Evaluating knowledge-based approaches for legal text analysis: A benchmark study

Computer Law & Security Review 61 (2026) 106279

Contents lists available at ScienceDirect

Computer Law & Security Review: The International Journal of Technology Law and Practice

journal homepage: www.elsevier.com/locate/clsr

Evaluating knowledge-based approaches for legal text analysis: A benchmark study☆

Awais Abdul Khaliq *, Davide Riva, Stefano Montanelli

Università degli Studi di Milano, Department of Computer Science, Via Celoria, 18, 20133 Milano, Italy

ARTICLE INFO

Keywords:
Legal NLP
Knowledge extraction
Legal benchmarking

ABSTRACT

Recent advances in transformer-based architectures and large-scale benchmarks have significantly improved Natural Language Processing performance. However, domain-specific areas such as law pose unique challenges due to specialized terminology, complex semantics, and limited annotated data.

In this paper, we benchmark ASKE, a knowledge extraction approach designed to capture latent semantic structures in legal texts, against state-of-the-art legal-domain models using the LexGLUE datasets. We analyze its performance across diverse legal tasks through pseudo-precision and pseudo-recall metrics, highlighting its strengths, limitations, and adaptability. Based on the experimental results, we discuss potential application scenarios where ASKE can effectively support legal text analysis and downstream legal information processing.

1. Introduction

Since the introduction of transformer-based models and large-scale benchmarks such as GLUE and SuperGLUE, Natural Language Processing (NLP) has made significant advancements. These benchmarks have created the foundation for standardized evaluations of general language understanding tasks, allowing consistent comparisons between models and fostering innovation [1,2]. Despite these advancements, particular domains such as law present unique challenges that general-purpose NLP models and evaluation datasets often fail to fully address [3].

Legal documents range from laws and court judgments to legal contracts, playing a vital role as the foundation for the smooth functioning of society. However, these documents are difficult to analyze due to their domain-specific vocabulary, complex sentence structures, and the specialized knowledge required [4]. Furthermore, legal texts typically contain technical information, often requiring tailored models for proper interpretation and processing. Traditional supervised learning approaches typically suffer from a lack of labeled data and the high cost of manual annotation, making them impractical for large-scale applications in the legal domain [5].

Transformer-based architectures, such as BERT and its variants, have demonstrated strong results in specialized domains due to significant advancements in machine learning over the past few years. Appropriate benchmarks are required to assess the capabilities of these models on the specific challenges of legal text, such as context-aware semantics and domain-specific vocabulary. The Legal General Language Understanding Evaluation (LexGLUE) is a benchmark of this kind. LexGLUE is a comprehensive benchmarking evaluation for NLP models across a wide range of legal tasks, including multi-label classification, multi-class classification, and question answering. It features datasets representative of actual legal practices, such as EUR-LEX, ECtHR cases, SCOTUS cases, and EU legislation [6]. At the same time, domain-specific models have been developed to improve legal text analysis (e.g., LegalNLP [7], LegalBench [8], IL-TUR [9], Lextreme [10], LawBench [11]). These models have the capacity to extract meaningful knowledge from legal documents without the need of large-labeled datasets. As a further solution in this context, we recently proposed ASKE [12]. The ASKE approach and related tool is conceived to exploit latent structures and semantics of legal texts to extract legal knowledge in the form of concepts and relationships with the goal to support a wide range of downstream tasks such as document classification and legal information retrieval.

In this paper, we benchmark ASKE against several state-of-the-art (SOTA) models for legal text analysis using the LexGLUE datasets. The goal is to frame ASKE with respect to main legal-domain models by also discussing strengths and limitations of the approach, and providing a comprehensive analysis of its performance in various legal tasks. This includes pseudo-precision and pseudo-recall metrics, which quantify how closely the knowledge extracted by ASKE matches the ground

☆ This article is part of a Special issue entitled: 'Legal Systems' published in Computer Law & Security Review: The International Journal of Technology Law and Practice.

* Corresponding author.
E-mail addresses: awais.abdul@unimi.it (A. Abdul Khaliq), davide.riva1@unimi.it (D. Riva), stefano.montanelli@unimi.it (S. Montanelli).

https://doi.org/10.1016/j.clsr.2026.106279

Available online 26 February 2026

truth. These metrics provide insights on the ASKE performance across different legal datasets, showcasing their versatility and adaptation to the new challenges posed by legal texts. As a further contribution of this work, based on the obtained experimental results, we outline possible application scenarios where ASKE can be effectively employed to address the current challenges of the legal field.

The paper is organized as follows. Section 2 reviews the relevant literature. Section 3 provides an overview of the essential ASKE features. Section 4 details the considered LexGLUE datasets, and Section 5 presents the SOTA models considered in the evaluation. Section 6 presents the experimental results. Section 7 discusses possible ASKE applications in Legal NLP. Finally, Section 8 highlights the limitations and future work, and Section 9 provides our concluding remarks.

2. Literature review

In the recent years, Legal Natural Language Processing (L-NLP) has become an active area of research, with advancements in tasks such as summarization [13], information retrieval [14,15], document segmentation [16,17], case prediction [6,16,18], semantic analysis, and information extraction [19,20].

Zhong et al. demonstrated that topological learning is an extremely effective method for legal judgment prediction, emphasizing the need for tailored models in the legal domain for Natural Language Understanding (NLU) [21]. In contrast, Chalkidis et al. proposed a large-scale multi-label dataset for classifying legal documents, which has helped advance research in this field [5]. Chalkidis et al. also proposed methods for improving the interpretability and accuracy of NLU models in the legal domain. More recently, Holzenberger et al. have further explored sophisticated neural architectures for legal document classification, achieving significant improvements in performance [22]. Nguyen et al. contributed a new approach to unsupervised legal text clustering and enhanced the quality of knowledge extraction in legal contexts [23].

Benchmarks have played a significant role in improving techniques and models in many domains, such as computer vision [24,25] and reinforcement learning [26–28]. In the NLP domain, several benchmarks have been proposed, such as CLUE [29], GLUE [1], IndicNLPSuite [30], XTREME [31], and Super-GLUE [32]. However, these benchmarks are oriented toward general NLP tasks, and models developed for the generic domain do not perform well in the legal domain [6,14,16].

Some researchers have proposed specific language models for the legal domain, such as LegalBERT [5] and CaseLawBERT [33]. However, legal LLMs have shown limited success so far, and their generalization and transfer learning capabilities have not been fully demonstrated [6, 14,16].

Efforts have also been made to develop benchmarks for the legal domain. For example, Chalkidis et al. created LexGLUE [6], a specialized English-language benchmark focused on the EU and US legal systems, to evaluate legal models in NLP. LexGLUE consolidates a range of existing datasets for various tasks, introducing six main tasks, all related to classification: topic prediction of contracts, identification of case holdings, concept identification, contractual terms, violated articles, case issue classification, and unfair identification.

Niklaus et al. contributed LEXTREME [34], a multilingual L-NLP benchmark covering 24 EU languages, with classification-based tasks focused on the jurisdictions of the EU and Brazil. Chalkidis et al. proposed multilingual benchmarking for evaluating the fairness of different models in legal judgment and topic prediction tasks, covering five languages and four jurisdictions [35]. Another approach by Hwang et al. proposed the LBOX benchmark [36], focusing on the legal system of Korea. This benchmark targets classification and summarization tasks using documents in Korean.

Chalkidis et al. also introduced FAIRLEX [35], a multilingual benchmark designed to evaluate fairness in legal text processing. Unlike the aforementioned performance benchmarks, FAIRLEX focuses on measuring bias and performance disparities across sensitive attributes. It comprises four datasets covering four jurisdictions (European Council, USA, Switzerland, and China) and five languages (English, German, French, Italian, and Chinese), evaluating whether models exhibit fairness across demographic attributes (gender, age, nationality), languages, and legal areas. This benchmark addresses a critical but distinct dimension of legal NLP which ensures equitable model performance across different groups and complementing general performance evaluation frameworks like LexGLUE.

Recently, Guha et al. introduced LegalBench [8], a large collaborative legal benchmark focused on the US legal system. It includes 162 tasks in English for testing the reasoning capabilities of LLMs, classified into six types of legal reasoning and corresponding to various stages in the litigation process. LegalBench primarily evaluates the aptitude of LLMs in handling legal processes, with documents averaging about 200 words.

To benchmark LLMs for Chinese law, Fei et al. released LawBench [11], a set of 20 tasks in Chinese designed to test LLMs' ability to memorize legal knowledge and understand the legal domain. Most of the texts in LawBench are much longer than those in LegalBench, averaging around 300 words.

Table 1 provides a summary view of the considered legal benchmarks based on jurisdictions, legal system orientations, supported task types, and languages.

Beyond these comprehensive benchmarks, various research contributions have also introduced specialized datasets tailored to specific legal NLP challenges and methodologies, as shown in Table 2. Further solutions in the field of L-NLP have been released, usually attached with specialized datasets used for testing the proposed methods on focused case studies.

Ren et al. introduce PatentMatch [37], a dataset designed to facilitate the matching of patent claims with prior art. Comprising 6 million labeled pairs, this dataset significantly enhances the ability of models to accurately match claims to relevant text passages in existing patents, thereby supporting improved patent examination processes. Ren et al. also worked on PatentGPT [38], a language model optimized for drafting patents. This model, which makes use of knowledge-based fine-tuning, combines large-scale training data from tens of millions United States Patent and Trademark Office (USPTO) patent texts and knowledge graph to show that it is possible to use it represent different patent histories in various industries. In another work, Ren et al. present PatCID [39], a dataset containing 81 million chemical structure images extracted from patents. This resource aims to surpass previous molecular retrieval databases in both coverage and quality, providing researchers with valuable tools for drug design and material science.

Ma et al. present a dataset focused on Korean legal language recognition, including 147,000 judicial precedents designed to evaluate judicial responsibilities such as judgment forecasting and summarization, and to promote the utilization of NLP in Korean legal applications [40]. Zuo et al. introduce PatentEval [41], a benchmark to evaluate the performance of automatic or semi-automatic patent generation systems, analyzing common error types and offering insights into the relative strengths and weaknesses of various LLMs.

Östling et al. introduce the Cambridge Law Corpus [42], a dataset containing over 258,000 UK court cases, which aims to advance legal AI research by providing information on case law and contributing to the development of legal reasoning models. Caldarola et al. present LAWSUIT [43], a dataset of 14,000 summaries of Italian court verdicts written by legal experts, designed to enhance summarization tasks in the legal domain and support the development of NLP models for generating concise legal summaries. Finally, Santos et al. explore the challenges and approaches for multilingual legal reasoning, proposing a multilingual benchmark for LLMs to support legal tasks in diverse contexts [44].

Table 1
Summary view of L-NLP benchmarks.

Ref	Dataset	Jurisdictions	System	Task types	Languages
[11]	LawBench	China	Civil law	Classification, Generation, Extraction	Chinese
[8]	LegalBench	Multiple	Common & civil law	Generation	English
[36]	LBOX	Korea	Civil law	Classification, Generation	Korean
[35]	FAIRLEX	E.U., U.S., China, Switzerland	Predominantly civil law	Fairness evaluation	E.U., Chinese
[34]	LEXTREME	E.U., Brazil	Predominantly civil law	Classification	E.U.
[6]	LexGLUE	U.S., E.U.	Predominantly civil law	Classification	English

Table 2
Summary view of L-NLP approaches and datasets.

Ref.	Title	ML approach	Dataset	Method
[37]	PatentMatch & Prior Art	Supervised learning, NLP, LLM	6 million labeled pairs for patent claim matching	Model training to enhance patent examination
[38]	PatentGPT	Fine-tuning with knowledge graphs, LLM	Tens of millions of USPTO patent texts and knowledge graphs	Knowledge-based fine-tuning for patent drafting
[45]	AI in Patent Field	Review of AI in patent classification	Various patent datasets for classification tasks	Analysis of AI's evolving role in patent law
[39]	PatCID	Multimodal learning, Image retrieval	81 million chemical structure images from patents	Retrieval tasks for molecular and drug design
[40]	Korean Legal Language Understanding	NLP, LLM for legal tasks	147k Korean judicial precedents	Evaluating legal judgment prediction and summarization
[41]	PatentEval	Evaluation benchmark for patent generation	Patent generation systems	Error typology analysis and evaluation of LLM performance
[42]	The Cambridge Law Corpus	Supervised learning, NLP, LLM	258,000 UK court cases	Legal reasoning and case outcome prediction
[43]	LAWSUIT	Supervised learning, NLP	14,000 Italian legal verdicts	Summarization tasks for legal verdicts
[44]	One Law, Many Languages	Multilingual legal reasoning, LLM	Multilingual legal datasets from the Swiss system	Multilingual legal tasks and LLM evaluation

Table 2 provides a summary view of L-NLP research approaches, highlighting specialized datasets introduced alongside novel methodologies addressing specific legal NLP challenges.

Among the available legal benchmarks, LexGLUE was selected for evaluating ASKE due to several methodological and practical considerations. First, LexGLUE is composed entirely of English datasets, aligning with ASKE approach current language capabilities. This excludes multilingual alternatives such as LEXTREME (24 EU languages), LawBench (Chinese), and LBOX (Korean), which would require multilingual processing capabilities beyond ASKE current scope. Second, LexGLUE emphasizes classification tasks—six of its seven datasets involve multi-label or multi-class classification—which directly align with ASKE core functionality of knowledge extraction and document categorization. In contrast, LegalBench focuses on LLM reasoning capabilities through 162 prompt-based tasks with short documents (averaging 200 words), which do not adequately test ASKE iterative knowledge extraction on longer legal texts. Third, while FAIRLEX provides valuable fairness evaluation capabilities across multiple jurisdictions and languages, it measures performance disparity across demographic groups rather than task performance, making it incompatible with ASKE evaluation objectives. Finally, LexGLUE offers diverse document types and lengths across multiple legal domains (EU law, US case law, contracts), enabling comprehensive assessment of ASKE adaptability, and provides well-established baselines from multiple SOTA models for direct performance comparison. These factors collectively make LexGLUE the most suitable benchmark for systematically evaluating ASKE legal knowledge extraction capabilities.

3. Overview of ASKE

In this section, we outline the ASKE approach for classifying and extracting legal knowledge from large corpora of legal documents.

Given a corpus of legal documents, the goal of ASKE is to extract a set of relevant concepts from documents with related sentences/paragraphs (called *document chunks*) that are associated, namely classified, with respective concepts. The result of this classification/extraction process is a concept view of the considered corpus, where a concept can be used to access the document chunks where that concept is actually appearing. ASKE is characterized by the use of *zero-shot learning techniques* and the adoption of *context-aware embedding models*. Zero-shot learning is a well-known unsupervised learning modality and it refers to the ASKE ability to perform classification tasks without any training. This solution is particularly appropriate to deal with situations in which labeled datasets are not available, which is the typical case of the legal field. ASKE is also characterized by the use of an initial set of *seed concepts* to initialize the classification/extraction process. A seed concept is a textual description (e.g., one or two phrases) of a topic of interest to consider for classification. Such a concept description can be manually defined by an interested user, or can be extracted from an external knowledge base that can be general-purpose (e.g., Wikipedia) or domain-specific (e.g., Westlaw). Concept extraction in ASKE is designed as an iterative, cyclic pipeline. Initially, ASKE classifies the documents according to the seed concepts. The results of the classification are analyzed with the aim to derive/extract additional concepts and to enrich the view with more concepts appearing in the documents. In terms of zero-shot learning, the use of seed concepts does not represent a limitation since it is not introducing the need of a labeled dataset for training.

A technical presentation of ASKE can be found in [12]. In the following, we provide a breakdown and description of each key step of ASKE, as shown in Fig. 1.

Data preparation. This step has the goal to build a unified, context-sensitive semantic space where both the documents of the considered corpus and the seed concepts are framed and represented. This is done by relying on a context-aware embedding model, and a shared vector space is generated as a result. The choice of using a context-aware embedding model has the goal aims to appropriately capture and manage the meaning of legal terminology by taking into account the context in which terms are used. This is particularly relevant in the legal field where technical terms are employed and some terms

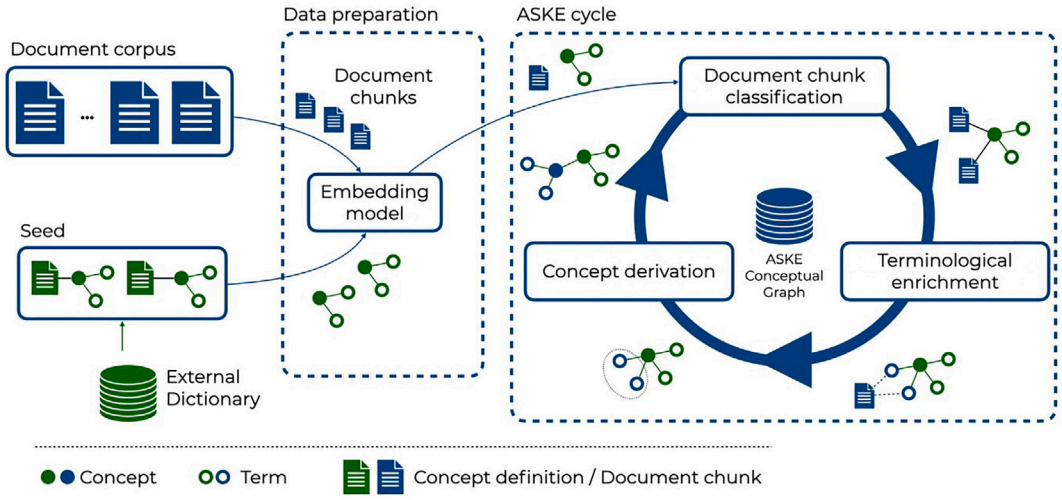

Fig. 1. Overview of the ASKE approach [12].

may also assume completely different meaning when used in different portions of a document like a court decision or a lawyer argument. ASKE is capable of supporting different embedding models, spanning from generic transformer-based models like BERT, to domain-specific pre-trained models like Legal-BERT.

In data preparation, each document in the corpus as well as each seed concept description is preprocessed using a tokenizer that splits the text in a set of document chunks, providing a lemmatized version of the terms therein contained. A document chunk should correspond to a logical unit within the text whose length can provide enough information content for its practical usage in application contexts. Moreover, the dimension of document chunks should fit the constraints of the adopted embedding model which generally works with a fixed-size window (e.g., 512 tokens). Considering the writing style featuring legal documents where sentences are generally rich and articulated, a single sentence is a good candidate to become a document chunk.

Document chunk classification. This step has the goal to associate document chunks to pertinent concepts. To this end, ASKE employs zero-shot, multi-label classification where the association of a chunk with a concept depends on their degree of similarity considering their embeddings generated during data preparation. The cosine metrics is exploited to calculate the chunk-concept similarity and a threshold is defined to decide when the chunk represents/contains an instance of the concept. As a result, it is possible to associate a chunk with multiple concepts (i.e., multi-label classification). This is particularly useful in legal documents, where a single chunk can contain references to multiple concepts or legal issues simultaneously.

Terminological enrichment. The goal of this step is to refine the definition of the considered concepts with additional terms taken from the chunks. The intuition is that the chunks classified as instances of a concept can provide domain-specific terms to enrich the concept specification to better capture the context where the concept is used. Given a chunk classified as instance of a concept, all the terms in the chunk are candidate to enrich the concept specification. For each term, the similarity between the term description and the concept description is calculated using their associated embeddings. A threshold is used to decide when to include a term in the specification of a concept for enrichment.

Concept derivation. The goal of this step is to consider the possible creation of new concepts for classifying the document chunks. The intuition behind ASKE is that the concept specifications changed due to terminological enrichment, then more fine-grained concepts can emerge and could be used to provide a more precise classification of chunks and related documents. ASKE generates a new concept through derivation from an existing one. For each concept, the creation of possible new concepts in ASKE is enforced by clustering the embedding vectors of terms featuring the concept description. The Hierarchical Clustering (HC) algorithm is adopted to this end, since it allows to detect the emergence of sub-groups of similar terms within the concept specification, without requiring to "a-priori define" the number of clusters to generate. A new concept is created for each cluster returned by HC on the terms of a concept. A derivation link is defined between the original concept and the new one to denote that they are somehow similar/related in content.

The ASKE cycle. The set of concepts obtained after derivation can trigger the execution of a new ASKE cycle based on *document chunk classification*, *terminological enrichment*, and *concept derivation*. New derived concepts can contribute to improve the classification of chunks with more fine-grained concepts. Further new concepts can be also discovered through a new execution of enrichment and derivation on the basis of a refined classification result. The ASKE process terminates after the execution of a given number of iterations when concept derivation does not provide the creation of new useful concepts.

ASKE at work on a concrete example. Fig. 2 illustrates document chunks, extracted terms, and concepts from a representative Terms-of-Service (UNFAIR-ToS) contract. On the left, document chunks are shown from the contract text. In the center, key legal terms are highlighted in bold as they appear within these chunks and serve to enrich conceptual terminology. On the right, terms are grouped into hierarchical concepts: the account intervention system at the top level, with breach enforcement, notification and exception, and liability exemption as subcategories. Derivation relationships visualize how detailed obligations and risks are organized under broader system-level concepts across ASKE refinement cycles.

This example demonstrates how effectively ASKE breaks down and organizes complex contract language. By clearly linking phrases from the contract to specific legal concepts, ASKE makes it possible to identify risks and check for compliance with greater accuracy. This level of transparency and careful structure is essential for reliable and advanced automation in analyzing consumer contracts.

4. The LexGLUE datasets

The LexGLUE datasets have been selected for the evaluation of ASKE on legal text analysis. The LexGLUE benchmark comprises seven datasets covering diverse legal NLU tasks. In this study, we evaluate ASKE on six of the seven LexGLUE datasets: ECtHR (Tasks A and

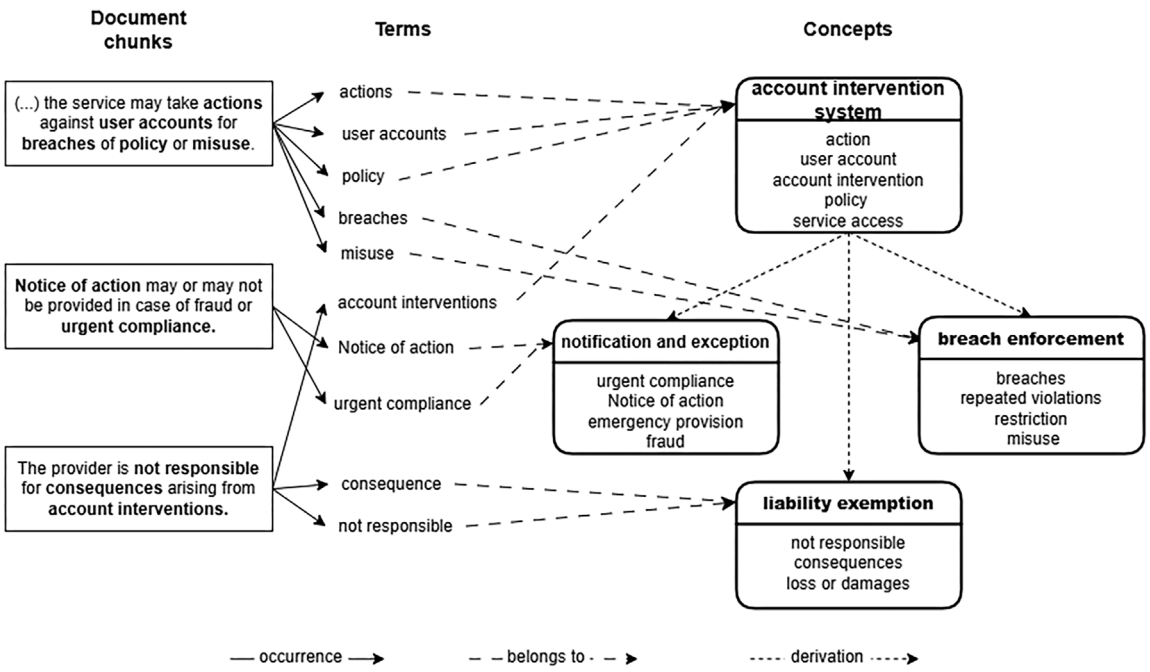

Fig. 2. Example of ASKE knowledge model instance.

Table 3
Overview of the LexGLUE datasets with modifications.

Dataset	Task type	Train/Dev/Test	Ref.	Sub-domain	Classes
LEDGAR	Multi-class	60K/10K/10K	[46]	Contracts	100
UNFAIR-ToS	Multi-label	5.5K/2.2K/1.6K	[47]	Contracts	8+1
EUR-LEX	Multi-label	55K/5K/5K	[48]	EU law	100
SCOTUS	Multi-class	5K/1.4K/1.4K	[49]	US law	14
ECtHR (Task A)	Multi-label	9K/1K/1K	[50]	ECHR	10+1
ECtHR (Task B)	Multi-label	9K/1K/1K	[51]	ECHR	10+1

B), SCOTUS, EUR-LEX, LEDGAR, and UNFAIR-ToS. We exclude CaseHOLD from our experiments because it is a multiple-choice question-answering task that requires identifying correct legal holdings from five candidate statements. This task format is fundamentally incompatible with ASKE knowledge extraction methodology, which is designed for document classification through iterative terminology enrichment and concept derivation. CaseHOLD evaluates reading comprehension and legal reasoning in a QA format, whereas ASKE focuses on zero-shot multi-label classification by extracting concepts from documents. The remaining six datasets all involve classification tasks (either multi-label or multi-class), making them suitable for assessing ASKE's ability to extract legal concepts and classify documents based on enriched terminology across diverse legal domains and jurisdictions. Table 3 provide key information about each dataset and the associated tasks.

SCOTUS. The SCOTUS dataset maps US Supreme Court opinions to 14 legal issue areas, such as civil rights, criminal procedure, and economic activity. The Supreme Court is the highest federal court in the United States, primarily hearing cases that are among the most controversial or difficult, which lower federal courts have not been able to resolve adequately. This dataset is built using the Supreme Court Database (SCDB) [49], which provides detailed metadata on cases decided between 1946 and 2020, including decisions, issues, and directions.

The dataset contains 7800 instances, divided chronologically into training (5K instances from 1946–1982), development (1.4K instances from 1982–1991), and test (1.4K instances from 1991–2016) sets. This chronological split reflects the natural evaluation of legal issues and ensures that the models are evaluated on cases that mirror more contemporary court rulings. Each case is assigned a single label, corresponding to one of 14 issue areas, covering a total of 278 different legal topics. The labels indicate the basic topic of each case, serving as a guide to the nature of the disputes.

This dataset provides a deeper understanding of how Supreme Court decisions are categorized and how legal issues have evolved over time. It offers a structured way to explore the complexities of judicial reasoning and the shifting priorities of the nation's highest court.

UNFAIR-ToS. The UNFAIR-ToS [47] dataset contains instances of linguistic descriptions of unfair contractual terms from the Terms of Service agreements of 50 well-known online platforms (e.g., Facebook, eBay, YouTube). These terms feature clauses that can seriously breach consumer rights under EU law. The dataset is labeled at the sentence level with eight classes of unfair terms, allowing for a fine-grained analysis of potentially exploitative language in user agreements. The dataset contains 9500 sentences, divided chronologically into training (5532 sentences from 2006–2016), development (2275 sentences from 2017), and test (1607 sentences from 2017) subsets. Each sentence is tagged with one or more labels indicating unfair terms, reflecting the potential ways in which user rights may be undermined.

This dataset emphasizes the language and structure of Terms of Service agreements that display patterns of unfair clauses. As such, it could prove to be a valuable tool for enhancing transparency and fairness in such documents.

LEDGAR. The LEDGAR [46] dataset focuses on classifying contract provisions into 100 unique categories. These provisions cover a wide range of topics, as can be seen from the selection of publicly available contracts. As such, the dataset is particularly valuable for understanding both the structure and content of legal agreements. It contains a total of 80,000 provisions, divided into training (60,000 provisions), validation (10,000 provisions), and test (10,000 provisions) subsets. This division allows for robust training and evaluation of models, ensuring generalization to new data.

The structured labels provide insights into the thematic prevalence of topics as well as their sequential occurrence within contracts. This is particularly valuable for legal experts and researchers seeking to streamline the analysis of contracts. The ability to categorize provisions into clear topics facilitates a better understanding of legal language,

especially when dealing with the diversity and complexity typical of legal agreements.

EUR-LEX. This dataset classifies EU legislation using EUROVOC (hierarchical vocabulary of 6900 concepts across 8 levels) [48]. LexGLUE uses 100 most frequent concepts on 65,000 documents: training (55K,1958–2010), development (5K,2010–2012), test (5K,2012–2016). Temporal split introduces concept drift, reflecting realistic generalization. EUR-LEX documents are longer, more complex than US case law with specialized civil law terminology. Averaging 2.2 labels per document, multi-label task reflects how EU legislation addresses multiple policy areas. Valuable for evaluating ASKE on long, complex texts with overlapping concepts.

ECtHR Tasks A and B. The ECtHR hears cases of alleged violations of the human rights provisions set forth in the European Convention on Human Rights (ECHR) by a state. In our work, we utilize the dataset generated by Chalkidis et al. [50]. This dataset includes approximately 11,000 cases from the open ECtHR database. The datasets from the European Court of Human Rights are related to two multi-label classification tasks. The first task involves identifying which articles of the ECHR were violated in the case, while the second focuses on predicting which articles the court considered during deliberation. These tasks are crucial for legal professionals, as they directly relate to the understanding and interpretation of human rights law in specific contexts.

The dataset comprises a total of 11,000 cases, evenly split into training (9000 cases), development (1000 cases), and test (1000 cases) subsets. This methodical composition allows for effective training and evaluation while ensuring generalization to previously unseen data. The cases include detailed factual paragraphs, providing a foundation for predictions and reflecting real-life legal complexities.

5. Models evaluated

The evaluation of ASKE has been performed by considering a set of SOTA pre-trained transformer models.

We fine-tuned these pre-trained models on the various LexGLUE tasks, specifically multi-label and multi-class classification. The second set of models, which includes the pre-trained transformer models, was used to evaluate the performance of ASKE. These models were chosen because they represent the challenges unique to legal texts, such as processing long documents, dealing with imbalanced datasets, and correctly interpreting fine-grained legal terms.

BERT [52] is a pre-trained language model based on masked language modeling and next-sentence prediction tasks during its pre-training phase. BERT has set the SOTA standard for diverse NLP tasks, excelling in contextual language understanding. BERT selected as a general-purpose baseline representing standard contextual language models in NLP. Its widespread adoption makes it a reference point for comparison. Its primary strengths lie in its ability to generate rich contextual representations at both sentence and paragraph levels, making it particularly effective for moderate-length texts up to 512 tokens.

RoBERTa [53] is an improved version of the BERT model. It discards the next-sentence prediction (NSP) task, utilizes dynamic masking, and pre-trains on larger datasets with a broader vocabulary. These enhancements lead to significant improvements across various NLP tasks. RoBERTa was selected because it is built on BERT's design by using more training data and dynamic masking, which often leads to better accuracy in complex language tasks. It is strong at understanding context in shorter legal texts, but like BERT, it cannot process documents longer than 512 tokens.

DeBERTa [54] is an advanced version of BERT. It incorporates disentangled attention mechanisms to better model token contents and positional relationships. Additionally, DeBERTa introduces a refined mask decoder, which outperforms both BERT and RoBERTa on several NLP tasks. DeBERTa was included because it introduces advanced attention and embedding techniques that further enhance context understanding compared to BERT and RoBERTa. This makes it well-suited for capturing subtle or complex language patterns often found in legal texts.

Longformer [55] is specifically designed for processing longer documents. It combines local and global attention mechanisms, using sparse attention to handle sequences with up to 4096 tokens. This model is well-suited for tasks that require long-range context. This feature is particularly valuable for analyzing statutes, contracts, or case files that exceed the input limits of standard transformers.

Legal-BERT [5] is a pre-trained BERT model trained on legal text datasets, such as contracts, legislation, and court opinions. It has achieved SOTA performance on various tasks involving complex legal language, thanks to its specialized vocabulary focused on legal terminology. Legal-BERT was included because it is pre-trained on a large collection of legal texts, allowing it to better understand domain-specific terminology and legal language patterns. By leveraging this specialized pre-training, Legal-BERT is expected to capture the nuances and subtleties of legal documents more effectively than general-purpose language models.

CaseLaw-BERT [56] is specifically tailored for the U.S. case law domain. It was pre-trained on over 3.4 million court documents, making it particularly effective for judicial opinions and case-specific corpus contexts. CaseLaw-BERT was selected for its specialized training on judicial opinions and case law, enabling it to recognize the language, structure, and context unique to legal reasoning and litigation. This specialization helps the model more accurately interpret and classify texts from court cases or related legal documents.

6. Experimental results and discussion

In this section, we first describe the setup of the experiments, then we discuss the obtained results in terms of F1, accuracy, precision, and recall. A public repository with scripts and complete experimental results is available at https://github.com/oziofficial5/Aske.

6.1. Experimental setup

In setting up the tools for the experiments, we tested various configurations and we adopted an empirical strategy by reporting and commenting the setup that provided the best results. All the experiments have been executed on our lab workstations with NVIDIA A100 GPUs, incorporating mixed-precision training to accelerate computation while maintaining numerical stability.

In particular, ASKE has been configured as follows. Documents and seed concept descriptions have been segmented in chunks of 512-tokens and Legal-BERT has been used for subsequent embedding in data preparation. The LexGLUE task labels have been considered as initial seed concepts to bootstrap zero-shot classification without any task-specific fine-tuning. Three ASKE cycles have been executed to refine the set of concepts to describe the given datasets and the related classification of extracted chunks. The ASKE thresholds have been tuned on the development sets; the threshold used in cosine similarity for concept-chunk classification has been set to 0.75 to optimize both precision and recall.

Baseline models, using BERT and RoBERTa, were fine-tuned for 10 epochs with the AdamW optimizer (learning rate of 2e−5, batch size of 16). To address class imbalance in multi-label tasks like UNFAIR-ToS, label-weighted loss functions were employed, ensuring that minority classes played a significant role in the learning process.

To prevent temporal data leakage, we strictly adhered to LexGLUE's pre-defined chronological splits. For example, the SCOTUS dataset was split temporally, with the training data spanning 1946–1982 and the

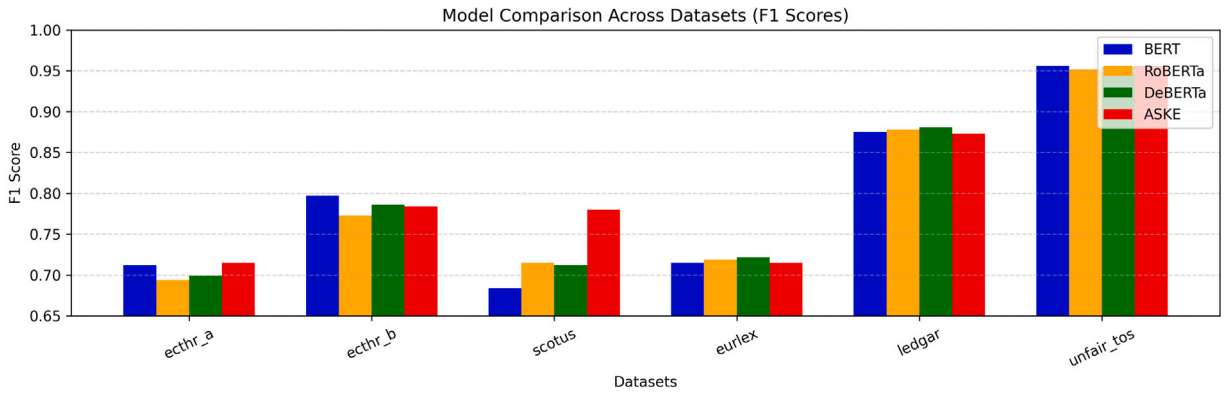

Fig. 3. F1 results on SOTA models over LexGLUE datasets.

test data spanning 1991–2016. This method simulates how models would operate in actual legal practice, where they must generalize to future cases.

Performance is evaluated using task-specific metrics. Macro-F1, precision, and recall were employed for both classification tasks to account for class imbalance, while accuracy was used as an additional measure for single-label scenarios. The quality of the embeddings is assessed using cosine similarity scores for semantically similar document pairs, reflecting how well the ASKE approach captured latent legal relationships. Statistically significant results are then verified using paired t-tests ($p < 0.05$) to ensure that any observed differences were not due to random variation.

6.2. F1 score

The comparison on F1 scores aims at highlighting the ability of the considered approaches at being both accurate and complete in their predictions. In this experiment, ASKE is compared against BERT, RoBERTa, and DeBERTa on ECtHR, SCOTUS, EUR-LEX, LEDGAR, and UNFAIR-ToS. ASKE outperformed the considered models across a diverse set of legal contexts. On the UNFAIR-ToS and LEDGAR datasets, ASKE is comparable to SOTA results (F1 = 95.6 and 87.6, respectively), indicating that the embedding model of ASKE successfully captures the relevant terms in the text to classify chunks with appropriate concepts (i.e., LexGLUE labels). ASKE substantially outperforms conventional models (e.g., TFIDF + SVM) on most datasets, especially on those that are more semantically demanding (e.g., UNFAIR-ToS). Moreover, the results show that ASKE outperforms conventional machine learning models and other transformer based architectures, including BERT, in cases where understanding the language and structure of law is a core issue for the task (see Fig. 3).

6.3. Accuracy

The accuracy results are illustrated in Fig. 4 and demonstrates that ASKE constantly achieves competitive performance across a range of L-NLP benchmarks, thus supporting the idea that also unsupervised methods like the zero-shot learning used by ASKE can provide effective results.

On the ECtHR-A dataset, ASKE achieved an accuracy of 0.70, closely following Longformer (0.73) but trailing the top-performing models such as RoBERTa (0.85), Caselaw-BERT, and Legal-BERT (both 0.83). On the LEDGAR dataset, ASKE records an accuracy of 0.63, demonstrating solid outcomes though not matching Caselaw-BERT's leading 0.86. Its results remain aligned with Legal-BERT and RoBERTa, each around 0.65, and only slightly below Longformer (0.74).

For the UNFAIR-ToS dataset, ASKE achieves an accuracy of about 0.63, competitive with RoBERTa and indicative of reliable generalization, while the highest scores are achieved by Legal-BERT (0.82). In the SCOTUS dataset, characterized by complex judicial reasoning, ASKE reaches 0.78, outperforming Caselaw-BERT(0.66) and closely matching Legal-BERT (0.79) and RoBERTa (0.75).

These findings highlight the robustness and adaptability of ASKE in legal document classification, particularly relevant given its lack of task-specific fine-tuning. Its ability to approximate the accuracy of domain-specialized transformers demonstrates its potential as an interpretable, transferable, and efficient framework for L-NLP, especially in settings with limited labeled data.

6.4. Precision

On precision, experimental results are illustrated in Fig. 5. In comparison with the LEDGAR dataset, ASKE achieves a precision score of 0.83, which is comparable to the performance of RoBERTa (0.85), while outperforming Legal-BERT (0.82), Caselaw-BERT(0.76) and Longformer (0.71). The quality performance on LEDGAR suggests that ASKE is particularly well-suited for datasets that require sensitivity to the nuanced understanding of the context in which legal language is used, particularly in the commercial and financial sectors. The close results to transformer-based models indicate that ASKE could offer exceptional quality for specialized workloads, such as L-NLP tasks.

On the ECtHR-A dataset, ASKE achieves a precision score of 0.72, surpassing RoBERTa (0.62) and closely aligning with Longformer (0.70). In comparison, Caselaw-BERT and Legal-BERT provide marginally lower precision scores of 0.65 and 0.63, respectively.

On the SCOTUS dataset, ASKE scores precision of 0.75, closely following RoBERTa (0.76) and outperforms Longformer (0.62), but scoring slightly lower than Caselaw-BERT (0.83) and Legal-BERT (0.78). This finding indicates that ASKE is well-equipped to process dense legal texts, such as Supreme Court case law, which involve nuanced legal reasoning and sophisticated argumentation. The competitive performance of ASKE suggests it can be valuable for high-level legal tasks, though there is still room for improvement to match the performance of models like Caselaw-BERT, and legal-BERT which are specifically fine-tuned for the legal domain.

The ASKE results on LEDGAR and ECtHR-A demonstrates its effectiveness in tasks requiring a deep understanding of legal language in specialized areas of law. Again, this result confirms that ASKE can generate high-quality embeddings that capture the meaning of legal text, achieving performance levels comparable to transformer-based models across a wide range of legal problem domains.

6.5. Recall

The results on recall are shown in Fig. 6 and they show that some performance differences can be observed depending on the dataset. ASKE scores 0.73 on the ECtHR-A dataset, which is considerably better than Longformer (0.65), Roberta (0.66), and Caselaw-BERT (0.63). This

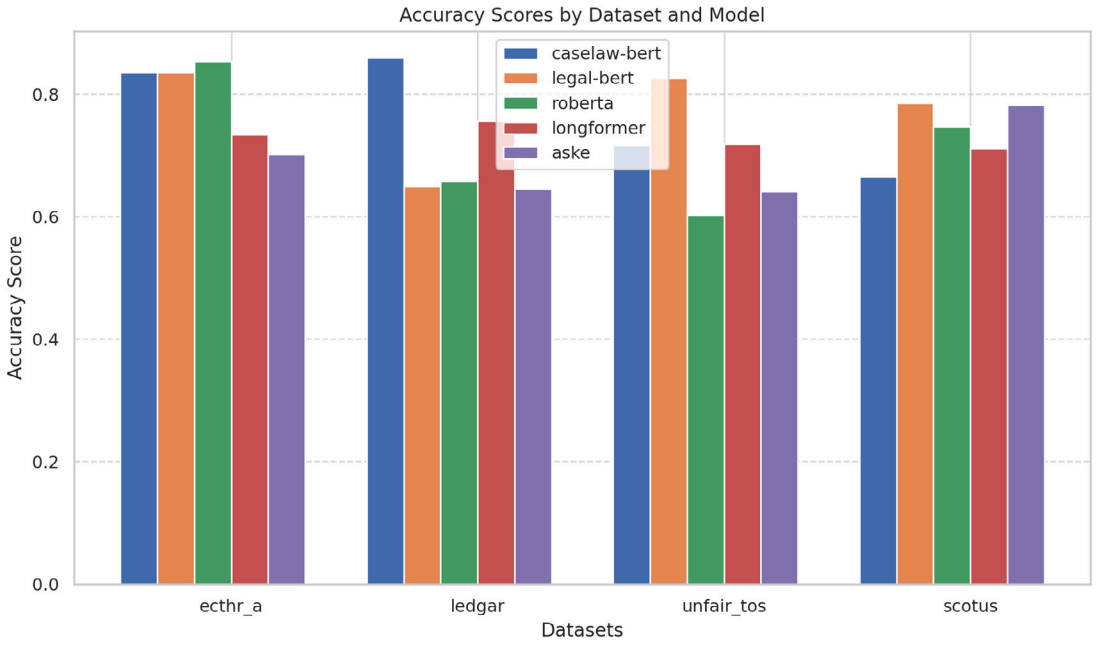

Fig. 4. Accuracy results on SOTA models over LexGLUE datasets.

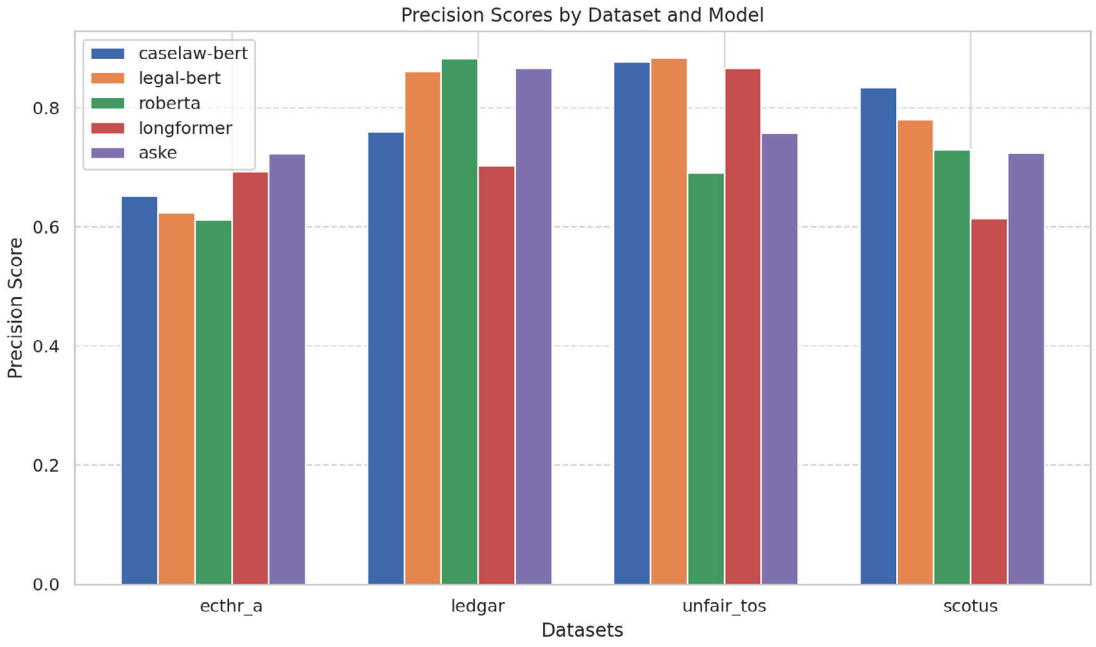

Fig. 5. Precision results on SOTA models over LexGLUE datasets.

suggests that ASKE is effective in retrieving relevant information from this dataset, with quality of results comparable to other considered models, but still falling behind Legal-BERT (0.85). Thus, this result indicates that ASKE is capable of accurately capturing relevant concepts appearing in legal documents, but further improvements are needed to match the outstanding performance of SOTA models in the field.

Within the UNFAIR-ToS dataset, recall analysis shows that ASKE achieves a score near to 0.70, closely aligned with both Caselaw-BERT (0.68) and Longformer (0.69). While RoBERTa demonstrates the highest recall at 0.83, and Legal-BERT follows with a score of 0.75, ASKE maintains robust recall despite not surpassing the leading transformer models.

These results demonstrate that ASKE is effective at retrieving relevant legal information from datasets that contain complex terms of service and extensive legal language, such as those found in Supreme Court cases. The high recall scores in these datasets highlight ASKE's focus on identifying important legal concepts, which are highly relevant to certain legal fields.

6.6. *Embedding similarity scores*

This experiment is focused on evaluating the effectiveness of document (chunk) embeddings in capturing the semantics of the considered legal text. The effectiveness of embeddings is crucial for tasks that require to discover latent similarities across different elements, such as for example document similarity and information retrieval. ASKE demonstrates promising performance on datasets like UNFAIR-ToS and LEDGAR, with mean similarity scores that are high enough to suggest

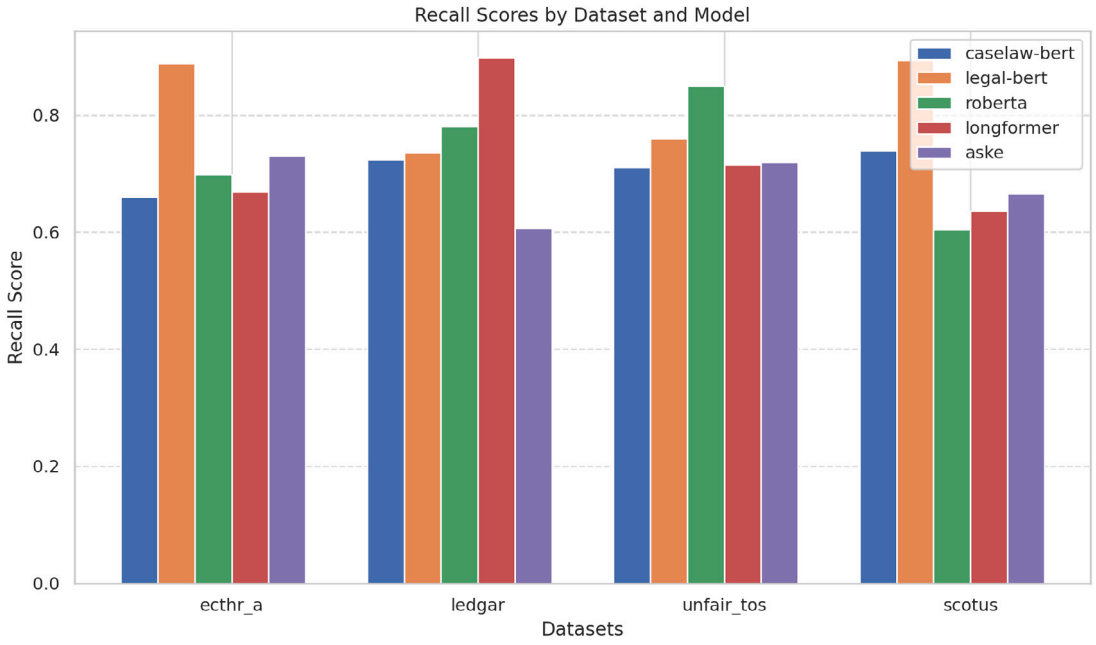

Fig. 6. Recall results on SOTA models over LexGLUE datasets.

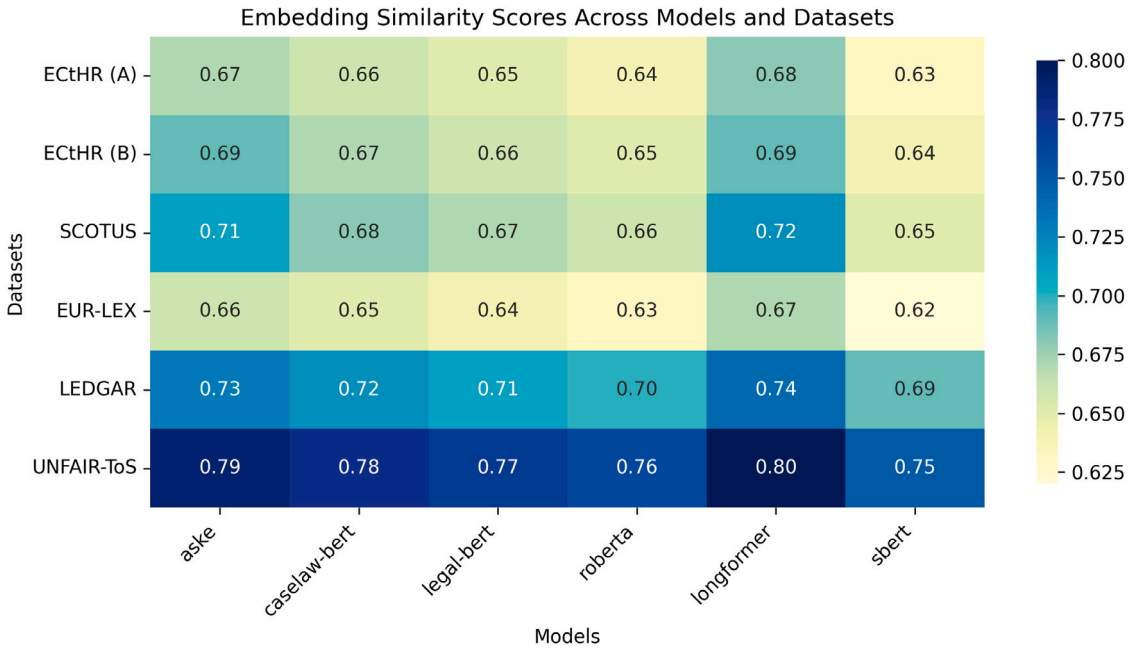

Fig. 7. Embedding similarity scores.

that the approach effectively captures the semantic nuances of legal documents. ASKE's embeddings outperform transformer embeddings, e.g., BERT and RoBERTa, specifically on legal tasks. BERT shows high quality on generic L-NLP tasks, but ASKE shows superior results on domain-specific legal embeddings and therefore can be crucial across specialized fields of L-NLP. ASKE captured legal text semantics and provided effective embeddings which improved its performance on the other downstream tasks like classification and document similarity. They are specifically designed for the legal domain and demonstrate a sophisticated and nuanced understanding of legal terms and concepts (see Fig. 7).

6.7. Performance metrics by models and dataset

Table 4 shows the performance of ASKE compared against the considered models (BERT, RoBERTa, DeBERTa, and Legal-BERT) on the LexGLUE datasets. On the ECtHR-A dataset, ASKE achieves a μ-F1 score of 75.7 and an m-F1 score of 70.3, surpassing all transformer-based baselines, including Legal-BERT and CaseLaw-BERT. A consistent pattern emerges on the SCOTUS dataset, where ASKE achieved a μ-F1 of 77.1 and an m-F1 of 67.8, underscoring its competitiveness with SOTA supervised transformer models in complex legal text classification tasks.

Table 4
Performance metrics by models and dataset for ECtHR and SCOTUS.

Method	ECtHR μ-F1	ECtHR m-F1	SCOTUS μ-F1	SCOTUS m-F1
TFIDF+SVM	69.6	58.4	78.2	69.5
BERT	75.4	68.6	68.3	58.3
RoBERTa	73.2	63.5	71.6	62.0
DeBERTa	74.0	66.9	71.1	62.7
Longformer	74.3	68.2	72.9	64.0
BigBird	74.4	66.9	72.8	62.0
Legal-BERT	75.2	70.2	76.4	66.5
CaseLaw-BERT	74.7	66.8	76.6	65.9
ASKE	75.7	70.3	77.1	67.8

Table 5
Performance metrics by models and dataset for EUR-LEX, LEDGAR, and UNFAIR-ToS.

Method	EUR-LEX μ-F1	EUR-LEX m-F1	LEDGAR μ-F1	LEDGAR m-F1	UNFAIR-ToS μ-F1	UNFAIR-ToS m-F1
TFIDF+SVM	71.3	51.4	87.2	82.4	95.4	78.8
BERT	71.4	57.2	87.6	81.8	95.6	81.3
RoBERTa	71.9	57.9	87.9	82.3	95.2	79.2
DeBERTa	72.1	57.4	88.2	83.1	95.5	80.3
Longformer	71.6	57.7	88.2	83.0	95.5	80.9
BigBird	71.5	56.8	87.8	82.6	95.7	81.3
Legal-BERT	72.1	57.4	88.2	83.0	96.0	83.0
CaseLaw-BERT	70.7	56.6	88.3	83.0	96.0	82.3
ASKE	73.0	58.5	88.7	83.5	96.2	83.5

ASKE performs remarkably well on additional legal datasets as shown in Table 5, reinforcing its superior ability across a variety of legal texts. ASKE achieves a μ-F1 score of 73.0, which is the highest score on EUR-LEX dataset, showing that ASKE generalizes well to the European legal domain, where there may be differences (in structure/terminology) to the other datasets (SCOTUS, ECtHR).

On the LEDGAR dataset, representing financial and regulatory law, ASKE leads with an μ-F1 of 88.7 over the other models, such as DeBERTa (88.2). This supports the idea that ASKE is able to flexibly deal with specialized legal sub-fields. For UNFAIR-ToS, ASKE achieves an excellent μ-F1 score of 96.2 compared to SOTA models such as Legal-BERT (96.0).

7. Possible ASKE applications in L-NLP

The experimental results on the LexGLUE datasets show that ASKE excels in various L-NLP tasks, including *document classification*, *concept extraction*, and *terminology enhancement*. In the following, we outline possible application tasks where ASKE can be effectively employed.

Contract analysis. Contract review often involves ensuring compliance, identifying potential risks, and extracting key information. The ASKE ability to extract pertinent concepts with related legal terminology enables to classify contract clauses, obligations, and essential terms. This ability significantly streamlines the contract review process, reducing the time and resources spent on manual analysis. As a result, legal teams can focus their efforts on addressing critical legal issues, thus increasing overall efficiency and minimizing the risk of overlooking key contractual elements.

Legal research and document review. Legal practitioners often deal with large volumes of legal documents, such as case law, statutes, regulations, and legal opinions, which require efficient retrieval and categorization. The classification abilities of ASKE enable legal practitioners to quickly locate and access relevant information. Furthermore, the extraction of legal concepts supports the identification of pertinent precedents, case law, and legal principles, thereby optimizing the research and review process.

Regulatory compliance and risk management. Legal texts such as regulations, statutes, and contracts often contain intricate language that can be difficult for non-experts to interpret. ASKE can be employed to extract key legal concepts from a document to check the compliance with relevant regulations by identifying regulatory terms and obligations. This functionality can improve compliance checks, mitigate the risk of non-compliance, and reduce the potential for legal penalties.

Legal knowledge extraction and ontology development. Legal ontologies serve as structured frameworks that organize and categorize legal knowledge, thereby facilitating the construction of (semi-)automated legal decision-making systems. ASKE can contribute to the creation of domain-specific knowledge bases, which can be leveraged in AI-driven legal applications such as automated legal assistants, contract negotiation tools, and compliance systems. These knowledge bases are essential for improving the accuracy and reliability of legal technology solutions and ensuring that they align with legal standards and practices.

8. Limitations and future work

At the moment, ASKE can only deal with English documents. The extension of ASKE to multilingual data is currently under consideration.

The experimental results revealed that ASKE is competitive against the considered SOTA models across many tasks. However, it is outperformed by specialized transformer-based models, such as Legal-BERT and CaseLaw-BERT, on tasks that require advanced reasoning or argumentation on articulated legal texts, such as those in the SCOTUS dataset. As a limitation, we observe that ASKE needs to improve when complex and nuanced legal texts are considered.

The need of domain-specific fine-tuning is a limitation for the ASKE generalizability, namely the applicability to different law areas (e.g., banking law, labor, frauds) as well as further knowledge domains other than law. This is a common issue in solutions based on domain-specific models with training data from a focused area. This introduces future directions of research on ASKE. First, ASKE could be combined with SOTA transformer based architectures like BERT or RoBERTa to exploit both the domain specific benefits of ASKE and the global context modeling ability of such architectures. ASKE can leverage this hybridization to provide higher results on complex legal documents. A further interesting research direction is to improve the ASKE interpretability, namely the ability to transparently communicate the motivations behind a certain classification result. The integration of ASKE with explainable AI techniques can enhance the support to decision-making processes, thus encouraging the adoption of ASKE in daily activities of lawyers and legal practitioners.

9. Concluding remarks

In this work, we evaluated our ASKE approach to legal knowledge extraction against reference SOTA models on selected legal datasets from the LexGLUE benchmark. The results revealed that ASKE outperforms the considered models on diverse legal datasets, including ECtHR, SCOTUS, LEDGAR, and UNFAIR-ToS. The ASKE experiments show that baseline models such as TFIDF+SVM are surpassed, while competitive results are provided when SOTA transformer models like BERT and Legal-BERT are considered. ASKE excels in document classification, terminology enrichment, and the derivation of legal concepts, achieving interesting F1 scores and precision on domain-specific datasets. Generalizability and explainability still remain open issues for future research work on ASKE. The goal is to improve the quality of results, with special focus on datasets where complex legal text and advanced legal reasoning are required. Further ongoing work is about the analysis and estimation of ASKE computational costs and scalability to provide guidance on concrete costs in case of a real-world deployment.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Yi W. AI in legal tech: Automating document review and legal research. SSRN Electron J 2025;1–21. http://dx.doi.org/10.2139/ssrn.5195526, URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5195526.

[2] Wang A, Pruksachatkun Y, Nangia N, Singh A, Michael J, Hill F, Levy O, Bowman SR. SuperGLUE: A stickier benchmark for General-Purpose language understanding systems. In: Advances in neural information processing systems 32 (neurIPS 2019). 2019, p. 3266–80, URL https://papers.nips.cc/paper_files/paper/2019/file/4496bf24afe7fab6f046bf4923da8de6-Paper.pdf.

[3] Chalkidis I, Androutsopoulos I, Kampas D. Deep learning in law: Early adaptation and legal word embeddings trained on large corpora. Artif Intell Law 2018.

[4] Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, Bernstein MS, Bohg J, Bosselut A, Brunskill Eea. On the opportunities and risks of foundation models. 2021, arXiv preprint arXiv:2108.07258.

[5] Chalkidis I, Fergadiotis M, Malakasiotis P, Aletras N, Androutsopoulos I. LEGAL-BERT: the muppets straight out of law school. In: Findings of the association for computational linguistics: EMNLP 2020. 2020, p. 2898–904. http://dx.doi.org/10.18653/v1/2020.findings-emnlp.261.

[6] Chalkidis I, Jana A, Hartung D, Bommarito M, Androutsopoulos I, Katz DM, Aletras N. LexGLUE: A benchmark dataset for legal language understanding in english. In: Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers). Dublin, Ireland: Association for Computational Linguistics; 2022, p. 4310–30, URL https://aclanthology.org/2022.acl-long.297/.

[7] Doe J, Smith J. Natural language processing in the legal domain. 2023, p. 1–12, arXiv preprint arXiv:2302.12039 URL https://arxiv.org/abs/2302.12039.

[8] Guha N, et al. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. In: Advances in neural information processing systems. vol. 36, 2024, p. 3220–9.

[9] Joshi A, et al. IL-TUR: Benchmark for Indian legal text understanding and reasoning. 2024, arXiv preprint arXiv:2407.05399.

[10] Niklaus J, et al. Lextreme: A multi-lingual and multi-task benchmark for the legal domain. 2023, arXiv preprint arXiv:2301.13126.

[11] Fei Z, Shen X, Zhu D, Zhou F, Han Z, Zhang S, Chen K, Shen Z, Ge J. Lawbench: Benchmarking legal knowledge of large language models. 2023, ArXiv Preprint arXiv:2309.16289.

[12] Castano S, et al. Enforcing legal information extraction through context-aware techniques: The ASKE approach. Comput Law Secur Rev 2024;52:105903.

[13] Moens M-F, Uyttendaele C, Dumortier J. Abstracting of legal cases: the potential of clustering based on the selection of representative objects. J Am Soc Inf Sci 1999;50(2):151–61.

[14] Joshi A, Paul S, Sharma A, Goyal P, Ghosh S, Modi A. IL-TUR: Benchmark for Indian legal text understanding and reasoning. In: Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: long papers). Bangkok, Thailand: Association for Computational Linguistics; 2024, p. 11460–99. http://dx.doi.org/10.18653/v1/2024.acl-long.618, URL https://aclanthology.org/2024.acl-long.618/.

[15] Jackson P, Al-Kofahi K, Tyrrell A, Vachher A. Information extraction from case law and retrieval of prior cases. Artificial Intelligence 2003;150(1–2):239–90.

[16] Malik A, Sharma R. Natural language processing (NLP)-Powered legal ATMs (LAMs). J Knowl Inf Syst 2024;15(2):123–45. http://dx.doi.org/10.1007/s13132-023-01450-2, URL https://ideas.repec.org/a/spr/jknowl/v15y2024i2d10.1007_s13132-023-01450-2.html.

[17] Kalamkar P, Tiwari A, Agarwal A, Karn S, Gupta S, Raghavan V, Modi A. Corpus for automatic structuring of legal documents. In: Proceedings of the thirteenth language resources and evaluation conference. Marseille, France: European Language Resources Association; 2022, p. 4420–9, URL https://aclanthology.org/2022.lrec-1.470/.

[18] Kapoor S, Henderson P, Narayanan A. Promises and pitfalls of artificial intelligence for legal applications. 2024, p. 1–18, arXiv preprint arXiv:2402.01656 URL https://arxiv.org/abs/2402.01656.

[19] Tran V, Nguyen ML, Satoh K. Building legal case retrieval systems with lexical matching and summarization using a pre-trained phrase scoring model. In: Proceedings of the seventeenth international conference on artificial intelligence and law. 2019, p. 275–82.

[20] Lagos N, Segond F, Castellani S, O'Neill J. Event extraction for legal case building and reasoning. In: International conference on intelligent information processing. Springer; 2010, p. 92–101.

[21] Zhong H, Xiao C, Tu C, Zhang T, Liu Z, Sun M. How does NLP benefit legal system: A summary of legal artificial intelligence. In: Proceedings of the 58th annual meeting of the association for computational linguistics. ACL, Online: Association for Computational Linguistics; 2020, p. 5218–30. http://dx.doi.org/10.18653/v1/2020.acl-main.466, URL https://aclanthology.org/2020.acl-main.466/.

[22] Guha N, Nyarko J, Ho DE, Ré C, Chilton A, Narayana A, Chohlas-Wood A, Peters A, Waldon B, Rockmore DN, Zambrano D, Talisman D, Hoque E, Surani F, Fagan F, Sarfaty G, Dickinson GM, Porat H, Hegland J, Wu J, Nudell J, Niklaus J, Nay J, Choi JH, Tobia K, Hagan M, Ma M, Livermore M, RasumovRahe N, Holzenberger N, Kolt N, Henderson P, Rehaag S, Goel S, Gao S, Williams S, Gandhi S, Zur T, Iyer V, Li Z. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. 2023, ArXiv Preprint arXiv:2308.11462.

[23] A. NV, B. TT. Legal judgment prediction using natural language processing and machine learning. SAGE Open 2025;15(1):1–15. http://dx.doi.org/10.1177/21582440251329663, URL https://journals.sagepub.com/doi/abs/10.1177/21582440251329663.

[24] Wu Y, Lim J, Yang M-H. Online object tracking: A benchmark. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2013, p. 2411–8.

[25] Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. ImageNet: A Large-Scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. IEEE; 2009, p. 248–55.

[26] Zhang A, Wu Y, Pineau J. Natural environment benchmarks for reinforcement learning. 2018, arXiv preprint arXiv:1811.06032.

[27] Cobbe K, Hesse C, Hilton J, Schulman J. Leveraging procedural generation to benchmark reinforcement learning. In: International conference on machine learning. PMLR; 2020, p. 2048–56.

[28] Laskin M, Yarats D, Liu H, Lee K, Zhan A, Lu K, Cang C, Pinto L, Abbeel P. URLB: Unsupervised reinforcement learning benchmark. 2021, arXiv preprint arXiv:2110.15191.

[29] Xu L, Hu H, Zhang X, Li L, Cao C, Li Y, Xu Y, Sun K, Yu D, Yu C, et al. CLUE: A Chinese language understanding evaluation benchmark. 2020, arXiv preprint arXiv:2004.05986.

[30] Kakwani D, Kunchukuttan A, Golla S, Gokul NC, Bhattacharyya A, Khapra MM, Kumar P. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and Pre-trained multilingual language models for Indian languages. In: Findings of the association for computational linguistics: EMNLP 2020. 2020, p. 4948–61.

[31] Hu J, Ruder S, Siddhant A, Neubig G, Firat O, Johnson M. XTREME: A massively multilingual Multi-task benchmark for evaluating Cross-lingual generalisation. In: International conference on machine learning. PMLR; 2020, p. 4411–21.

[32] Wang A, Singh A, Michael J, Hill F, Levy O, Bowman SR. GLUE: A Multi-Task benchmark and analysis platform for natural language understanding. 2018, arXiv preprint arXiv:1804.07461.

[33] Zheng L, Guha N, Anderson BR, Henderson P, Ho DE. When does pretraining help? Assessing Self-Supervised learning for law and the CaseHOLD dataset of 53,000+ legal holdings. In: Proceedings of the eighteenth international conference on artificial intelligence and law. ICAIL'21, New York, NY, USA: Association for Computing Machinery; 2021, p. 159–68.

[34] Niklaus J, Matoshi V, Rani P, Galassi A, Stürmer M, Chalkidis I. Lextreme: A Multi-lingual and Multi-task benchmark for the legal domain. 2023, arXiv preprint arXiv:2301.13126.

[35] Chalkidis I, Pasini T, Zhang S, Tomada L, Schwemer S, Søgaard A. FairLex: A multilingual benchmark for evaluating fairness in legal text processing. In: Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers). Dublin, Ireland: Association for Computational Linguistics; 2022, p. 4389–406.

[36] Hwang W, Lee D, Cho K, Lee H, Seo M. A multitask benchmark for Korean legal language understanding and judgement prediction. 2022, arXiv preprint arXiv:2206.05224.

[37] Ren R, Ma J. PatentMatch: A dataset for matching patent claims & prior art. 2020, ArXiv arXiv:2012.13919 URL https://arxiv.org/abs/2012.13919.

[38] Ren R, Ma J. PatentGPT: A large language model for patent drafting using Knowledge-based Fine-tuning method. 2024, ArXiv arXiv:2406.19465 URL https://arxiv.org/pdf/2406.19465.

[39] Ren R, Smith A. PatCID: An Open-Access dataset for chemical structures in patent documents. Nat Commun 2024. URL https://www.nature.com/articles/s41467-024-50779-y.

[40] Ma J. A Multi-Task benchmark for Korean legal language understanding and judgement prediction. In: NeurIPS. 2022, p. 3521–36, URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/d15abd14d5894eebd185b756541d420e-Abstract-Datasets_and_Benchmarks.html.

[41] Zuo Y, Gerdes K, de la Clergerie E, Sagot B. PatentEval: Understanding errors in patent generation. 2024, ArXiv arXiv:2406.06589 URL https://arxiv.org/abs/2406.06589.

[42] Östling A, Sargeant H, Xie H, Bull L, Terenin A, Jonsson L, Magnusson M, Steffek F. The cambridge law corpus: A dataset for legal AI research. 2024, ArXiv URL https://www.cambridge.org/core/journals/ai-and-law/article/cambridge-law-corpus-a-dataset-for-legal-ai-research/.

[43] Caldarola N, Melilli M. LAWSUIT: a large expert-Written summarization dataset of Italian constitutional court verdicts. 2024, ArXiv URL https://openreview.net/pdf?id=r30thTMcaM.

[44] Santos FDM, Arrieta RA. One law, many languages: Benchmarking multilingual legal reasoning for judicial support. 2024, ArXiv URL https://openreview.net/pdf?id=r30thTMcaM.

[45] Ren R. Artificial intelligence exploring the patent field. 2024, ArXiv arXiv:2403.04105 URL https://arxiv.org/abs/2403.04105.

[46] Tuggener D, von Daniken P, Peetz T, Cieliebak M. LEDGAR: A large-scale multi-label corpus for text classification of legal provisions in contracts. In: Proceedings of the 12th language resources and evaluation conference. Marseille, France: European Language Resources Association; 2020, p. 1235–41.

[47] Lippi M, aw Pał ka P, Contissa G, Lagioia F, Miclitz H-W, Sartor G, Torroni P. CLAUDETTE: an automated detector of potentially unfair clauses in online terms of service. Artif Intell Law 2019;117–39.

[48] Chalkidis I, Fergadiotis M, Kotitsas S, Malakasiotis P, Aletras N, Androutsopoulos I. An empirical study on large-scale multi-label text classification including few and zero-shot labels. In: Proceedings of the 2020 conference on empirical methods in natural language processing. EMNLP, Online; 2020, p. 7503–15.

[49] Spaeth HJ, Epstein L, Segal JA, Martin AD, Ruger TJ, Benesh SC. Supreme court database, version 2020 release 01. 2020.

[50] Chalkidis I, Androutsopoulos I, Aletras N. Neural legal judgment prediction in English. In: Proceedings of the 57th annual meeting of the association for computational linguistics. Florence, Italy: Association for Computational Linguistics; 2019, p. 4317–23.

[51] Chalkidis I, Fergadiotis M, Tsarapatsanis D, Aletras N, Androutsopoulos I, Malakasiotis P. Paragraph-level rationale extraction through regularization: A case study on European court of human rights cases. In: Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies. Online: Association for Computational Linguistics; 2021, p. 226–41, URL https://aclanthology.org/2021.naacl-main.22/.

[52] Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies. 2019, p. 4171–86.

[53] Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V. Roberta: A robustly optimized BERT pretraining approach. 2019, ArXiv Preprint arXiv:1907.11692.

[54] He P, Liu X, Gao J, Chen W. DeBERTa: Decoding-enhanced BERT with disentangled attention. In: Proceedings of the 9th international conference on learning representations (ICLR 2021). 2021, p. 1–20, URL https://arxiv.org/abs/2006.03654.

[55] Beltagy I, Peters ME, Cohan A. Longformer: The long-document transformer. 2020, CoRR arXiv:2004.05150.

[56] Zheng L, Guha N, Anderson BR, Henderson P, Ho DE. When does pretraining help? Assessing Self-Supervised learning for law and the CaseHOLD dataset. In: Proceedings of the 18th international conference on artificial intelligence and law. Association for Computing Machinery; 2021, p. 159–68. http://dx.doi.org/10.1145/3462757.3466073.