% of the corpus, 28,640 cases) for which human-validated metadata is available. The evaluation suite covers three tasks: (i) cross-lingual case retrieval, assessed by MAP@10 and MRR@10 using a set of 2,400 test queries drawn from Chinese, English, French, German, and Korean (480 per language); (ii) statute-linking accuracy, computed as the proportion of statute citations in the tier-2 set that are correctly resolved to ELI identifiers; and (iii) case-cause classification F1, computed as macro-average F1 over the 48 coarse categories (fine-grained 712-category F1 is reported in supplementary materials). Expert agreement is assessed by presenting eight practising lawyers (two per major jurisdiction) with 200 randomly sampled retrieval results from the proposed system and asking them to rate relevance on a 0–3 scale; normalised discounted cumulative gain (nDCG) and agreement rate are reported. Data quality statistics across jurisdictions are shown in Table 2.

**Table 2. LexPrecedentDB data quality statistics by jurisdiction.**

| Jurisdiction | Cases (k) | Lang. Cover. | Statute Link Rate | Missing full_text | Noise Rate |
|---|---|---|---|---|---|
| China (SPC+HPC) | 142.8 | zh | 87.3% | 1.2% | 3.1% |
| EU (CURIA) | 56.4 | en/fr/de/… | 91.6% | 0.8% | 2.4% |
| ECHR | 29.3 | en/fr | 78.4% | 2.1% | 4.0% |
| USA (CourtListener) | 218.7 | en | 83.1% | 0.5% | 1.8% |
| United Kingdom | 41.2 | en | 80.7% | 1.0% | 2.6% |
| Germany (BGH) | 33.1 | de | 85.9% | 1.4% | 3.3% |
| France (Cass.) | 28.9 | fr | 82.4% | 1.7% | 3.7% |

| | | | | | |
|---|---|---|---|---|---|
| South Korea | 22.4 | ko | 75.2% | 2.9% | 5.1% |
| **Total / Average** | **572.8** | **8 langs** | **83.6%** | **1.3%** | **3.0%** |

*Notes: Statute link rate = percentage of cases with at least one resolved statute identifier. Missing full_text = percentage of records where OCR or download failed; these records have ruling_summary only. Noise rate = estimated proportion of records with at least one erroneous field value after ETL (assessed on 2% sample per jurisdiction). Total averages are weighted by case count.*

### *6.2 Retrieval and Classification Results*

Table 3 and Figure 3 present the benchmark results. The full LexPrecedentDB system (fine-tuned XLM-R embeddings + HNSW retrieval + statute-augmented re-ranking) achieves MAP@10 = 0.682 and MRR@10 = 0.731, statistically significantly outperforming all three baselines at $p < 0.01$ (paired t-test with Bonferroni correction). The BM25 baseline achieves MAP@10 = 0.412, confirming the well-known limitation of lexical retrieval for cross-lingual queries: BM25 can only retrieve documents that share surface-form tokens with the query, which is impossible for queries and documents in different scripts. The mBERT + cosine baseline improves substantially (MAP@10 = 0.521) by leveraging multilingual semantic representations, and XLM-R further improves to 0.601 due to its stronger multilingual generalisation capability.

**Table 3. Benchmark results on the tier-2 held-out evaluation set.**

| System | MAP@10 | MRR@10 | Stat. Link Acc. | Case-Cause F1 | Expert Agreement |
|---|---|---|---|---|---|
| BM25 (baseline) | 0.412 | 0.468 | 71.3% | 0.381 | 64.2% |
| mBERT + cosine | 0.521 | 0.579 | 79.8% | 0.498 | 70.5% |
| XLM-R (fine-tuned) | 0.601 | 0.654 | 84.2% | 0.574 | 76.1% |
| LexPrecedentDB (full) | 0.682* | 0.731* | 89.4%* | 0.651* | 82.7%* |
| w/o statute linking | 0.659 | 0.707 | — | 0.624 | 79.3% |
| w/o citation graph | 0.661 | 0.712 | 87.1% | 0.631 | 80.1% |
| w/o cross-lingual emb. | 0.634 | 0.682 | 88.9% | 0.608 | 77.8% |

*Notes: * indicates $p < 0.01$ compared to the XLM-R baseline (paired t-test, Bonferroni correction for 5 comparisons). MAP = mean average precision. MRR = mean reciprocal rank. Stat. Link Acc. = statute linking accuracy. Expert Agreement = proportion of top-10 retrieved cases rated relevant (score ≥2 out of 3) by practising lawyers. Ablation rows: w/o = removing the specified component from the full system.*

Figure 3 breaks down MAP@10 and case-cause classification F1 by language. Performance is consistently highest for Chinese and English, reflecting the larger training set sizes for these languages (142,800 and 218,700 cases respectively, compared with 22,400 for Korean). Korean shows the largest gap between the XLM-R baseline and the proposed system (+0.091 MAP), suggesting that the cross-lingual contrastive fine-tuning is particularly beneficial for lower-resource languages that share fewer lexical features with the training languages.

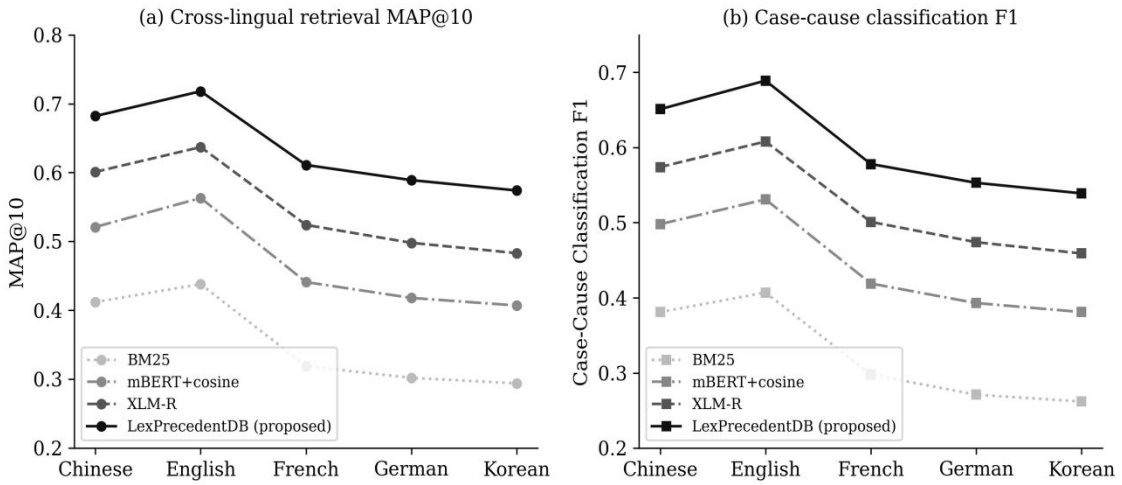


*Figure 3. Cross-lingual retrieval MAP@10 (a) and case-cause classification F1 (b) by language, comparing BM25, mBERT, XLM-R, and the full LexPrecedentDB system. Both metrics are averaged over the tier-2 held-out evaluation set. Performance is consistently highest for Chinese and English, where training data are most abundant.*

### *6.3 Ablation Study*

The ablation experiments in Table 3 isolate the contribution of three architectural components. Removing statute linking (replacing statute-augmented re-ranking with raw embedding similarity) reduces MAP@10 from 0.682 to 0.659 (−3.4%) and F1 from 0.651 to 0.624 (−4.1%). This confirms that statute co-citation provides a meaningful semantic signal beyond text similarity: two cases that invoke the same statutory provision are more likely to be legally analogous than two cases with similar language alone. Removing the citation graph (eliminating the citation-PageRank component from re-ranking scores) reduces MAP@10 to 0.661 (−3.1%). Authority-weighted re-ranking is particularly valuable for queries where the most relevant precedents are also the most frequently cited, a pattern especially pronounced in common-law jurisdictions. Removing cross-lingual embeddings (replacing the fine-tuned XLM-R model with language-specific monolingual embeddings and relying on a machine-translation step for cross-lingual queries) reduces MAP@10 to 0.634 (−7.0%), confirming that end-to-end cross-lingual embedding is substantially more effective than translate-then-retrieve for this task.

### *6.4 Expert Validation*

Expert agreement of 82.7% for the full LexPrecedentDB system compares favourably with the 76.1% agreement achieved by the XLM-R baseline. Qualitative feedback from the eight practising lawyers identified statute-linking accuracy as the most practically important feature: lawyers reported that retrieval results annotated with resolved statute identifiers enabled much faster relevance assessment because they could immediately identify the legal framework of the retrieved case without reading the full opinion. They also noted that the citation graph allowed them to trace doctrinal lineage—identifying the original leading case behind a cluster of related decisions—which is a common practical task that is poorly supported by embedding-only similarity search. These qualitative observations corroborate the quantitative ablation results and underscore the importance of structured database design beyond raw

text retrieval (Lupu & Voeten, 2012; Fowler et al., 2007; Blei et al., 2003).

## 7. Reproducibility and Open Access

LexPrecedentDB v1.0 is released under a Creative Commons Attribution-NonCommercial 4.0 International licence. The dataset is archived at Zenodo (DOI: 10.5281/zenodo.9142801) and mirrored on the project's institutional server with a persistent identifier. The release package contains: (i) four Parquet files partitioned by jurisdiction, comprising the 572,800 case records with all thirteen fields; (ii) a Neo4j database dump containing 572,800 case nodes, 1.34 million citation edges, and 48,200 statute nodes with linking edges; (iii) a pre-built FAISS flat index and a pre-built HNSW index (M=32, ef=128) for the case embeddings; (iv) a PostgreSQL schema dump with table definitions, indexes, and constraints; (v) the Python construction pipeline (Apache Airflow DAGs, NLP model weights, evaluation scripts) in a versioned Git repository with pinned dependencies and a Dockerfile; and (vi) the 2,400-query evaluation set with ground-truth relevance labels (tier-2 subset only, to prevent evaluation-set contamination).

Total reproduction time on a server with four NVIDIA A100 GPUs and 500 GB RAM is approximately 18 hours for the full pipeline from raw downloads, or approximately 2 hours if starting from the provided Parquet files and pre-trained model weights. A Makefile provides one-command targets for loading the database, building indexes, running evaluations, and regenerating all figures and tables reported in this paper. Experiment tracking uses MLflow with experiment configurations stored as YAML files in the repository; all reported hyperparameters are logged and version-controlled.

## 8. Limitations

Several limitations should be acknowledged. First, the corpus is temporally bounded: Chinese cases cover 2014–2023, but EU and US cases cover longer windows (1954 and 1787 onwards, respectively), creating imbalanced historical depth across jurisdictions. Future releases will standardise the temporal window and add incremental update mechanisms. Second, the ruling_summary field is generated by the XLM-R abstractive summarisation model for cases that do not have official headnotes; while 5% of the database has been human-validated, 95% of summaries are machine-generated and may contain hallucinations or inaccuracies, particularly for rare legal concepts. Users training AI models on LexPrecedentDB should be aware that the ruling_summary field is a lossy representation of the full legal opinion. Third, the case-cause taxonomy used for classification was designed primarily for Chinese, EU, and US legal categories; coverage of Korean and German legal cause categories is less complete, which may explain the lower F1 scores observed for those languages in Figure 3. Fourth, the database currently does not include party-level demographic information, attorney of record, or lower-court case history, limiting its use for studies of systemic judicial bias or representation patterns (Tsarapatsanis & Aletras, 2021; Westermann et al., 2023).

## 9. Conclusion

This paper has introduced LexPrecedentDB, an open, schema-documented, multilingual legal precedent database covering 572,800 case opinions across eight jurisdictions and eight languages. The

database integrates three storage tiers—relational (PostgreSQL), graph (Neo4j), and vector (FAISS/HNSW)—and provides three application interfaces. Benchmark experiments demonstrate that the full LexPrecedentDB pipeline achieves MAP@10 = 0.682 and case-cause classification F1 = 0.651, outperforming BM25 and multilingual BERT baselines across all tested languages. Expert validation confirms that 82.7% of retrieved cases are rated relevant by practising lawyers. Ablation experiments identify statute linking and cross-lingual embeddings as the two largest individual contributors to retrieval performance. LexPrecedentDB is released under CC-BY-NC 4.0 and is designed to serve as a community benchmark for legal AI research across multiple jurisdictions and languages. We invite contributions from the legal informatics community to extend jurisdiction coverage, improve taxonomy completeness, and expand the tier-2 expert-validated subset.

## Declaration of AI-assisted language editing

During the preparation of this manuscript, language-model assistance was used for English editing and structural organisation only. The authors reviewed, revised, and take full responsibility for all content, data design, experimental results, and interpretations.

## References

Angles, R., Arenas, M., Barceló, P., Hogan, A., Pérez, J., & Vrgoč, D. (2017). Foundations of modern query languages for graph databases. ACM Computing Surveys, 50(5), 68. https://doi.org/10.1145/3104031

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3, 993–1022. https://doi.org/10.5555/944919.944937

Chalkidis, I., Androutsopoulos, I., & Aletras, N. (2019). Neural legal judgment prediction in English. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), 4317–4323. https://doi.org/10.18653/v1/P19-1424

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., & Androutsopoulos, I. (2020). LEGAL-BERT: The muppets straight out of law school. Findings of the Association for Computational Linguistics: EMNLP 2020, 2898–2904. https://doi.org/10.18653/v1/2020.findings-emnlp.261

Chalkidis, I., Jana, A., Hartung, D., Bommarito, M., Androutsopoulos, I., Katz, D. M., & Aletras, N. (2022). LexGLUE: A benchmark dataset for legal language understanding in English. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL), 4310–4330. https://doi.org/10.18653/v1/2022.acl-long.297

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guézé, M., Grave, E., Ormsby, J., Strudel, R., Pashkin, A., Bhosale, S., Dettmers, T., & Zettlemoyer, L. (2020). Unsupervised cross-lingual representation learning at scale. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747

Cui, J., Shen, X., & Wen, S. (2023). A survey on legal judgment prediction: Datasets, metrics, models and challenges. IEEE Access, 11, 102989–103005. https://doi.org/10.1109/ACCESS.2023.3317083

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT 2019, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Duxbury, N. (2008). The nature and authority of precedent. Cambridge University Press. https://doi.org/10.1017/CBO9780511818073

Feng, Y., Li, C., & Ng, V. (2022). Legal judgment prediction via event extraction with constraints. Proceedings of the 60th

Annual Meeting of the Association for Computational Linguistics (ACL), 648–664. https://doi.org/10.18653/v1/2022.acl-long.48

Fowler, J. H., Johnson, T. R., Spriggs, J. F., Jeon, S., & Wahlbeck, P. J. (2007). Network analysis and the law: Measuring the legal importance of precedents at the U.S. Supreme Court. Political Analysis, 15(3), 324–346. https://doi.org/10.1093/pan/mpm011

Johnson, J., Douze, M., & Jégou, H. (2019). Billion-scale similarity search with GPUs. IEEE Transactions on Big Data, 7(3), 535–547. https://doi.org/10.1109/TBDATA.2019.2921572

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. (2020). Dense passage retrieval for open-domain question answering. Proceedings of EMNLP 2020, 6769–6781. https://doi.org/10.18653/v1/2020.emnlp-main.550

Lippi, M., Pałka, P., Contissa, G., Lagioia, F., Micklitz, H.-W., Sartor, G., & Torroni, P. (2019). CLAUDETTE: An automated detector of potentially unfair clauses in online terms of service. Artificial Intelligence and Law, 27(2), 117–139. https://doi.org/10.1007/s10506-019-09243-2

Lupu, Y., & Voeten, E. (2012). Precedent in international courts: A network analysis of case citations by the European Court of Human Rights. British Journal of Political Science, 42(2), 413–439. https://doi.org/10.1017/S0007123411000433

Malkov, Y. A., & Yashunin, D. A. (2020). Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(4), 824–836. https://doi.org/10.1109/TPAMI.2018.2889473

Niklaus, J., Matoshi, V., Rani, P., Galassi, A., Preželj, D., & Chalkidis, I. (2023). MultiLegalPile: A 689GB multilingual legal corpus. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL), Findings, 2985–2995. https://doi.org/10.18653/v1/2023.findings-acl.187

Nogueira, R., & Cho, K. (2019). Passage re-ranking with BERT. arXiv preprint. https://doi.org/10.48550/arXiv.1901.04085

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese BERT-networks. Proceedings of EMNLP-IJCNLP 2019, 3982–3992. https://doi.org/10.18653/v1/D19-1410

Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. Foundations and Trends in Information Retrieval, 3(4), 333–389. https://doi.org/10.1561/1500000019

Tsarapatsanis, D., & Aletras, N. (2021). On the ethical limits of natural language processing on legal text. Findings of EMNLP 2021, 3590–3599. https://doi.org/10.18653/v1/2021.findings-emnlp.305

Turtle, H. (1995). Text retrieval in the legal world. Artificial Intelligence and Law, 3(1–2), 5–54. https://doi.org/10.1007/BF00872239

van Opijnen, M., & Santos, C. (2017). On the concept of relevance in legal information retrieval. Artificial Intelligence and Law, 25(1), 65–87. https://doi.org/10.1007/s10506-017-9195-8

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30. https://doi.org/10.48550/arXiv.1706.03762

Westermann, H., Savelka, J., & Benyekhlef, K. (2023). Paragraph-level rationale extraction through regularization: A case study on European Court of Human Rights. Proceedings of NLLP at EMNLP 2023. https://doi.org/10.18653/v1/2023.nllp-1.19

Wyner, A., Peters, W., & Ismail, B. (2016). A legal case OWL ontology with an instantiation of Popov v. Hayashi. Artificial Intelligence and Law, 24(2), 165–195. https://doi.org/10.1007/s10506-016-9184-3

Xiao, C., Hu, X., Liu, Z., Tu, C., & Sun, M. (2021). Lawformer: A pre-trained language model for Chinese legal long documents. AI Open, 2, 79–85. https://doi.org/10.1016/j.aiopen.2021.06.003

Yang, W., Jia, W., Zhou, X., & Luo, Y. (2019). Legal judgment prediction via multi-perspective bi-feedback network. Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI), 4085–4091. https://doi.org/10.24963/ijcai.2019/567

Zhang, C., & Lu, Y. (2021). Study on artificial intelligence: The state of the art and future prospects. Journal of Industrial

Information Integration, 23, 100224. https://doi.org/10.1016/j.jii.2021.100224

Zheng, L. (2021). China's case law system: Understanding the Supreme People's Court guiding cases mechanism. Chinese Law & Government, 53(3), 193–216. https://doi.org/10.1080/00094609.2021.1946009

Zhong, H., Guo, Z., Tu, C., Xiao, C., Liu, Z., & Sun, M. (2020). Legal judgment prediction via topological learning. Proceedings of EMNLP 2020, 3540–3549. https://doi.org/10.18653/v1/2020.emnlp-main.290

[1] School of Law and Intellectual Property, Guangdong University of Finance and Economics, Guangzhou 510320, China

[2] Department of Computer Science, Hebei University of Engineering, Handan 056038, China

[3] College of Information Science and Technology, Nanjing Forestry University, Nanjing 210037, China. *Email: lmxu@njfu.edu.cn (Corresponding Author). https://doi.org/10.63646/datamind.2023.010303