Leveraging LLMs and LegalDocML to extract legal interpretations: a case study on UK legislation and case law

Safia Kanwal[1], Livio Robaldo[1], Joseph Anim[1], Davide Liga[2]

[1] School of Law, Swansea University, UK
{safia.kanwal,livio.robaldo,joseph.anim}@swansea.ac.uk
[2] Department of Computer Science, University of Luxembourg
davide.liga@uni.lu

Abstract. The increasing volume and complexity of legal texts pose challenges in bridging legislative language with judicial interpretation. This paper introduces a novel methodology, along with a corresponding tool, that leverages Large Language Models (LLMs) and the LegalDocML format in a two-phase approach aimed at extracting legal interpretations of UK legislation within UK case law. The UK Publication Office (The National Archives, TNA) is the single institution in the world providing all its legislation and case law in LegalDocML format. Therefore, the tool is not currently applicable to other jurisdictions. By linking legal terms to their interpretations, the methodology enhances the precision and scalability of legal document annotations, reducing the time and effort typically required for manual annotation. Evaluation results demonstrate high accuracy in identifying and extracting key phrases, showcasing the methodology's effectiveness in addressing the diverse contextual meanings of legal language. This work offers a scalable solution for enriching TNA's LegalDocML datasets while establishing connections between legislation and case law through semantically aligned terms. The tool source code can be accessed through the GitHub repository https://github.com/SafiaK/Odyssey-Terms-Extraction

Keywords: Legal Interpretations, Large Language Models, LegalDocML, Information Extraction

1 Introduction

The law is a fundamental component of modern society, providing mechanisms for governance, formalizing conflict settlement, and safeguarding individual interests. As society continues to advance, the volume of legal texts required to regulate its processes grows exponentially, presenting unprecedented challenges in efficiently processing this vast body of knowledge.

The primary challenge faced by legal practitioners is arguably the interplay between legislative norms and their *legal interpretation* in context. In most instances, the application of the law (legislation) requires interpreting legislation, which is the purview of the courts. It is this interpretative function that makes case law significant to most, if not all, legal systems. In many cases, the interpretation of a provision within a piece of legislation is what resolves a particular legal issue, whether in or outside the courts. Thus, case law is instrumental to both legal practice and the functioning of legal systems. Through decisions in specific cases, judges elucidate legislative intent, establish precedents, and refine the law, ensuring its consistent application in future disputes. Lawyers, on the other hand, play a crucial role in shaping legal interpretations by presenting arguments to persuade judges toward interpretations of the legislative text that favor their clients; they craft arguments to demonstrate how specific excerpts of legislation should or should not apply to the context under scrutiny, trying to guide judges toward decisions aligned with their clients' interests.

As a result, the work of lawyers involves a quite detailed analysis of previous case law to understand how relevant excerpts of legislative text have been interpreted in the past and, consequently, which arguments have been used (or can be used) to advocate for interpretations that support their client's position. Judges must also undertake this analysis, but for a different purpose: harmonizing legal interpretations, i.e., ensuring that legislative text is consistently applied across similar contexts. The considerable time required to analyze legal interpretations and engage in debates surrounding them often results in trials that can span months or even years. This protracted process not only delays justice but can also likely contribute to inequalities and undermine public trust in the legal system.

This calls for LegalTech solutions capable of linking legislative texts with the case law in which they are interpreted, thereby enabling the development of applications that help lawyers and judges navigate and understand past legal interpretations. At the core of these developments is Natural Language Processing (NLP), which enables the automated extraction, classification, and understanding of legal texts. The rise of Large Language Models (LLMs) has of course opened new opportunities in NLP for LegalTech. However, recent studies have highlighted that general-purpose LLMs still fall short compared to smaller, fine-tuned legal domain models, underscoring the need for domain-specific models and tailored training datasets to address the unique challenges of legal language [7], [14], [20], [15].

Nevertheless, creating a training dataset that captures how excerpts of text are legally interpreted in case law appears to be an enormous challenge, given the complexities of legal language, context-dependent interpretations, and the evolving nature of judicial decisions [1]. These challenges require sophisticated methods that go beyond basic text analysis, demanding models capable of understanding context, resolving ambiguities, and identifying subtle relationships between legislative texts and case law [8]. Establishing a standardized approach to annotating datasets that link legislative texts to their legal interpretations is particularly difficult, if not impossible, due to the nuanced and dynamic nature of legal reasoning.

This paper presents a two-phase methodology, along with a corresponding tool, in which general-purpose LLMs are employed to identify legal interpretations of phrases within legislative texts. The tool was implemented and evaluated using UK legislation and case law; indeed, the current implementation *only* operates on UK legislation and case law, for the reasons explained below. The methodology proposed in this paper aims at *incrementally*, though not necessarily *exhaustively*, building the desired training dataset of legal interpretations, which would be too labor-intensive to create through standard human-centered annotation methodologies for corpus building.

In the first phase, general-purpose LLMs are used to identify relevant excerpts of text, specifically to *filter* the original legal documents in order to focus on the excerpts of text that convey legal interpretations, while excluding conventional phrases included for formal reasons, as well as reported background facts and other non-substantive content. The result is a set of pairs "(`lg`, `cl`)", where `lg` represents an excerpt of text from a legislative act, and `cl` is an excerpt of text from a case law that references the act and likely provides a legal interpretation of `lg`. In the second phase, the general-purpose LLM is prompted to extract the specific phrase in `lg` that is legally interpreted in `cl`, if any.

As mentioned before, the reason why the current implementation of the proposed tool only works with UK legislation and case law is that these are available in XML format, which greatly facilitates the first phase of the methodology. Developing similar tools for other jurisdictions, where legislation or case law are likely available only in PDF or HTML formats, would require an additional module to process these formats. On the one hand, this would require extensive low-level programming to manage the complexities of these formats; on the other hand, it would likely result in a higher error rate in the overall results.

The work presented here is part of the research activity of the Innovate UK project "Odyssey" (see acknowledgments). A key member of the project consortium is The National Archives (TNA)[3], the UK publication office, which is the single publication office in the world that makes all country's primary legislation and jurisprudence accessible online in LegalDocML[4], a.k.a. Akoma Ntoso [16], [17]. LegalDocML is generally thought as *the* XML standard format for the legal domain. It is a Committee specification by OASIS that defines a set of simple, technology-neutral representations of legislative and judiciary documents in XML.

Every UK act may be easily downloaded in LegalDocML from https://www.legislation.gov.uk while every case law from 2003 onward may be downloaded in LegalDocML from https://caselaw.nationalarchives.gov.uk. The LegalDocML files, which were prepared and validated by a team of human annotators at TNA, clearly structure the legal texts into sections, paragraphs, etc., and contain explicit references between the legal documents. We therefore tailored our prototype to work with the LegalDocML files from TNA, while leveraging the information already present within them.

The rest of the paper is organised as follows: The next section introduces the input data, specifically the LegalDocML files from TNA, and discusses how phrases from legislation are legally interpreted in case law, i.e., what we aim for our methodology and tool to extract. The methodology is then presented in Section 3, which contains the core of the research presented in this paper. The next section 4 presents the analysis of experiments and evaluation of the results. After that, the related work is described in Section 5 and the last section 6 concludes this study.

[3] https://www.nationalarchives.gov.uk

[4] https://www.oasis-open.org/committees/legaldocml

2 The input data: the LegalDocML files from The National Archives (TNA)

As explained in the previous section, all UK legislation and case law are publicly available in LegalDocML format through TNA's portals. Each UK act published on the portal https://www.legislation.gov.uk can be downloaded in LegalDocML format simply by appending "`data.akn`" (where "`akn`" stands for "Akoma Ntoso") to the end of the URL. For example, the Child Abduction and Custody Act 1985, accessible online via the first URL in (1), can be downloaded in LegalDocML format by following the second URL in (1).

(1) https://www.legislation.gov.uk/ukpga/1985/60
https://www.legislation.gov.uk/ukpga/1985/60/data.akn

Due to space constraints, we are unable to provide details about the various LegalDocML tags used in the "`data.akn`" file or include substantial excerpts of the act's XML annotation in this paper[5]. However, the XML format is relatively intuitive. By following the second URL in (1), the reader can observe how the XML format neatly organizes legal texts into parts, sections, subsections, etc., each associated with a specific `eId`. The format also explicitly annotates titles (tag `<heading>`), indexes (tag `<num>`), headings, and references (tag `<ref>`), as well as abbreviations (tag `<abbr>`), among other elements. As previously mentioned, the XML structure provided by the LegalDocML format facilitates the straightforward programmatic retrieval of meaningful legal text, a task that would be significantly more labor-intensive in HTML or PDF files.

Similarly, case law can be downloaded in LegalDocML from the portal https://caselaw.nationalarchives.gov.uk, but in a different way. At the bottom of each case law web page, e.g., https://caselaw.nationalarchives.gov.uk/ewhc/fam/2020/3257, a link labeled "Download this judgment as XML" allows users to download the LegalDocML file of the case. However, note that while LegalDocML structures legislative acts into sections (using the `<section>` tag), it structures case law into paragraphs (using the `<paragraph>` tag).

In the LegalDocML files of the acts, certain phrases denoting key concepts within the scope of the act are tagged as `<term>`. For example, the phrase "rights of custody", which denotes a key concept in the Child Abduction and Custody Act 1985, is tagged in the LegalDocML file as follows:

(2)
```
<p>‘‘<term refersTo="#term-rights-of-custody" eId="term-rights-of-custody">rights of
custody</term>’’ shall include rights relating to the care of the person of the child
and, in particular, the right to determine the child’s place of residence;</p>
```

Given that they denote key concepts, many of these phrases are legally interpreted in case law. For example, the phrase "rights of custody" is legally interpreted in the case law "[2020] EWHC 3257 (Fam)", which references the act; the fourth paragraph[6] of the case law, for instance, states that in the Mother's opinion the Father's rights of custody were not breached:

(3) *The Mother opposes the Application on the basis:*

(1) That the children's retention in the UK was not in breach of the Father's rights of custody and so, she says, the retention (or their removal) was not "wrongful" within the meaning of Article 3 of the Convention; alternatively

(2) Etc.

Other paragraphs of the case law may include additional legal interpretations (e.g., it is likely that the Father disputes the Mother's interpretation of his rights of custody), as well as the arguments presented by the lawyers to support these interpretations, and ultimately the judges' decision based on the facts and arguments. It is clear, however, that this case law is highly relevant for legal practitioners who must argue similar cases in court, where the question of whether someone's rights of custody have been violated is at issue. The methodology presented in this paper represents the first step towards the creation of an enhanced repository where the links between key phrases in legislative acts and the paragraphs in case law that legally

[5] The full vocabulary of LegalDocML is available at http://docs.oasis-open.org/legaldocml/akn-core/v1.0/akn-core-v1.0-part2-specs.html.

[6] https://caselaw.nationalarchives.gov.uk/ewhc/fam/2020/3257#para_4

interpret them are made explicit. LegalDocML already includes tags to link acts with relevant jurisprudence[7], which could be utilized to store these connections once identified by the NLP module.

Nevertheless, the example discussed in (2) and (3) is a relatively simple one. First, most of the relevant phrases from UK legislation that are legally interpreted in case law are *not* tagged as `<term>` in the LegalDocML file of the act. TNA defined a set of regular expressions to help annotators identify `<term>`s; for instance, the phrase "rights of custody" in (2) was tagged as a `<term>` because it appears within quotes (`‘‘...’’`) and is followed by the text "shall include". However, the majority of key phrases, such as the two discussed below in this section, do not appear in the act according to a fixed pattern that can be associated with an obvious regular expression. For this reason, TNA annotators do not tag them as `<term>`, even though they should, as these phrases are legally interpreted in at least one case law.

Secondly, and more importantly, contrary to the example in (2) and (3), many key phrases are not repeated verbatim in case law, which makes their identification more challenging. The use of LLMs to identify these phrases is therefore highly promising for developing a recommendation system that can suggest potential `<term>`s to TNA annotators with greater coverage and accuracy than regular expressions. LLMs are capable of *paraphrasing* text, enabling them to effectively identify linguistic variants of the target key phrases.

An example is the phrase "physical, emotional and educational needs" from section 1(3)(b) of the Children Act 1989[8]. This phrase is not tagged as `<term>`, like "rights of custody" in the previous example. Still, it is key for the domain of the act as it is legally interpreted in several case law. One of these is "[2024] EWHC 17 (Fam)", specifically its 66th paragraph[9]:

(4) *By contrast to the position of a German court seised of proceedings, whilst the parties have engaged in proceedings in this jurisdiction concerning X's welfare, in the current circumstances, the English court would not have as easy access to the educational and health care professionals engaged with X, and the information concerning his physical, educational and emotional welfare, that will most fully inform the assessment of X's best interests. Etc.*

Note that the phrase "physical, emotional, and educational needs" is paraphrased as "physical, educational, and emotional welfare" in this case law. Nevertheless, both expressions clearly refer to the same concept, which is legally interpreted in (4): the judge determined that the information concerning physical, emotional, and educational needs is not easily accessible to the English court in the context of the trial.

Identifying links such as the one between "physical, emotional, and educational needs" and (4) using regular expressions would be too difficult, if not impossible. By contrast, LLMs are capable of making these connections, as demonstrated below in this paper.

A third final example is the phrase "controlling or coercive behaviour" from section 1(3)(c) of the Domestic Abuse Act 2021[10], which is legally interpreted in paragraph 115 of "[2023] EWHC 2983 (Fam)"[11]:

(5) *I find that the Father did coerce the Mother into travelling to the UK and signing documents with the effect of fraudulently procuring UK tax credits and that this constituted financial abuse of a controlling and coercive nature. Etc.*

In the judge's opinion, what the Father did can be categorized as "financial abuse of a controlling and coercive nature", which contextualizes the phrase "controlling or coercive behaviour" within the legal discussion of the trial. Once again, LLMs are currently the single available technology capable of recognizing the link between these two excerpts of text.

2.1 Selected case law and corresponding acts

In this paper, we focus on Family Law cases from the past five years, specifically from 2020 to 2024. Running our developed tool on *all* case law from https://caselaw.nationalarchives.gov.uk is considered future work.

[7] Specifically, the `<judicial>` tag, see http://docs.oasis-open.org/legaldocml/akn-core/v1.0/os/part2-specs/os-part2-specs_xsd_Element_judicial.html.

[8] https://www.legislation.gov.uk/ukpga/1989/41/section/1#section-1-3

[9] https://caselaw.nationalarchives.gov.uk/ewhc/fam/2024/17#para_66

[10] https://www.legislation.gov.uk/ukpga/2021/17/section/1#section-1-3

[11] https://caselaw.nationalarchives.gov.uk/ewhc/fam/2023/2983#para_115

Furthermore, we only considered case law that references at least one UK act. Other cases that reference only case law or other secondary materials are excluded for simplicity, as including them would require the implementation of an additional module to identify which UK acts they (indirectly) reference. Table 1 shows the breakdown of cases containing legislative references by year.

Year	2020	2021	2022	2023	2024	**Total**
N. of cases	26	22	29	80	40	**197**

Table 1: Cases processed per year

As mentioned earlier, the LegalDocML format structures case law into `<paragraph>`s. These may contain `<subparagraph>`s. However, during initial experimentations, we found that `<subparagraph>`s often lacked sufficient context when analyzed in isolation. Therefore, we selected `<paragraph>` as the primary unit for data processing in our pipeline to maintain context and ensure accurate analysis. A similar rationale applies to the LegalDocML files of the acts, where `<section>` was chosen as the primary unit for data processing.

For each `<paragraph>`, our developed tool must determine whether the `<paragraph>` legally interprets a concept denoted by a phrase occurring within the UK legislation. To this end, as will be explained in the next section, each `<paragraph>` is linked to a `<section>` of a UK act referenced in the case law. As mentioned earlier, we only consider case law that references at least one UK act, ensuring that each `<paragraph>` will be associated with a `<section>`. Several `<paragraph>`s also contain explicit references to sections or subsections of a UK act through the LegalDocML tag `<ref>`. These `<ref>`s will, of course, be utilized by the module that associates a `<section>` with each `<paragraph>`, as the search will be restricted to only those sections mentioned in the `<paragraph>` (if any).

3 Methodology

The methodology detailed in this research is structured around leveraging LLM-based NLP techniques to identify and extract legal terms from UK legislation that are interpreted in case law. Legal interpretation, in this context, refers to the process of clarifying, explaining, or applying legislative language to specific case contexts.Our approach dynamically identifies interpretations without without relying on predefined labels, extracting structured relationships from texts, and incrementally building a dataset of pairs in the form (`phr`, `cl`), where `phr` is a phrase occurring in legislation (typically a short noun phrase) and `cl` is a `<paragraph>` from the LegalDocML file of a case law.

Our methodology consists of two Phases. In the first Phase, we filter the `<paragraph>`s of the input case law to extract only those that may contain a legal interpretation; these are then paired with a `<section>` of the UK acts referenced in the case law. The second Phase then extracts the specific phrase from the `<section>` that is legally interpreted in the case law.

By leveraging advanced LLM capabilities, the methodology addresses the complexity of legal language and its nuanced contextual relationships. Specifically, our approach builds on the advancements of Large Language Models (LLMs), integrating few-shot learning and chain-of-thought [21] reasoning to facilitate filtering of key phrases (Phase 1) and the extraction tasks (Phase 2).

3.1 Phase 1: matching `<paragraph>`s from case law with `<section>`s from UK acts

As explained in the Introduction above, the LegalDocML format eliminates the need to pre-process the input documents, which is required in non-UK jurisdictions where legislation and case law are only available in PDF and HTML formats. From a LegalTech perspective, this provides the UK with a significant advantage over other jurisdictions. As is well known, pre-processing HTML and, even more so, PDF files is highly labour-intensive, which can easily result in an error rate that propagates through the subsequent steps.

On the other hand, in LegalDocML files, the text is already structured and easily accessible. Therefore, the "pre-processing" of our methodology simply involves collecting all `<paragraph>`s from the input case law directly from the LegalDocML files.

The workflow of the first phase is depicted in Figure 1.

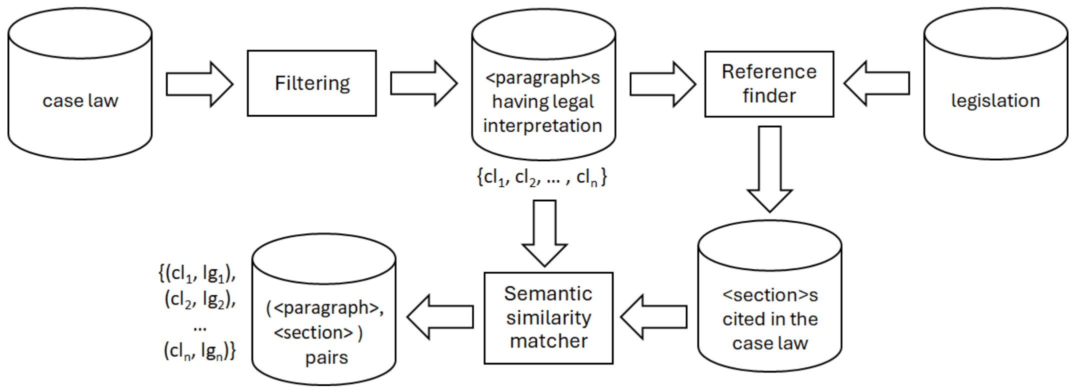

Fig. 1: Workflow for Phase 1 of the proposed methodology

The first step of the methodology is called "Filtering". In this step, the `<paragraph>`s extracted from the case law are filtered to retain only those that convey legal interpretations. This step was introduced because most `paragraph`s in a case consist of conventional phrases included for formal purposes, along with reported background facts and other non-substantive content. Such paragraphs will not be associated with any section of the UK acts cited in the case and are therefore excluded a priori.

In order to identify the `<paragraph>`s of interest, we used `GPT-4o mini`. The model was prompted to determine whether the text contains a legal interpretation. Based on our analysis of the input data and the outcomes of our initial experiments, we decided to instruct the model through the following system role:

(6) *You are a legal language model designed to analyze UK case law for paragraphs that contain legal interpretations. Your task is to identify text that interprets or explains legislative terms and concepts.*
   - *Accurately identify and analyze any legal interpretations within given texts, focusing on how courts, tribunals, or authoritative bodies explain or clarify the meaning or scope of UK legislation.*
   - *Distinguish between mere citations/references and actual legal interpretations. Text that simply cites a statute (e.g., "pursuant to s.100(2)(b)") without any explanatory reasoning or discussion of its meaning does not qualify as interpretation.*
   - *Focus on:*
     - *UK legislation (i.e., Acts of Parliament or other UK statutory instruments)*
     - *Judicial and statutory interpretation principles (e.g., purposive approach, mischief rule)*
   - *Do not consider text as legal interpretation when it:*
     - *Merely mentions the law or quotes statutory wording without explaining it.*
     - *Refers to non-UK conventions, treaties, or rulings.*
     - *Discusses jurisdictional or procedural issues without interpreting legislative language.*
     - *Recites the law verbatim (e.g., "Art. 8 provides...") without additional interpretive commentary.*

Along with the prompt in (6), which is structured following a *chain-of-thought* strategy, we have included a few examples to leverage the few-shot learning technique and enhance model performance [2]. This combined approach—integrating chain-of-thought reasoning with few-shot learning—enables *in-context learning*, where examples within the prompt guide the model towards better outputs. By providing demonstrations and step-by-step reasoning, we effectively steer the statistical distribution of the generated output in a direction that better aligns with our expectations, ultimately improving the model's ability to produce accurate and consistent responses for complex tasks.

To validate the filtering classifier, we adopted a comprehensive approach. First, we manually annotated 12 case laws to create a ground truth dataset. Then, we employed a leave-one-out cross-validation strategy: for each iteration, the model was trained using examples from 11 case laws, while the remaining case law was used as the test set. This process was repeated for all 12 case laws, ensuring a thorough and robust evaluation across the entire dataset.

The validation results are presented in Table 2, on the left. The recall is relatively high, whereas the precision is comparatively lower. We consider these results satisfactory because this module is primarily designed to reduce the number of <paragraph>s processed in subsequent phases. Additionally, as explained in the introduction, our methodology does not aim to *exhaustively* identify all legal interpretations occurring in the case law, but only as many as possible. These interpretations are then intended to be validated by TNA human annotators, and if confirmed, the phrase that is legally interpreted will eventually be tagged as a <term> in the LegalDocML file of the act and linked with the corresponding <paragraph>.

In light of this, it is more critical to ensure that the majority of <paragraph>s conveying a legal interpretation are included in the output, even if this means that the results also contain many <paragraph>s that do not convey a legal interpretation. While a higher risk of false positives (due to high recall and low accuracy) is not a strong concern for the goals of our work, we nevertheless expect that these results might improve by simply increasing the number of examples provided to the model.

On the right, Table 2 presents the number of filtered <paragraph>s relative to the total number of <paragraph>s extracted from the 197 case laws. As shown in the table, the Filtering classifier proves to be highly effective, reducing the set of considered <paragraph>s by 88.27%.

Metric	**Value**
Precision	0.538461538461
Recall	0.913043478260
F1 Score	0.677419354838

Metric	**Value**
Total <paragraph>s	20,724
<paragraph>s with possible legal interpretations	2,430
% of <paragraph>s with possible legal interpretations	11.73%

Table 2: Validation results of the Filtering classifier

Once the set $\{\texttt{cl}_1, \texttt{cl}_2, \dots, \texttt{cl}_n\}$ of <paragraph>s that potentially contain legal interpretations has been identified, all corresponding <section>s from the LegalDocML files of the relevant UK acts are retrieved ("Reference finder" in Figure 1). This step is relatively straightforward as it involves following the <ref> links provided in the LegalDocML files of the case law. Once again, the use of LegalDocML proves highly advantageous, as the files already include the references that, in the case of plain text processing, would need to be identified automatically, thus introducing a potential error rate that could propagate through to the final result. The result of this step is a set of <section>s associated with each <paragraph> that possibly conveys a legal interpretation. For those <paragraph>s that include <ref> links to specific <section>s of the acts, only those referenced <section>s are considered. However, only 355 <paragraph>s, i.e., only 14.6% of the 2,430 filtered ones, have an explicit reference to the legislation. <paragraph>s that do not contain any <ref> link are associated with *all* <section>s of any act referenced anywhere in the case law.

The final step of Phase 1 is called the "Semantic Similarity Matcher". The outcome of this step is a list of pairs $\{(\texttt{cl}_1, \texttt{lg}_1), (\texttt{cl}_2, \texttt{lg}_2), \dots, (\texttt{cl}_n, \texttt{lg}_n)\}$, where $\{\texttt{cl}_1, \texttt{cl}_2, \dots, \texttt{cl}_n\}$ are the <paragraph>s returned by the Filtering classifier, and $\{\texttt{lg}_1, \texttt{lg}_2, \dots, \texttt{lg}_n\}$ are the corresponding <section>s that best match the <paragraph>s, among all those associated with each <paragraph> by the Reference Finder. For this step, we employed OpenAI's embedding models[12], specifically text-embedding-ada-002. For simplicity, we selected a *single* <section> for each <paragraph>, i.e., the <section> with the highest semantic similarity score, even if multiple acts are referenced in the case.

The resulting set of pairs $\{(\texttt{cl}_1, \texttt{lg}_1), (\texttt{cl}_2, \texttt{lg}_2), \dots, (\texttt{cl}_n, \texttt{lg}_n)\}$ serves as the input for Phase 2, described in the next subsection. The objective of the subsequent Phase 2 is to identify the specific legislative term (typically a short noun phrase) that occurs verbatim within the <section> $\texttt{lg}_i$ and is legally interpreted in the <paragraph> $\texttt{cl}_i$. These key legislative terms are potential candidates for annotation as new <term> elements in the LegalDocML files, similar to the term "rights of custody" in (2) above.

[12] https://platform.openai.com/docs/guides/embeddings/embedding-models

3.2 Phase 2: Identifying key legislative terms

In Phase 2, the selected pairs (`<paragraph>`, `<section>`), where the `<paragraph>` belongs to a case law and potentially represents a legal interpretation of a key legislative term found in the associated `<section>`, are processed by an Extractor module. This module employs GPT-4o with *chain-of-thought* techniques. The *chain-of-thought* approach enables the model to break down complex legal reasoning into explicit steps, making its analysis more transparent and reliable.

Therefore, the hypothesis underlying Phase 2 is that, by verbalizing its thought process before reaching conclusions, the model can more effectively identify logical connections between legislative text and its interpretations in case law. The key component of Phase 2's chain of thought are the following:

– **Analysis**: the LLM is tasked with identifying a textual chunk, i.e., a context-providing snippet, within the `<paragraph>` and aligning it with a corresponding textual chunk in the `<section>`. The key legislative phrases are then extracted from the selected chunk in the `<section>` by prompting the LLM to identify the minimal core phrase or phrases. This approach ensures that the system not only extracts key phrases but also considers their interpretation within the given context.

– **Extraction Criteria**: phrases are extracted based on semantic equivalence and interpretative value, as not all text within the `<paragraph>` contains the legal interpretation. The system is guided by the following extraction criteria:
  - **Textual and Semantic Overlap**: phrases must directly reference or semantically align with the same legal context.
  - **Interpretative Relationships**: extracted chunks should focus on meaningful legal interpretations rather than irrelevant mentions.
  - **Specificity**: key phrases should explain a legal concept rather than a generic concept.

– **Validation and Explainability**: To enhance the system's awareness of its extractions, it is required to provide a confidence score and its reasoning for each extraction. The confidence level ("High", "Medium", or "Low") assigned to each match reflects the degree of alignment between the `<paragraph>` and the `<section>`, showcasing a reference-free evaluation approach. Similarly to the approach in [23], our system assesses the quality and relevance of legal text pairs without relying on reference annotations. While [23] employs question-answering for summary evaluation, we adapt this concept to evaluate the semantic alignment between chunks and phrases within the `<paragraph>` and the `<section>`.

– **Structured Output**: the system is required to structure the output in the following JSON format:

```
{ "case_law_chunk": "text from <paragraph>",
  "legislation_chunk": "text from <section>",
  "key_phrases": ["phrase1","phrase2","phrase3"],
  "reasoning": "reason of extraction",
  "confidence": "level of confidence" }
```

The prompt that implement the above chain of thought is shown in (7):

(7) *You are a specialized legal analyst with expertise in matching legal interpretations between case law and legislation. Follow this systematic process:*

– ***ANALYSIS Phase:***
  - *Identify specific (not overly general) legal concepts or phrases in the case law.*
  - *Find the corresponding, equally specific portion in the legislation. This should be a somewhat longer, context-providing phrase.*
  - *From that longer legislative phrase, also extract the key noun phrase(s) or core concept(s)—the minimal expression that captures the critical legal idea.*

– ***MATCHING CRITERIA:***
  - *Direct textual overlap or near-verbatim references (no paraphrasing).*
  - *Semantic equivalence in the same legal context (avoid purely generic wording).*
  - *Clear interpretative relationship (case law explains or applies the legislation).*
  - *Substantive connection (not merely tangential mentions).*

- ***VALIDATION RULES:***
  - *Only extract text that actually appears in each source (verbatim).*
  - *For "`legislation_chunk`", use the longer snippet that captures context.*
  - *For "`key_phrases/concepts`", extract the essential, shorter noun phrase(s) from within that legislation snippet.*
  - *Ensure the match has legal interpretive or explanatory value (avoid trivial or broad phrases).*
- ***OUTPUT STRUCTURE:***
  - *Return a **JSON array** of objects. Each object must contain:*
    - *"`case_law_chunk`": exact phrase from the case law (no rewording).*
    - *"`legislation_chunk`": a longer, context-inclusive phrase from the legislation.*
    - *"`key_phrases/concepts`": the shorter core phrase(s)—verbatim—taken from within "`legislation_chunk`" that most directly capture the legal concept (often a noun phrase).*
    - *"`reasoning`": brief explanation of how the term interprets/applies to the legislation.*
    - *"`confidence`": "`High`", "`Medium`", or "`Low`" based on how closely they match in legal meaning.*
- ***RULES:***
  - *Extract only exact phrases from source texts.*
  - *No rephrasing or inference.*
  - *Include only paired matches with clear legal interpretation.*
  - *Return raw JSON without formatting or explanation.*

A distinctive feature of our approach is its emphasis on explicit reasoning by the LLM. By requiring explanations and encouraging the model to "think aloud," chain-of-thought prompting has been shown to enhance performance across various tasks [18].
Tables 3, 4, and 5 show case law chunks that legally interpret the key legislative terms "child's welfare", "rights of custody", and "controlling or coercive behaviour", which appear in the Children Act (1989), the Child Abduction and Custody Act 1985, and the Domestic Abuse Act 2021 of UK legislation, respectively. These tables show how the same legal concept can be expressed in different forms within the legal narrative.

The first column of the table contains the year, the case law ID, and the paragraph ID (pId) within the case law. For example, '2020/877-10' refers to paragraph 10 in [2020] EWHC 877 (Fam)[13].

Year/Id-pId	`case_law_chunk`
2020/877-10	welfare of the child, while a primary consideration, is not the paramount consideration
2020/3496-10	J's best interests
2020/2878-141	welfare analysis itself involves a balance of interference with and promotion of her rights
2024/1156-44	the judge must consider the child's welfare now, throughout the remainder of the child's minority and into and through adulthood
2021/33-82	adoption was the only realistic option for this child
2021/2931-124	welfare questions in circumstances where moving the child by reason of an unacceptable delay in securing registration may conflict with the child's wider welfare needs

Table 3: child's welfare (Children Act 1989)

4 Analysis and Evaluation

The evaluation of our methodology primarily focused on assessing the accuracy and reliability of extracting the final key terms, such as "child's welfare", rights of custody", and "controlling or coercive behaviour". This section presents the analysis of the results along with their evaluation, including some limitations of our work, which set the basis for future improvements.

[13] https://caselaw.nationalarchives.gov.uk/ewhc/fam/2020/877#para_10

Year/Id-pId	case_law_chunk
2020/1599-86	the child herself objects to being returned
2020/3257-113	breach of the Father's rights of custody
2020/1903-59	removal was indeed in breach of the mother's rights of custody
2022/1827-32	father did not, at the time P was removed from the Republic of Ireland, have rights of custody
2024/1282-14	judge considering a return order
2023/2082-100	the exercise of the discretion under the Convention

Table 4: rights of custody (Child Abduction and Custody Act 1985)

Year/Id-pId	case_law_chunk
2022/2755 - 9	controlling, coercive or threatening behaviour, violence or abuse
2023/2983 - 115	financial abuse of a controlling and coercive nature
2023/505 - 44	cutting her off from friends and family

Table 5: controlling or coercive behaviour (Domestic Abuse Act 2021)

4.1 Analysis of the results

Of the 2,430 <paragraph>s processed in Phase 1, our system successfully extracted one or more key phrases from 2,066 <paragraph>s, accounting for 85% of the total. Each extracted phrase was linked to the JSON template shown above in "Structured Output."

Conversely, for 364 <paragraph>s, the system either failed to associate a <section> with the <paragraph> or did not identify any key phrases within the text. This occurred either because the <paragraph> did not contain any legal interpretation (recall that the precision of Phase 1 is 0.54%), or because the interpreted content does not appear in UK legislation. On a deeper analysis, we came to know that 308 paragraphs actually do not have any legal interpretation. One of the other reasons we found that other legal documents, such as the Hague Convention or the European Convention on Human Rights, are frequently referenced in case law and could be subject to interpretation. However, since our study focuses exclusively on UK legislation and we extract only <section>s from UK laws, the system was unable to process these <paragraph>s.

3008 key legislative terms were extracted from the 2066 <paragraph>s, with several <paragraph>s yielding more than one key term, as explained earlier. These terms occur verbatim in the <section> associated with each <paragraph>. Most of the extracted key phrases are short noun phrases. Their average length is 4.14 words. 49 legislative acts were mentioned in the selected <paragraph>s, with the most frequently mentioned act being the Children Act 1989. This is unsurprising, as the majority of case law in Family law indeed pertains to the custody of minors.

Finally, the system assigned confidence levels ("High", "Medium", "Low") to its outputs based on the semantic alignment between case law and legislative text. Of the extracted key terms, 74.1% (2256 terms) were classified as "High" confidence, 25.8% (747 terms) as "Medium" confidence, and only 0.2% (5 terms) as "Low" confidence. These results underscore the system's confidence in identifying meaningful connections.

The next subsection discusses the evaluation of these results. For this purpose, we hire a PhD student in Law to review the system's outputs, verifying each of the 3008 extracted key legislative terms as well as the 364 <paragraph>s for which the system did not return any key legislative terms.

4.2 Evaluation

The quality and relevance of key phrases were assessed under the supervision of a legal expert with a PhD in law, ensuring a high standard of evaluation. Given the complexity and nuanced nature of legal language, the expert's review was indispensable in ensuring both accuracy and consistency with established legal principles. The evaluation was based on two key criteria: the relevance of each key phrase to the specific case law described in the <paragraph>, and its validity as a recognised legal concept within the context of the act to which the <section> associated with the <paragraph> pertains.

The expert review process involved multiple steps. First, the extracted key phrases, along with their associated <paragraph>s, reasoning, and contextual details, were compiled into an organised spreadsheet. The spreadsheet was structured to enable a systematic evaluation and included two dedicated columns (`key_phrases_check` and `reasoning_check`) with drop-down options of "yes" or "no" for streamlined assessment. Specifically, the legal expert was tasked not only with evaluating whether the key phrase extracted from the <section> was indeed legally interpreted in the <paragraph>, but also with assessing whether the reasoning provided by the LLM about *why* the <paragraph> conveyed a legal interpretation of that key phrase was sound. We consider the evaluation of the latter to be even more critical than the former, as it assesses the *explainability* of our results and lays the groundwork for characterizing, in future research, different sub-categories of legal interpretations.

This spreadsheet was then provided to the expert for annotation and validation. The expert was instructed *not* to seek additional information from TNA's portals or any external sources to ensure that their evaluation was based solely on the information available to the LLM for each <paragraph>-<section> pair.

The legal expert spent approximately four weeks conducting an exhaustive review of the spreadsheet. As explained above, the expert not only verified the accuracy of the extracted phrases but also scrutinised the underlying reasoning, ensuring that the logical connections drawn between the key phrases and their legal implications were sound and well-supported by established legal principles.

As shown in Table 6 on the left, the legal expert marked the great majority of spreadsheet cells as "yes", thereby confirming the LLM's ability to identify and explain legal interpretations of key terms from UK legislation within UK case law. Notably, the model achieves a reasoning accuracy of 98.29% when it is highly confident, demonstrating its capability to handle complex legal interpretation tasks effectively.

Table 6 on the right, by contrast, presents the legal expert's analysis of the 364 <paragraph>s that the system discarded, either due to a failure to associate a <section> with the <paragraph> or because no key phrases were identified within the text. According to the expert, 84.6% of these paragraphs do not, in fact, contain any legal interpretation, while 8.52% of them do include one but not of key phrases occurring in UK legislation. In both cases, the system correctly discharged them. Only 6.8% of the <paragraph>s were mistakenly discharged; in other words, for only 6.8% of them the system failed to recognize either that they contained a legal interpretation or the specific key phrase from UK legislation being interpreted.

Confidence	**Key Phrase Accuracy**	**Reasoning Accuracy**
Low	100%	20%
Medium	99%	32%
High	99.60%	98.29%

Reason discharged <paragraph>	**%**
Paragraphs do not have legal interpretation	84.6%
Interpretation is not of UK legislation	8.52%
Have legal interpretation but system failure	6.8%

Table 6: Analysis of Accuracy and Failure Reasons

In addition, when consulted on the legal soundness of our overall research endeavour, the expert noted that, while the extracted key phrases were interpreted accurately within their respective <paragraph>, this approach may not capture the full complexity of legal reasoning. Legal analysis is inherently multifaceted, often requiring the simultaneous interpretation of multiple legislative <section>s. A single <paragraph> in a case law document may encompass legal concepts influenced by several <section>s, as practitioners frequently consider the combined effect of different provisions to construct arguments or derive conclusions.

This interconnected nature of legal interpretation presents a significant challenge for the methodology, which we aim to address in future work. The current version of our methodology relies on mapping a <paragraph> to a single corresponding <section>. By restricting key phrases to a single <section>, the analysis may overlook nuances arising from cross-references and interdependencies within the legal text. The expert emphasized that this limitation is a critical factor that could affect the depth and comprehensiveness of the extracted legal interpretations.

The second key observation highlighted by the legal expert was the challenge posed by very short <paragraph>s. In some cases, these <paragraph>s lacked sufficient context to enable a complete legal interpretation, necessitating a reference to preceding <paragraph>s within the case law document. This reliance on preceding text underscores the contextual nature of legal language, where meaning often emerges from earlier arguments or explanations. In the current version of our methodology, a consistent unit of processing was defined, with <paragraph>s chosen as the standard unit for analysis. While this approach effectively addresses the majority of <paragraph>s, it does present limitations in situations where context is fragmented across multiple <paragraph>s. However, the expert noted that such cases were relatively rare and did not significantly impact the overall efficacy of the process. For the purposes of this study, this trade-off was considered acceptable. However, it highlights an area for refinement in future iterations of the methodology, where the LLM should be enabled to also examine the <paragraph>s that precede the one under analysis.

5 Related Work

The legal technology domain has seen remarkable advancement in recent years, with researchers addressing various challenges in legal text processing. Significant progress has been made in fundamental tasks such as legal document classification [14], [22], argument mining [10], [24], [5], and information extraction [19]. These studies have demonstrated the potential of both traditional NLP approaches and modern language models in handling legal text complexities.

Legal reasoning and argumentation typically require practitioners to connect multiple sources of legal information. [4] provides a comprehensive review of the transformer-based models in the context of legal applications. They specifically explore how these models can analyze both case law and legislation, with an emphasis on organizing and understanding legal texts. Complementing this, another study [9] highlights the challenges of integrating multidimensional heterogeneous knowledge from legal provisions, judicial interpretations, and case law. By employing a knowledge graph-based approach using a joint knowledge enhancement model (JKEM), the study demonstrates how embedding prior knowledge into large language models can significantly improve the extraction and organisation of legal knowledge. While effective for direct references, their approach does not capture implicit semantic relationships or contextual interpretations. Another study [3] proposed an unsupervised method that calculates topic similarity between legal documents using Sentence Transformers, employing the Akoma Ntoso XML format for annotation. This method is particularly useful for comparing EU regulations with national bills, aiming to facilitate the legislative process. While these methods demonstrate the value of connecting legislative intent with judicial interpretation, they focus primarily on document-level classification, overlooking specific phrase-level interpretations.

While prior research has made significant strides in legal document classification, argument mining, and knowledge integration through advanced NLP techniques, much of the existing work remains limited to document-level analysis or relies on explicit, syntactic matches. These approaches fall short in addressing the complex, phrase-level semantic relationships between legislation and case law, particularly when legal concepts are paraphrased or contextually reinterpreted. Moreover, to the best of our knowledge, despite some recent effort to explore the intersection between legal decisions, argumentation, and generative AI [6], [13], [5], none of these works have yet addressed the problem of interpreting legal terms from legislation within case law. Our study seeks to bridge these gaps by introducing an Information Extraction methodology using LLM that enables the identification and linking of legal phrases across legislative and judicial contexts, even when they are expressed in semantically diverse forms. Furthermore, we have shown how leveraging LegalDocML (with the trustworthy information that they already encode in their structure) proved invaluable for bypassing an otherwise cumbersome preprocessing phase. This highlights the broader importance of providing legal data in a structured, computationally accessible format, where vital metadata embedded in XML can be exploited by sub-symbolic data-driven agents like LLMs.

In this sense, from a higher-level perspective, our work addresses a deeper fundamental need in AI: bridging the gap between symbolic and subsymbolic AI paradigms, searching for a hybrid solution [11]. While data-driven/bottom-up methods (e.g., subsymbolic neural models like LLMs) exhibit strong scalability and performance across many tasks but can sometimes be unreliable, top-down/knowledge-driven tools such as ontologies, LegalXML standards, and logical systems offer reliability and explainability but lack scalability. In this regard, the synergy between LLMs (subsymbolic) and standards like LegalDocML (symbolic) serves at least two critical purposes. First, it enhances the trustworthiness of AI systems by combining the scalability of subsymbolic models with the clarity and rigor of symbolic formalisms. Second, it anticipates the growing legal requirements for transparency and accountability, which is expected to affect, under the growing impact of the AI Act, the whole sphere of Automated Decision Making [12]. Therefore, the specific contribution of this work can be seen as part the more general attempt to bridge the gap between symbolic and sub-symbolic AI towards trustworthy AI systems, which are even more compelling in the legal domain.

6 Conclusions

This paper presented a two-phase methodology to extract legislative text excerpts and connect them to their legal interpretations found in case law, using general-purpose Large Language Models (LLMs) alongside structured legal data in XML format (LegalDocML). The methodology was implemented and evaluated on UK legislation and case law, specifically on the LegalDocML files provided by The National Archives (TNA).

In the first phase, the `<paragraph>`s from the case law that convey legal interpretations were identified and filtered, as well as associated with a corresponding `<section>` of the acts cited in the case law. In the second phase, the relationships between specific key phrases occurring (verbatim) in the `<section>` and their legal interpretations in case law were identified.

The motivation for this work stems from the need to increase the number and quality of legally annotated terms in UK legislation. Our approach effectively reduces the time and labor required for the manual annotation of these links, offering a scalable solution for enriching the annotation of the LegalDocML files.

Currently, only a limited number of such terms are annotated in TNA's LegalDocML files using rigid regular expressions (regex), which often fail to account for the linguistic diversity and contextual nuances of legal texts. For instance, concepts like "child's welfare" are expressed in various ways across different case laws, ranging from "welfare of the child" to more nuanced descriptions such as "welfare questions in circumstances where moving the child by reason of an unacceptable delay in securing registration may conflict with the child's wider welfare needs" (see Table 3 above). Similarly, the concept of "controlling or coercive behavior" may be described in broader contexts, such as "cutting her off from friends and family" (see Table 5 above), which regex-based approaches struggle to identify effectively.

The proposed methodology addresses these challenges by identifying these diverse linguistic forms and linking them to their legal interpretations. This is made possible by LLMs' ability to paraphrase text, effectively capturing linguistic variations while preserving their meaning. This marks an important step toward the development of systems that not only detect such terms more accurately than regular expressions but also connect them to specific case law paragraphs where these terms are legally interpreted.

This work has significant implications for LegalTech, as it paves the way to innovations such as applications that assist lawyers in preparing cases or help judges harmonize the legal interpretation of legislation across cases. The ability to identify and link legal terms to their interpretations enhances the efficiency and accuracy of legal reasoning, reducing the time and effort required for legal analysis and potentially contributing to the fairer and more consistent application of the law.

The results of the methodology demonstrated strong performance, underscoring its potential for broader adoption. The success of this approach is attributable not only to the capabilities of LLMs but also to the availability of well-structured LegalDocML data prepared and validated by TNA's expert annotators. The structured format of LegalDocML files allowed us to bypass the need for additional pre-processing modules to extract relevant information, as the key elements of legal documents were already tagged. For jurisdictions where legal texts are available only in formats like PDF or HTML, additional pre-processing modules would be required, thus increasing the complexity and potential error rate of the system.

While the results are promising, there is still room for improvement. In the future, we plan to expand this work to cover more cases and additional areas of law beyond Family Law. Furthermore, the methodology will be refined to identify all `<section>`s relevant to a `<paragraph>`. Additionally, a tool will be developed for the TNA annotators to quickly validate the extracted key legislative terms and tag them as `<term>` in the LegalDocML file, while also linking them to the relevant case law, thereby refining the annotated datasets released by TNA. As the annotated dataset grows, it could enable the training of domain-specific LLMs, further enhancing the precision and scalability of our proposed tool.

Acknowledgments

This research was supported by Innovate UK project 10106412: "Odyssey - Opening the National Archive's Legal Data to AI for Access to Justice (A2J)".

References

1. Ashley, K.D.: Artificial Intelligence and Legal Analytics: New Tools for Law Practice in the Digital Age. Cambridge University Press (2017)
2. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners (2020), https://arxiv.org/abs/2005.14165
3. Corazza, M., Zilli, L., Palmirani, M.: Topic similarity of heterogeneous legal sources supporting the legislative process. In: CEUR Workshop Proceedings (2024), https://ceur-ws.org/Vol-3878/28_main_long.pdf
4. Greco, C.M., Tagarelli, A.: Bringing order into the realm of transformer-based language models for artificial intelligence and law. Artificial Intelligence and Law **32**, 863–1010 (2024). https://doi.org/10.1007/s10506-023-09374-7, https://doi.org/10.1007/s10506-023-09374-7
5. Grundler, G., Galassi, A., Santin, P., Fidelangeli, A., Galli, F., Palmieri, E., Lagioia, F., Sartor, G., Torroni, P., et al.: Amelia-argument mining evaluation on legal documents in italian: A calamita challenge. In: Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), Pisa, Italy (2024)
6. Grundler, G., Santin, P., Galassi, A., Galli, F., Godano, F., Lagioia, F., Palmieri, E., Ruggeri, F., Sartor, G., Torroni, P.: Detecting arguments in CJEU decisions on fiscal state aid. In: Proceedings of the 9th Workshop on Argument Mining. pp. 143–157 (2022)
7. Jayakumar, T., Farooqui, F., Farooqui, L.: Large Language Models are legal but they are not: Making the case for a powerful LegalLLM. In: Preoțiuc-Pietro, D.e.a. (ed.) Proc. of the Natural Legal Language Processing Workshop 2023. Association for Computational Linguistics (2023)
8. Katz, D.M., Bommarito, M.J.: Measuring the complexity of the law: The united states code. Artificial Intelligence and Law **22**(4), 337–374 (2014)
9. Li, J., Qian, L., Liu, P., Liu, T.: Construction of legal knowledge graph based on knowledge-enhanced large language models. Information **15**(11) (2024), https://www.mdpi.com/2078-2489/15/11/666
10. Liga, D., Palmirani, M.: Transfer learning with sentence embeddings for argumentative evidence classification. In: 20th Workshop on Computational Models of Natural Argument. CEUR-WS.org (2020)
11. Liga, D.: Hybrid artificial intelligence to extract patterns and rules from argumentative and legal texts (2022)
12. Liga, D.: The interplay between lawfulness and explainability in the automated decision-making of eu administration. In: Governance of Automated Decision-Making and EU Law. Oxford University Press (09 2024). https://doi.org/10.1093/9780198919575.003.0009
13. Liga, D., Fidelangeli, A., Markovich, R.: Using ontological knowledge and large language model vector similarities to extract relevant concepts in vat-related legal judgments. In: JSAI International Symposium on Artificial Intelligence. pp. 115–131. Springer (2023)
14. Liga, D., Robaldo, L.: Fine-tuning GPT-3 for legal rule classification. Computer Law & Security Review **51**, 105864 (2023)
15. Niklaus, J., Matoshi, V., Stürmer, M., Chalkidis, I., Ho, D.: MultiLegalPile: A 689GB multilingual legal corpus. In: Ku, L.W., Martins, A., Srikumar, V. (eds.) Proc. of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics (2024)
16. Palmirani, M.: Legislative Change Management with Akoma-Ntoso. Springer Netherlands, Dordrecht (2011)
17. Palmirani, M., Vitali, F.: Akoma Ntoso for Legal Documents, pp. 75–100. Springer Netherlands, Dordrecht (2011)
18. Qiao, S., Ou, Y., Zhang, N., Chen, X., Yao, Y., Deng, S., Tan, C., Huang, F., Chen, H.: Reasoning with language model prompting: A survey (2023), https://arxiv.org/abs/2212.09597
19. Sleimi, A., Sannier, N., Sabetzadeh, M., Briand, L., Ceci, M., Dann, J.: An automated framework for the extraction of semantic legal metadata from legal texts (2020), https://arxiv.org/abs/2001.11245
20. Stern, R., Rasiah, V., Matoshi, V., Bose, S.B., Stürmer, M., Chalkidis, I., et al.: One law, many languages: Benchmarking multilingual legal reasoning for judicial support (2024), https://arxiv.org/abs/2306.09237
21. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E.H., Le, Q., Zhou, D.: Chain of thought prompting elicits reasoning in large language models. CoRR **abs/2201.11903** (2022), https://arxiv.org/abs/2201.11903
22. Xie, Y., Li, Z., Yin, Y., Wei, Z., Xu, G., Luo, Y.: Advancing legal citation text classification a conv1d-based approach for multi-class classification. Journal of Theory and Practice of Engineering Science **4**(02) (2024). https://doi.org/10.53469/jtpes.2024.04(02).03, https://centuryscipub.com/index.php/jtpes/article/view/490
23. Xu, H., Ashley, K.: A Question-Answering Approach to Evaluating Legal Summaries. IOS Press (Dec 2023). https://doi.org/10.3233/faia230977, http://dx.doi.org/10.3233/FAIA230977
24. Zhang, G., Nulty, P., Lillis, D.: Enhancing legal argument mining with domain pre-training and neural networks (2022), https://arxiv.org/abs/2202.13457