Received: 24 February 2026 | Revised: 7 April 2026 | Accepted: 14 May 2026 | Published online: 29 May 2026

RESEARCH ARTICLE

Journal of Computational Law and Legal Technology
2026, Vol. 1, A1, 18 Pages
DOI: 10.47852/bonviewJCLLT62029448

Identifying How UK Legislation Is Applied in Case Law: An Ensemble LLM Approach Using LegalDocML

Safia Kanwal[1], Livio Robaldo[2,*], Stergios Aidinlis[3], Joseph Anim[2] and Davide Liga[4]

[1] *Faculty of Creative and Digital Arts and Sciences, Anglia Ruskin University, UK*

[2] *School of Law, Swansea University, UK*

[3] *School of Law, Durham University, UK*

[4] *Department of Computer Science, Luxembourg University, Luxembourg*

Abstract: Legal judgments derive their value not only from stating what the law is but also from showing how legal principles are applied to the facts of specific disputes. This applicative step, central to doctrines such as stare decisis, remains underexplored in legal artificial intelligence (AI) research, which has largely focused on tasks such as retrieval and classification. Yet effective AI support for legal practice requires transparent methods that trace how statutes are applied in case law. This paper introduces a methodology for bridging UK statutory law with its judicial application by combining large language models (LLMs) with structured LegalDocML data. We process official LegalDocML files published by The National Archives, meticulously curated and validated by legal experts as part of a nationwide modernization of legislative publishing. The UK is among the few countries to provide all legislation and case law in LegalDocML, and to our knowledge, this study is the first substantial academic use of this resource with LLMs for the analysis of how legislation is applied in case law. Our results show that integrating bottom-up neural inference with top-down expert-curated XML data allows the proposed framework to identify phrase-level applications of legislation in case law with high accuracy and explainability. This approach advances practitioner-oriented legal AI and lays the foundation for next-generation LegalTech tools that support precedent analysis and traceable legal reasoning.

Keywords: judicial application of statutes, phrase-level linking of legislation and case law, large language models, LegalDocML

1. Introduction

For legal scholars, the value of legal judgments lies not merely in stating what the law *is* but in explaining how general legal principles are applied to the background facts of specific disputes. This applicative dimension, often referred to as the "A" step in the IRAC (Issue–Rule–Application–Conclusion) framework, a widely used pedagogical tool in legal education, is central to legal analysis. This step lies at the core of legal reasoning, connecting the general and the particular, and underpins doctrines such as *stare decisis*. Scholars emphasize that such reasoning is defeasible and must evolve with technological and societal changes [1]. Legal rules are often vague, analogies and distinctions are crucial, and exceptions may override general principles [2].

In this light, traditional legal theory has long described legal reasoning as a two-step process: first, the determination of the relevant facts and, second, the application of the law to those facts [3, 4]. Despite the centrality of this second step, it remains relatively underexplored in the legal AI literature, which has historically focused more on tasks such as classification, retrieval, and knowledge enrichment.

However, there is growing recognition that effective legal AI must move beyond these tasks and address how judges select, interpret, and apply norms to specific factual matrices [5–7]. Large language models (LLMs) offer promising support in this regard, particularly through their capacity to rephrase and abstract textual content in ways that align underlying meaning across cases [8–10], opening the door to applications that assist legal practitioners in this crucial task. Yet realizing this potential requires careful attention to how legal concepts are expressed in legislation and subsequently applied in judicial reasoning.

Many legal concepts applied in case law are not denoted by a single term or a few specific keywords, but rather by fuller, structured phrases found in legislation; it is these entire phrases that are often applied in judicial reasoning.

We therefore define *phrase-level linking* as the task of identifying, for a given paragraph of a judicial decision, the specific

***Corresponding author:** Livio Robaldo, School of Law, Swansea University, UK. Email: livio.robaldo@swansea.ac.uk

© The Author(s) 2026. Published by BON VIEW PUBLISHING. This is an open access article under the CC BY License (https://creativecommons.org/licenses/by/4.0/).

multi-word excerpt(s) of statutory text whose meaning is being applied to the facts at hand, together with an explicit justification of how the statutory phrase is being applied. Because such phrases are frequently reformulated, paraphrased, or only implied in case law, identifying their application requires going beyond surface-level lexical matching to capture semantic equivalence at the phrase level.

Moreover, legal AI must be *transparent and explainable*: it must not only identify which legal concepts are applied but also show how they are applied and produce a report clarifying the rationale [11]. This is increasingly required by data protection laws and proposed AI-specific regulations. Yet LLMs often hallucinate, invent citations, and struggle with the Application step of IRAC, producing shallow justifications [5, 12]. A system that bridges this gap must therefore combine accuracy with traceability.

This paper addresses the following research objectives. First, we aim to develop and evaluate a methodology for automatically identifying phrase-level applications of UK statutory law within case law, combining LLMs with structured LegalDocML data. Second, we seek to assess the effectiveness of ensemble voting and adjudication across multiple LLMs for both classification and extraction tasks in the legal domain. Third, we investigate the role of LegalDocML's machine-readable structure in anchoring LLM outputs to formal legal structures, thereby supporting explainability.

To contribute to this direction, our study introduces a new information extraction and linking methodology explicitly targeting legal applications. We combine two established methodologies, known as ensemble voting and adjudication and chain-of-thought prompting, applied to state-of-the-art LLMs, with LegalDocML, the XML-based legal standard used by The National Archives for publishing UK legislation and case law, to anchor rules and facts within a machine-readable structure and guide LLMs during inference. This hybrid approach enables LLMs to dynamically incorporate legal knowledge at generation time and identify semantically equivalent or related legal phrases across legislation and case law. By integrating structured symbolic representations with statistical models, our methodology bridges bottom-up neural inference with top-down legal reasoning, aligning outputs with formal legal structures while preserving scalability.

We focus on UK law as an ideal testbed; the methodology, however, is designed to generalize to any common law jurisdiction that adopts structured legal document standards. Using The National Archives API, we downloaded 25,406 decisions encoded in the LegalDocML format from 2000 to 2025. We then applied a stratified sampling strategy to select 426 cases across 11 court jurisdictions, ensuring representative coverage for classification and extraction tasks. This dataset provides the empirical basis for training and evaluating our pipeline.

The remainder of the paper is organized as follows. Section 2 reviews recent natural language processing (NLP) approaches in the legal domain, with particular emphasis on LLMs, to situate our contribution within the current literature. This section focuses on works not cited above, which address the issues explored here from a broader perspective. To the best of our knowledge, this is one of the first studies to systematically use AI to identify and explain how legislative phrases are applied in UK case law at the phrase level.

Section 3 introduces the dataset, describing the stratified sampling strategy applied to 25,406 case law documents from The National Archives. Section 4 presents the two-phase system architecture and its experimental evaluation. Section 5 concludes with a discussion of contributions and directions for future research.

2. Related Work

In theory, applying general rules to case-specific facts follows the structure of syllogistic reasoning: the major premise encapsulates statutes or precedents, the minor premise describes the facts, and a conclusion follows from matching the minor premise to the major premise. In practice, however, aligning real-world factual contexts with legal rules is far more complex. It often requires identifying higher-order legal concepts and drawing analogies to prior cases in which those concepts were applied, even when the textual descriptions of the facts differ significantly. Addressing this complexity computationally has become an active area of research.

The legal domain has recently seen a significant surge in the use of artificial intelligence (AI), particularly NLP, for the automated analysis and understanding of legal texts.

Over the past few years, substantial progress has been made in key foundational areas such as legal document classification [13, 14], argument mining [15–17], and information extraction from statutory and case law documents [18]. These efforts highlight the versatility of both traditional NLP pipelines and, more recently, pre-trained LLMs, which show strong potential for handling the complex linguistic and structural characteristics of legal texts thanks to their ability to rephrase, abstract, and generalize textual content.

Despite the improved accuracies achieved in these contributions, they remain largely task-specific and fall short of enabling a deeper, holistic understanding of legal terminology. Law inherently requires engagement with historical context and the social values that shape legal systems. Legal terms often carry nuanced meanings that evolve through interpretation, making their understanding highly context-dependent. In other words, applying the law cannot be reduced to processing legal documents in isolation but requires attention to the diverse sources that inform legal reasoning, extending beyond statutory texts.

To facilitate the application of the law to the background facts of disputes, legislation is often structured to explicitly define key legal concepts. For instance, UK Acts commonly include clear definitions of important legal terms relevant to the Act [19]. This pattern is also observed in other jurisdictions; many European regulations, for example, include a dedicated article that compiles and defines the most significant legal terms. An example is Article 4 of the European General Data Protection Regulation.

Beyond legislation itself, case law plays a defining role in common law jurisdictions such as the UK, where judicial decisions establish binding precedents that interpret and extend legislative provisions.

However, the interpretive landscape extends even further. A crucial, though more diffuse, layer of legal understanding arises from what is broadly termed "legal doctrine." This encompasses authoritative commentaries on legislation, legal treatises, scholarly articles, ISO standards, and more. While these doctrinal sources are not binding, they carry significant persuasive weight in court.

In light of the above, several approaches in legal NLP have sought to bridge the multiple layers of legal knowledge found in legislation, case law, and legal doctrine. Transformer-based models have proven particularly effective in addressing this challenge; Greco and Tagarelli [20] and Dehghani et al. [21], among others, provide a comprehensive survey of transformer architectures tailored to legal

applications, emphasizing their capacity to analyze heterogeneous legal sources.

A notable example in this line of research is the work by Corazza et al. [22], which proposes an unsupervised method for assessing topic similarity using Sentence Transformers on documents in or converted to the LegalDocML XML format—the same format used by The National Archives and in the system developed in this paper. However, their approach operates at the document level and does not capture fine-grained, phrase-level paraphrasing or latent semantic shifts that frequently occur in judicial reasoning, where courts often express legal concepts using language that diverges from the original legislative text.

Progress in this direction has also been made by Li et al. [23], who introduced the Joint Knowledge Enhancement Model, a system designed to populate a legal knowledge graph that integrates statutory provisions, judicial interpretations, and precedents. Similarly, Arvin [24] presents the CaseHOLD benchmark, which evaluates AI models by requiring them to summarize the holding of a decision, that is, a concise description of the main legal rules established or applied in the case, along with the factors, that is, the specific considerations judges use to determine case outcomes.

While identifying rules, factors, and other key elements can significantly reduce the effort involved in manually reviewing lengthy judicial opinions [25], many more applications of legal concepts from legislation in case law still need to be identified. This is important both to, on the one hand, assist lawyers in analyzing prior case law and understanding which arguments have been or could be used to support their client's position and, on the other hand, to help judges achieve harmonization by applying legislative provisions consistently across similar contexts involving those legal concepts. Discussions with legal experts at The National Archives reinforced the need to look beyond isolated keywords and to capture the longer phrases through which legislative provisions are typically applied in case law.

Moreover, all the systems reviewed above tend to overlook the crucial aspect of explainability. This becomes increasingly important when direct, surface-level links, such as exact lexical matches, are absent. Without clear traceability to the original texts or an explanation of the underlying rationale, users lack a meaningful understanding of how or why these connections were made across heterogeneous legal sources.

Recent investigations have started to incorporate explainability, particularly at the intersection of generative AI and argumentation. For example, Fink and Finck [26] and Liga [27] highlight the importance of designing AI systems that support the "right to explanation," a principle recognized as fundamental for ensuring transparency in automated decision-making under regulatory frameworks such as the European AI Act. Within this emerging area, two main directions can be identified: using LLMs to extract and analyze arguments [28] and integrating argumentation frameworks to enhance the explainability of LLMs.

More broadly, the research community has addressed the limitations of purely data-driven approaches by developing neuro-symbolic approaches that combine LLMs with structured, symbolic formalisms. In the legal domain, these pipelines leverage the expressive power of transformers while introducing reasoning mechanisms to overcome the limitations of purely statistical models [7]. By grounding generation in explicit legal rules and extracted facts, they promote interpretable, traceable decisions and help reduce hallucinations [11]. However, existing benchmarks in legal NLP rarely assess a model's ability to perform analogical reasoning based on case law [29], revealing a significant unmet need. Building dedicated resources that trace how legal concepts are applied in specific cases would provide a crucial foundation for both training and evaluating such systems.

The system developed in this study incorporates both a neuro-symbolic design and explainability. It combines LLM inference with the structured, expert-validated XML of LegalDocML to anchor outputs in formal legal structures, while also generating an explanation of the rationale underlying each connection between legislation and case law.

However, the core contribution of this work lies in leveraging LegalDocML. Its rich, machine-readable structure reduces the need for extensive preprocessing and enhances the interpretability and reliability of the extracted information. Building on this foundation, our system adopts a hybrid approach combining XML-based symbolic reasoning with sub-symbolic methods, allowing LegalDocML to effectively align LLM outputs with formal legal structures.

LegalDocML was also used in the work by Corazza et al. [22] mentioned above. However, most of the LegalDocML files employed in Corazza et al. [22] are not official; instead, they were converted from other formats sourced from various repositories (e.g., the EUR-Lex website maintained by the European Publication Office) using ad-hoc converters developed by the authors.

In contrast, the present work processes the official LegalDocML files published by The National Archives, the UK Publication Office. These files have been edited in great detail and validated by legal experts at The National Archives for several years to date, as part of a broader initiative to modernize the UK's legislative publishing infrastructure and support advanced computational legal analysis. The UK is among the few countries worldwide whose Publication Office publishes all legislation and case law in LegalDocML, and to the best of our knowledge, this is the first study to combine these official LegalDocML files [30, 31] with LLMs for the analysis of how UK legislation is applied in case law [32].

To summarize, while the studies reviewed above have made valuable contributions, a significant gap remains in linking legislation and case law at the phrase level with transparent explanations. Our work addresses this gap by grounding the analysis in official LegalDocML files maintained by The National Archives, enabling systematic, phrase-level identification of how legal concepts are applied across cases. The next section introduces the dataset we constructed to support this methodology.

3. Dataset

The National Archives publishes all UK legislation and case law online through two official portals: *legislation.gov.uk* and *caselaw.nationalarchives.gov.uk*. The former, launched in 2010, hosts UK legislation, while the latter, launched in April 2022, provides access to judgments from several UK courts. Case law is updated daily, reflecting ongoing judicial output, whereas legislation is updated less frequently.

All documents on these portals can be downloaded in LegalDocML, a rich and structured XML standard for legal documents[1] [31, 32]. The LegalDocML annotations are prepared by legal experts at The National Archives, supported by assistive technologies that convert original documents (typically PDF or Word) into structured XML drafts and help with manual validation.

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

Table 1
Sample composition across court jurisdictions

Court jurisdiction	N_i	n_i
EWHC (England and Wales High Court)	14,359	234
EWCA (England and Wales Court of Appeal)	8396	137
UKFTT (UK First-tier Tribunal)	848	14
EWFC (England and Wales Family Court)	459	7
UKUT (UK Upper Tribunal)	424	7
EWCOP (England and Wales Court of Protection)	374	6
UKSC (UK Supreme Court)	176	5
EAT (Employment Appeal Tribunal)	269	5
UKPC (UK Privy Council)	71	5
UKAIT (UK Asylum and Immigration Tribunal)	29	5
EWCC (England and Wales Crown Court)	1	1
Total	25,406	426

Given the complexity of LegalDocML, a detailed explanation of its structure is beyond the scope of this paper. For understanding this work, it suffices to know that legislative acts are organized into indexed <section> elements, as illustrated[2] in (1), while case law documents are structured into indexed <paragraph> elements, as illustrated[3] in (2). LegalDocML also supports references to other acts or case law via the tag, as shown again in (2).

The National Archives also provides an API to programmatically download case law from 11 different court jurisdictions (listed in Table 1). We used this API to retrieve all case law published over the past 25 years, up to May 2025. To ensure traceability in our experiments, we retained only those case law in which every paragraph was properly indexed with the eId attribute, resulting in a final set of 25,406 case law. Regarding legislative acts, we programmatically downloaded all those referenced within the case law via the tag.

(1) <body eId="body">

```
<part eId="part1">
  <num><b>Part 1</b></num>
  <heading>Socio-economic inequalities</heading>
  <section eId="section-1">
    <num>1</num>
    <heading>Public sector duty regarding socio-
    economic inequalities</heading>
    <subsection eId="section-1-1">
      <num>(1)</num>
      <content><p>An authority to which this section
      applies must, when making... </p></content>
    </subsection>
  ...
  </section>
</part>
...
</body>
```

(2) <judgmentBody>

```
<paragraph eId="para 1">
  <num>1.</num>
  <content><p>The provisions of the <ref href="http://
    www.legislation.gov.uk/id/ukpga/1992/34">Sexual
    Offences (Amendment) Act 1992</ref> apply to...
    </p></content>
</paragraph>
...
<paragraph eId="para 27">
  <num>27.</num>
  <content><p>...<ref href="https://caselaw.nationalar
    chives.gov.uk/ewca/crim/2018/185">[2018] EWCA
    Crim 185 </ref>.</p></content>
</paragraph>
...
</judgmentBody>
```

The experiments presented in the following sections were not conducted on the entire dataset of 25,406 downloaded case law, but rather on a reduced, statistically representative sample of 426 case law selected from it, along with the legislative acts they reference. These 426 documents were selected using the sampling methodology described in the next subsection.

3.1. Selection of the experimental sample

For this study, we employed a stratified sampling approach to select case law from the UK legal database, following the methods proposed in Cochran [33] and Lohr [34]. The methodology was designed to ensure representative coverage across court jurisdictions while maintaining statistical validity for classification tasks.

3.1.1. Sample size determination

For a large population, the minimum sample size for proportion estimation in classification tasks is calculated using Cochran's formula [33]:

$$n = \frac{Z^2 \cdot p \cdot (1-p)}{e^2}$$

where Z represents the Z-score for the confidence level, p is the expected proportion, and e is the margin of error. While

[2]This XML excerpt is taken from the LegalDocML version of the Equality Act 2010, available at https://www.legislation.gov.uk/ukpga/2010/15

[3]This XML excerpt is taken from the LegalDocML version of "R v Adrian Revill," available at https://caselaw.nationalarchives.gov.uk/ewca/crim/2025/762

a 95% confidence level with a 5% margin of error is standard, some sources[4] recommend slightly higher confidence, for example, 96% ($Z = 2.05$), in fields requiring greater certainty such as the legal domain. The value of p is set to 0.5, as we had no prior knowledge of the proportion of paragraphs containing legal applications of law. According to Israel [35], $p = 0.5$ provides the most conservative sample size estimate.

Applying Cochran's formula with these parameters gives:

$$n = \frac{(2.05)^2 \cdot 0.5 \cdot (1 - 0.5)}{(0.05)^2} = \frac{4.2025 \cdot 0.25}{0.0025} = 420.25 \approx 421$$

Cochran's formula assumes an infinite population. Since our dataset is finite (25,406 case law documents), we apply the finite population correction:

$$n = \frac{n_0}{1 + \frac{n_0 - 1}{N}}$$

where $n_0 = 421$ and $N = 25{,}406$, yielding an adjusted sample size of $n = 414$. To ensure statistical representativeness across the 11 court jurisdictions, we apply the stratified sampling strategy proposed by Lohr [34], described next.

3.1.2. Stratified sampling strategy

With 25,406 case law documents across 11 court jurisdictions, we applied the stratified sampling approach from Lohr [34], using three constraints: (1) proportional allocation: cases sampled in proportion to each jurisdiction's representation; (2) minimum representation: jurisdictions with at least 5 cases were assigned at least 5; and (3) complete coverage: all jurisdictions included. The proportional allocation formula is:

$$n_i = \max\left(5, n \cdot \frac{N_i}{N}\right) \text{if } N_i \geq 5,\ n_i = N_i \text{ if } N_i < 5$$

where N_i is the population size of stratum i (Table 1), $N = 25{,}406$, and $n = 414$.

Table 1 shows the distribution of case law across court jurisdictions in both the original dataset and the final sample. Applying the minimum representation constraint increased the sample size to 426, due to the 5-case minimum for jurisdictions with sufficient representation (EAT, UKPC, UKAIT). This added five cases beyond proportional allocation, ensuring robust representation across all court types for reliable classification analysis.

This sampling strategy ensured the final dataset remained representative of the court jurisdictions while maintaining sufficient statistical power for classification analysis. The resulting sample of 426 cases represents approximately 1.68% of the total population, with each jurisdiction included in proportion to its prevalence in the original dataset, subject to the minimum representation constraint. The text was extracted from each <paragraph> and <section> element by collecting all text nodes from the XML, effectively stripping markup tags while preserving the textual content of inline references. No additional text cleaning, stemming, or normalization was applied, as the official LegalDocML files are already well-structured and free of the formatting artifacts typical of PDF-derived corpora. The code for downloading and parsing cases and legislation is available at our GitHub repository[5], including the annotations verified by the human evaluators.

[4] https://mathblog.com/statistics/definitions/z-score/ci/96-to-z/

[5] https://github.com/SafiaK/BridgingCaseLawAndLegislation

3.2. Examples

Before proceeding, it is helpful to present a few examples of the texts that will be processed through the LLMs, following the methodology described in the next section. We also include the corresponding results to illustrate the task at hand, that is, identifying how excerpts from legislation are applied within paragraphs of case law, as discussed in the Introduction.

The first example we consider is reported in (3).

(3) **Phrase from legislation:** "(a) the defendant has a criminal lifestyle but has not benefited from his general criminal conduct, or (b) the defendant does not have a criminal lifestyle and has not benefited from his particular criminal conduct."

Act/section: Proceeds of Crime Act 2002, Section 20[6]

Case law paragraph where the phrase is applied: "13. We refused permission to adduce this evidence. CIL's pleaded case, and its case as advanced in its evidence and at trial, has consistently been for damages representing the rent it lost as a result of the Appellant's unlawful occupation of the Land. The benefits that the Appellant obtained through his criminal tipping on the Land are irrelevant to that claim or its quantification. Moreover, the argument to be made on the strength of this evidence would appear to involve a claim to share in the Appellant's proceeds of crime, which might well be open to objection in principle."

Case law/paragraph: [2018] EWCA Civ 162, Paragraph 13[7]

In (3), the legal question is whether fresh evidence can be adduced to support the claims of the Respondent on the basis of the provision about benefit from criminal conduct in the Proceeds of Crime Act 2002, Section 20. The Court applies this provision by conceptually distinguishing benefit from criminal conduct under the provision from a private damages action framed in loss of revenue. In other words, it is a negative application of the legislative provision to the facts of this case by rejecting its applicability. The specific benefits obtained by the Appellant through his criminal tipping of the land are legally distinct from the loss of revenue action currently considered by the Court.

Another example is reported in (4):

(4) **Phrase from legislation:** "(a party to arbitral proceedings) may not raise that objection later, before the tribunal or the court, unless he shows that, at the time he took part or continued to take part in the proceedings, he did not know and could not with reasonable diligence have discovered the grounds for the objection."

Act/section: Arbitration Act 1996, Section 73[8]

Case law paragraph where the phrase is applied: "81. To the extent that in his oral Reply Mr Saunders submitted that there was any unfairness from the issue of polyamorphism being raised late, the Tribunal's response was: "I think the fact they tried to put in evidence later on and did not succeed and so on, that is all irrelevant. They cannot be worse off as a result of having done that than if they had not."

Case law/paragraph: [2023] EWHC 910 (Comm), Paragraph 81[9]

[6] https://www.legislation.gov.uk/ukpga/2002/29/section/20

[7] https://caselaw.nationalarchives.gov.uk/ewca/civ/2018/162#para_13

[8] https://www.legislation.gov.uk/ukpga/2003/41/section/31

[9] https://caselaw.nationalarchives.gov.uk/ewhc/comm/2023/910##para_81

In (4), the legal question is whether it is consistent with fairness standards in the context of procedural issues related to the introduction of evidence to raise an objection after a certain time under the Arbitration Act 1996, Section 73. This is an *implicit* application of the legal provision, as the Court does not directly refer to the wording of the section, but explains that the rise of a specific objection (regarding polyamorphism) that failed does not prejudice the interests of the objecting party, thus not creating complications related to unfairness for them. It is also an *indirect* application of the law, in the sense that the Court refers to an application of the law made by the Tribunal (i.e., the lower-tier Court) and not the High Court itself.

A final example is reported in (5):

(5) **Phrase from legislation:** "the judge must in particular have regard—(a) to the nature and source of the document"

Act/section: Extradition Act 2003, Section 84[10]

Case law paragraph where the phrase is applied: "17. However, Section 84(2) of the Act does not require that something that on its face is a statement within the ordinary meaning of that term should take the exact form of statements made for the express purpose of prosecution in the United Kingdom. Under Section 84(3a) the court must consider, in determining whether a statement should be treated as admissible evidence, the nature of the document. In this case, all the relevant statements contained averments of fact and fell within the criteria of Section 84(2a). Furthermore, they were signed in the presence of the prosecutor as an indication of the witness's belief in the accuracy of their contents."

Case law/paragraph: [2012] EWHC 1098, Paragraph 17[11]

In (5), the legal question is whether a statement made by a person in a document should be considered as admissible evidence by the Court in accordance with the Extradition Act 2003, Section 84. The Court applies this provision by conceptually clarifying what is meant by the "nature" and "source" of the document. In the specific facts of this case, this can involve assessing whether the document was signed "in the presence of the prosecutor as an indication of the witness's belief in the accuracy of their contents." In other words, it is a positive application of the legislative provision to the facts of this case by confirming that the scope of the legislative provision includes considerations such as those mentioned by the Court in the last sentence of the case law paragraph.

4. Identify Phrases from Legislation Linked to Case Law Paragraphs via LLMs

This section presents the core contribution of this research: the system we developed within the Innovate UK project Odyssey to link excerpts from UK legislation to paragraphs in UK case law.

The system takes as input the <paragraph>(s) from UK case law and the <section>(s) from the UK Acts referenced in the case law, as tagged in the LegalDocML format by The National Archives. As discussed earlier, the XML structure of these files offers a straightforward means of extracting the relevant elements, which is significantly easier to implement than processing documents in PDF or Microsoft Word format. It is well known that preprocessing such formats, particularly PDF, is highly labor-intensive and susceptible to introducing errors. These errors can easily propagate through subsequent stages of analysis. In contrast, LegalDocML files provide a well-structured and semantically rich format in which key legal elements are explicitly marked and readily accessible. This significantly reduces the complexity and practically eliminates the potential error rate of preprocessing.

Once the <paragraph>(s) and <section>(s) are extracted, the system operates in two sequential phases, illustrated in Figure 1. In this work, each phase was trained and evaluated exclusively on the 426 case law documents selected in Subsection 3.1. The next two subsections delve into the details of each phase.

The system described in Figure 1 employs two LLM methodologies: ensemble voting and adjudication, used in both phases, and chain-of-thought prompting, used only in the second phase.

Ensemble voting and adjudication is a well-established two-step machine learning methodology that improves accuracy and reduces the biases of individual models, thereby enhancing overall robustness [36, 37]. This approach combines multiple LLMs through voting and adjudication. In the first step, two or more LLMs independently process the same input using identical prompts, each generating its own output based on its inference capabilities. These outputs are then aggregated via a voting mechanism, with the most frequent result becoming the ensemble's preliminary decision. In the second step, if there is disagreement among the models, a third, typically more advanced LLM acts as an adjudicator by reviewing the conflicting outputs and selecting the most appropriate final result.

Chain-of-thought prompting is a strategy that enhances the reasoning capabilities of LLMs by guiding them through intermediate steps to produce a final output. This approach enables *in-context learning*, where examples embedded in the prompt shape the model's behavior toward more accurate and consistent responses [38]. By encouraging models to "think aloud," chain-of-thought prompting enhances performance on complex tasks that require multi-step reasoning. The structured reasoning process not only improves output quality but also facilitates transparency and explainability, as the decision-making pathway becomes explicit and the logical connections across textual inputs are more easily traced.

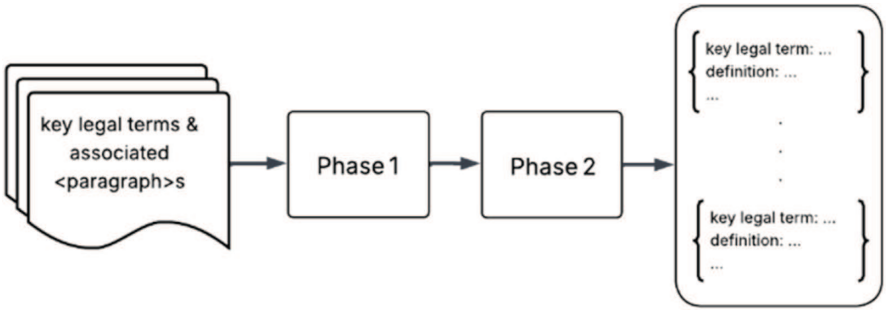
Drawing on extensive literature on ensemble disagreement resolution [39–41], we adopted an informed judge approach to resolve instances where the voting models disagreed. Prior research has consistently shown that ensemble methods incorporating diverse model perspectives, followed by an LLM-as-judge, outperform single-model judges and exhibit reduced intra-model bias. Moreover, these methods achieve better alignment with human judgment when using pairwise comparisons rather than direct scoring, and presenting multiple perspectives to a judge model has been shown to reduce overconfidence in ambiguous cases [42]. In the legal domain, disagreement samples often contain valuable interpretive insights that should be preserved rather than discarded, since legal texts are inherently open to multiple valid interpretations, and a nuanced understanding of context and judicial reasoning is essential [43, 44].

We selected general-purpose state-of-the-art LLMs rather than domain-specific models such as SaulLM or LegalBERT for two reasons. First, our task requires generative capabilities—producing reasoning explanations and extracting arbitrary text spans—which encoder-only models like LegalBERT cannot perform. Second, using models from different providers (OpenAI, Meta, DeepSeek) ensures architectural diversity in the ensemble, reducing shared biases in the voting mechanism.

[10] https://www.legislation.gov.uk/ukpga/2003/41/section/84

[11] https://caselaw.nationalarchives.gov.uk/ewhc/admin/2012/1098#para_17

Figure 1
The two sequential phases of the developed system



4.1. Phase 1: classification of <paragraph>(s)

Many paragraphs in case law documents primarily consist of formal phrases, background information, or other non-substantive content that do not convey any direct legal application of legislative text. To help the system focus on paragraphs that contain actual applications of UK legislation, such non-substantive paragraphs are excluded during the first phase.

This phase thus functions as a filter to retain only those <paragraph> elements likely to contain substantive legal applications. This allows the system to concentrate solely on content relevant to the legislative context.

To ensure robust and reliable classification of the <paragraph>s, we implemented a two-step process combining ensemble voting with adjudication, as illustrated in Figure 2. The following two subsubsections describe each step of the process in detail.

4.1.1. Step 1: ensemble classification

In the initial classification step, a voting mechanism was employed using two state-of-the-art LLMs: GPT-4o-mini[12] and Llama 3.3 70B Versatile[13]. These models were applied to the selected sample of 426 case law documents. Each model independently evaluated <paragraph>s from the cases to determine whether they contained an application of law to specific facts, based on a standardized prompt shown below. To ensure consistency, both models were given identical inputs and followed the same prompting strategy throughout the classification task. Furthermore, all LLM inferences in this phase were conducted with a temperature of 0 to ensure deterministic outputs. GPT-4o-mini was accessed via the OpenAI API, and Llama 3.3 70B Versatile was accessed via the Groq API.

The following few-shot prompt was used consistently across both models to ensure standardized classification criteria:

You are analyzing paragraphs from UK case law to determine if they contain an application of law to specific facts.

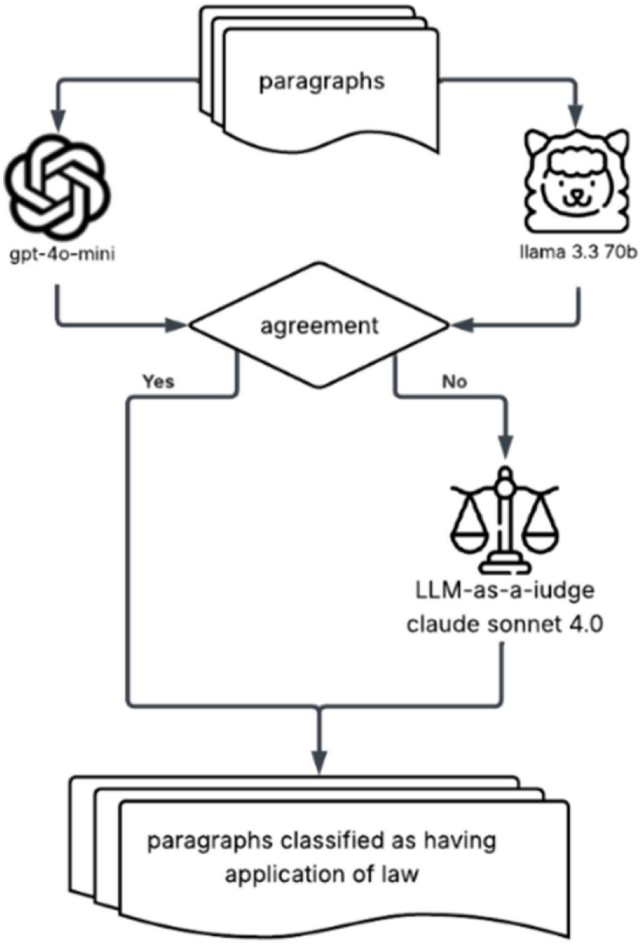
Application of law definition: An application of law is where statutory legal provisions are applied to the specific facts of the case at hand. This goes beyond merely citing or discussing law in the abstract and without specific reference to the facts of the case at hand.

Important: Legal provisions can be applied BOTH explicitly (with direct statutory citation) AND implicitly (without naming specific statutes). Courts often apply legal principles without explicitly citing the statute number.

Indicators of application of law:

1) The judge connects specific statutory legal provisions to the specific factual circumstances.
2) The text shows reasoning that explains how the law resolves or addresses the unique facts.
3) The paragraph contains the judge's analysis leading to a conclusion based on legal principles.
4) Legal tests or criteria are being applied to the case facts.

Figure 2
The ensemble voting-adjudication process in Phase 1



[12] https://developers.openai.com/api/docs/models/gpt-4o-mini

[13] https://console.groq.com/docs/model/llama-3.3-70b-versatile

Not applications of law:

1) Mere citations of statutes, cases, or legal principles without application to facts.
2) Background procedural information or case history.
3) Statements about jurisdiction or general legal explanations.
4) Summaries of arguments made by parties without judicial analysis.
5) Restatements of previous cases without connecting them to current facts.

For each paragraph, determine if it contains an application of law, identify the specific phrases showing application, and provide a brief explanation for your decision. Return a valid JSON.

The LLMs were then provided with a few few-shot examples, one of which was:

"para id": "ewfc 2025 41#para 42",
"para content": "42. In this respect, I recognise that there would indeed...",
"if law applied": 1,
"reason": "The judge applies legal principles regarding parental financial..."

The complete list of few-shot examples used in Phase 1 is available on GitHub.

The para id field serves as a unique identifier that links each classification back to the original LegalDocML paragraph, ensuring full traceability and enabling verification of results against the source documents. The Boolean if law applied field represents the core classification decision: "1" indicates that the paragraph contains a legal application, and "0" indicates that it does not. Most importantly, the reason field provides the model's natural language justification for its classification decision.

Both GPT-4o-mini and Llama 3.3 70B Versatile agree on the <paragraph>s from examples (3) and (5) presented above. In contrast, they disagree on the <paragraph> corresponding to example (4): GPT-4o-mini argues that it contains a legal application, while Llama 3.3 70B Versatile contends that it does not. Each model supports its conclusion with the following reasoning:

(6) GPT-4o-mini: The Tribunal addresses the fairness standards in the context of procedural issues related to the introduction of evidence, applying principles of fairness to the specific circumstances of the case at hand.

Llama 3.3 70B Versatile: This paragraph reports a statement from the Tribunal but does not contain an application of law to specific facts; it is a quote without analysis or reasoning based on legal principles.

4.1.2. Step 2: adjudication

After the ensemble classification, we analyzed the agreement rates between the models. In most cases, GPT-4o-mini and Llama 3.3 70B produced identical classifications (see detailed results in the next subsubsection). These instances were considered reliably classified and retained their consensus labels.

For cases where GPT-4o-mini and Llama 3.3 70B disagreed, Claude Sonnet 4[14] acted as an independent adjudicator (LLM-as-a-judge). Claude Sonnet 4 was selected for its proven capabilities in complex reasoning tasks and its independence from the initial classification models. It was accessed via the Anthropic API with a temperature of 0. This choice helps mitigate potential systematic biases while harnessing the complementary strengths of diverse model architectures, ultimately yielding a final authoritative judgment for each disputed paragraph.

Claude Sonnet 4 received the original paragraph text along with both models' classification and reasoning, that is, the JSON fields "if law applied" and "reason" from the previous phase. In other words, the adjudication prompt was designed to clearly present the conflicting outputs, including each model's Boolean decision, explanatory reasoning, and the textual evidence they relied on. This approach follows best practices in legal NLP, where discrete classifications supported by detailed justifications are generally considered more reliable than continuous scoring systems.

For each pair of classification decisions and reasoning outputs from GPT-4o-mini and Llama 3.3 70B (referred to as Model A and Model B below), the prompt submitted to Claude Sonnet 4 was structured as follows:

You are analyzing paragraphs from UK case law to determine if they contain an application of law to specific facts.

Application of law definition: <same as above>.
Your task is to resolve disagreements between two models by:

1) Carefully analyzing both models' reasoning.
2) Evaluating which model's interpretation better aligns with the definition above.
3) Providing your own classification with detailed justification.

Focus on evidence-based decision-making and provide clear reasoning for your final classification.
Model A Decision: 0/1
Model A Reasoning: [detailed explanation] Model B Decision: 0/1
Model B Reasoning: [detailed explanation]

One example of disagreement submitted to Claude Sonnet 4.0 for resolution is shown in (6) above. The <paragraph>, along with the two conflicting classifications and their accompanying reasoning, was presented to Claude Sonnet 4.0, which sided with the decision made by GPT-4o-mini (referred to as Model A):

Claude Sonnet 4.0: Model A's analysis is more accurate. While this paragraph does contain a quote from the Tribunal, it represents the Tribunal's judicial reasoning applying fairness principles to the specific procedural circumstances of the case. The Tribunal is addressing Mr Saunders' submission about unfairness regarding the late raising of the polyamorphism issue and providing legal analysis that connects fairness standards to the specific facts. The Tribunal's statement demonstrates the application of legal principles (fairness in procedural matters, particularly regarding late evidence introduction) to resolve the specific factual dispute. The fact that it's presented as a quote doesn't negate that it contains substantive judicial reasoning applying law to facts. Model B incorrectly focuses on the format (being a quote) rather than the substantive legal analysis contained within the quote.)

4.1.3. Results and evaluation

This section presents the comprehensive results of our two-phase classification process, including the performance of individual models, patterns of agreement between models, and the effectiveness of our LLM-based adjudication.

Inter-model agreement analysis. The initial classification stage revealed substantial agreement between Llama 3.3 70B Versatile and GPT-4o-mini across the dataset. A total of 23,983

[14] https://platform.claude.com/docs/en/api/overview

paragraphs were extracted from the 426 case law documents selected as described in Subsection 3.1. As shown in Table 2, the two models agreed on 81.77% of paragraphs (19,611 cases), while they disagreed on 18.23% (4372 cases). This high level of consensus demonstrates the robustness of our classification criteria and the consistency of both models in interpreting legal texts according to the established framework.

Table 2
Agreement analysis between Llama 3.3 70B and GPT-4o-mini

Prediction	Count	%
False (no law applied)	14,981	62.5
True (law applied)	9,002	37.5

The agreement rate of approximately 82% indicates that the majority of paragraphs presented clear-cut cases where both models could confidently determine whether legal provisions were applied to specific facts or not. The remaining ~18% of disagreement cases represented more nuanced scenarios requiring expert adjudication, justifying our two-stage methodology.

Model prediction patterns. Analysis of the individual model predictions reveals distinct classification patterns, as summarized in Table 3. Llama 3.3 70B took a more conservative stance, labeling 62.85% of paragraphs (15,074 cases) as not containing legal application and 37.15% (8909 cases) as positive. By contrast, GPT-4o-mini exhibited a slightly more liberal approach, identifying 43.2% of paragraphs (10,352 cases) as containing legal application and 56.8% (13,631 cases) as negative.

Table 3
Distribution of positive and negative predictions

	False (no law applied)		True (law applied)	
Model	Count	%	Count	%
Llama 3.3 70B	15,074	62.85	8909	37.15
GPT-4o-mini	13,631	56.8	10,352	43.2

This difference in classification thresholds indicates that Llama 3.3 70B applied stricter criteria to identify legal application, while GPT-4o-mini showed greater sensitivity to implicit legal reasoning. These approaches highlight the strength of the employed ensemble methodology in capturing diverse perspectives on legal text interpretation.

Adjudication outcomes. After Claude Sonnet 4 resolved the disagreement cases, we obtained the final dataset, composed as shown in Table 4. The adjudication process classified 62.5% of the disagreement paragraphs as not containing legal application and 37.5% as containing legal application.

Table 4
Final predictions after Claude-based disagreement resolution

Agreement Type	Count	%
Agreement	19,611	81.77
Disagreement	4372	18.23

This final distribution aligns with Llama 3.3 70B's initial predictions, as further evidenced by the results in Table 5, which show Claude's strong preference for Llama 3.3 70B's classifications during disagreement resolution.

Claude agreed with Llama in 80.8% of contested cases (3569 instances), compared to 19.2% agreement with GPT-4o-mini (849 instances). This marked bias toward Llama's classifications suggests that Claude's interpretation of the legal application criteria was more in line with Llama's conservative approach, potentially reflecting a shared inclination toward stricter standards for identifying genuine law-to-fact applications. Importantly, the adjudicator only resolves disagreement cases, which represent approximately 18% of the total paragraphs; the remaining 82% are determined by consensus between the two voting models without adjudicator involvement.

Table 5
Claude's agreement with base models in disagreement resolution

Agreement with	Count	%
Llama 3.3 70B	3532	80.8
GPT-4o-mini	840	19.2
Total	4372	100.0

The 9002 <paragraph>(s) identified in Phase 1 were further processed in Phase 2 to determine which specific excerpts from the UK Act, as referenced in the case law containing these <paragraph>(s), are actually applied within them.

4.2. Phase 2: identification of legislative text applied within the <paragraph>(s)

Following Phase 1, which identified 9002 <paragraph>(s) likely to contain applications of law from the selected sample of 426 case law documents, Phase 2 focuses on determining the specific excerpts of UK legislation actually applied within each <paragraph>. Phase 2 produces the final results of the proposed system, some of which have already been illustrated in Subsection 3.2 for demonstration purposes. Figure 3 describes the specific steps of Phase 2.

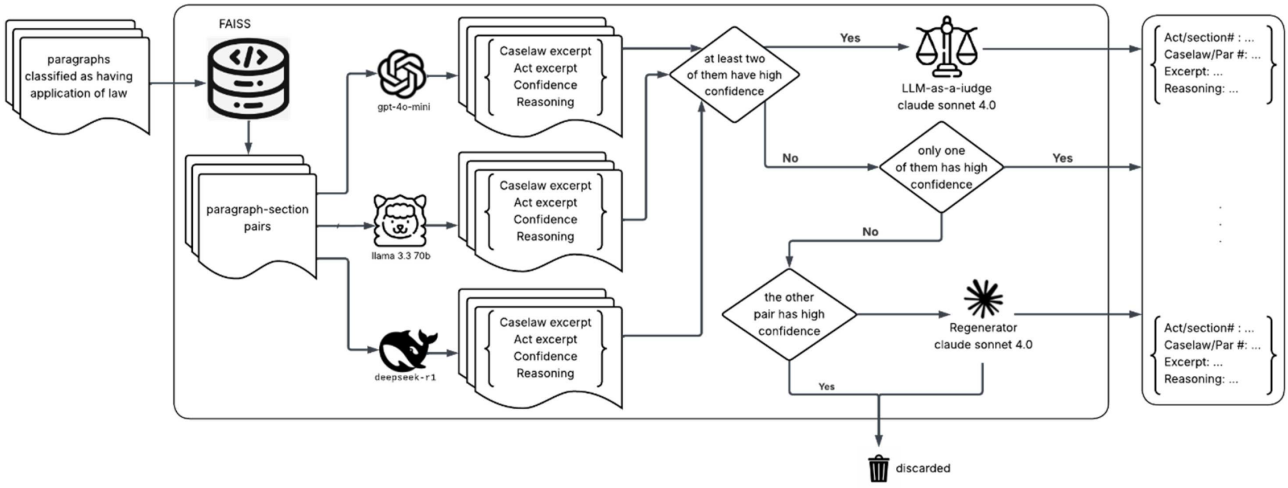
Phase 2 begins by using the FAISS (Facebook AI Similarity Search) library[15], a widely used library designed for efficient similarity search and clustering of dense vectors (embeddings), to identify, based on textual similarity, the <section>(s) from legislative acts referenced in the case law that are most likely applied within the <paragraph>(s).

Each resulting <paragraph>–<section> pair is then submitted to an ensemble of three LLMs, which are responsible for identifying the specific excerpt of text from the <section> that is applied in the <paragraph>, if any. In contrast to Phase 1, which employed two voting LLMs, Phase 2 relies on a three-model ensemble. A further distinction is that each LLM must indicate its level of confidence in the result, marked as either "low" or "high." These two technical enhancements were introduced because the task of extracting a specific text excerpt is clearly more demanding than binary classification.

In the final step, the outputs of the three voting LLMs are analyzed. If a single LLM identifies an excerpt with high confidence, that excerpt is included in the output. If multiple excerpts are returned with high confidence, Claude Sonnet 4.0 is again called upon to act as an LLM-as-a-judge and select the one that

[15] https://github.com/facebookresearch/faiss

Figure 3
The ensemble voting-adjudication process in Phase 1



best matches the <paragraph>. If no excerpt is identified with high confidence, Claude Sonnet 4.0 is instead asked to extract an excerpt on its own. If this new extraction is returned with high confidence, it is added to the output; otherwise, the <paragraph>–<section> pair is deemed not to reflect a genuine legal application and is therefore discarded.

The next three subsections describe each step of Phase 2 in further detail.

4.2.1. Step 1: textual similarity

In the first step, we matched the 9002 case law <paragraph>(s) containing legal applications with the relevant <section>(s) from UK legislation. These <section>(s) will serve as potential "candidates" from which to extract the specific textual excerpts applied in the <paragraph>(s) in the following steps.

To perform this matching, we again employed FAISS[16]. Since FAISS operates on vector representations, we first converted the 9002 <paragraph>(s) and the <section>(s) into embeddings using OpenAI's text-embedding-ada-002 model[17]. This model produces 1536-dimensional vector representations. The FAISS index was configured using the default flat L2 (Euclidean distance) index via the LangChain FAISS wrapper, performing exact nearest-neighbor search with $k = 2$.

However, we did not match each <paragraph> with <section>(s) from all UK Acts. Instead, we again leveraged the LegalDocML format to significantly reduce the number of <section>(s) to be considered for each <paragraph>.

Specifically, for a given <paragraph>, we only considered <section>(s) from those acts that are referenced via the LegalDocML tag within the case law document in which the <paragraph> occurs. Furthermore, if the <paragraph> itself contains explicit references to specific acts or to parts, chapters, etc. of acts, we prioritized the <section>(s) contained within those references.

Each <paragraph> was associated with *two* <section>(s), specifically the two that achieved the highest FAISS semantic similarity scores, based on the prioritization strategy described above.

By increasing the number of <section> candidates per <paragraph>, we expanded the pool of potential excerpt matches and thereby reduced the likelihood that no applicable excerpt would be found. As shown below, only 511 <paragraph>(s) were ultimately discarded, a number we consider negligible. This figure could potentially be reduced even further by associating three or more <section>(s) with each <paragraph>.

This process described above resulted in 18,004 <paragraph>–<section> pairs; that is, 9002 <paragraph>(s) each linked to two <section>(s), which serve as the input for the subsequent step.

4.2.2. Step 2: ensemble extraction of the excerpts

As in Phase 1, each of the 18,004 selected pairs undergoes an ensemble voting and adjudication process, although for a more complex task: identifying the specific excerpts of text from the <section>(s) that apply to the <paragraph>(s).

Since selecting a specific excerpt of text is more challenging than the binary classification performed in Phase 1, we opted to use an ensemble of three LLMs rather than just two. The selected models were GPT-4o-mini [6], LLaMA 3.3 70B Versatile [38], and DeepSeek R1[18]. A three-model ensemble was adopted in Phase 2 to increase robustness for the more demanding extraction task. DeepSeek R1 was selected for its reasoning-oriented architecture, which complements the other two models. All three models were run with a temperature of 0.

For each <paragraph>–<section> pair, each model was independently asked to identify the excerpt within the <section> that best represents the legal applicability to the <paragraph>.

Once again, we employed a chain-of-thought prompting strategy to achieve this goal; that is, we asked the models to verbalize their reasoning process to better guide them in identifying meaningful and contextually relevant connections. The chain-of-thought prompting used in Phase 2 is more advanced than the one in Phase 1, again reflecting the increased complexity of the task. Specifically, Phase 2's approach goes beyond providing a brief explanation of the result (as in Phase 1's binary classification) but rather requires a more detailed reasoning process, articulated through the following components:

[16] https://github.com/facebookresearch/faiss

[17] https://developers.openai.com/api/docs/models/text-embedding-ada-002

[18] https://www.deepseek.com/en/

Identification of a textual excerpt within the <paragraph>: In addition to selecting an excerpt from the <section>, the LLM is tasked with identifying a corresponding textual chunk, that is, a context-providing snippet, within the <paragraph>, aligning it with the legislative excerpt. This ensures that the system not only extracts key legislative passages but also pinpoints precisely where they are applied within the <paragraph>.

Task-specific prompts: The models receive prompts containing specific extraction criteria focused on semantic equivalence and interpretative value, aiming to help them recognize that not all text within the <paragraph> is relevant to the legal application. In particular, the models are instructed to follow these criteria:

1) Specificity: Extract key phrases that explain a legal concept rather than a generic one.
2) Textual and semantic overlap: Select phrases that directly reference or semantically align with the same legal context.
3) Interpretative relationships: Focus on meaningful legal interpretations rather than irrelevant mentions.

Self-assessment: To enhance the system's awareness of its own outputs, the models are required to provide a binary confidence score ("High" or "Low"). In the legal domain, where accuracy and accountability are paramount, this self-rating is especially valuable. It aids error detection by flagging potentially unreliable results and supports more effective human oversight, particularly in cases of model disagreement or ambiguity.

Explicit explanation: As in Phase 1, models must provide a clear, plain-language explanation justifying why, in their view, the <paragraph> represents a legal application of the legislative excerpt from the <section>. In legal contexts, such justifications are crucial for transparency, interpretability, and traceability. They further facilitate human validation and error analysis and serve as a chain-of-thought prompt that encourages structured and reliable reasoning from the models.

To implement the above component, we submitted to each of the three models the following prompt:

You are a specialized legal analyst with expertise in finding which part/provision of the law is applied in the case law. You will be given a section from the legislation act and a paragraph from the case law with their identifiers. Your task is to extract the exact text excerpt from the Legislation text which is applied in the paragraph. Also extract the exact piece of text from the case law where the law is applied. Follow this systematic process:

1) Analysis phase

a. Identify specific legal expressions or actions in the case law.
b. Find the corresponding, equally specific portion in the legislation. This should be a somewhat longer, context-providing phrase.

2) Matching criteria

a. Clear interpretative relationship (case law explains or applies the legislation).
b. Direct textual overlap or near-verbatim references.
c. Semantic equivalence in the same legal context.

3) Validation rules

a. **Only** extract text that actually appears in each source (verbatim).
b. For "legislation excerpt," include a context-complete snippet (not just keywords).
c. For "case law excerpt," extract the exact text that invokes or applies the law.
d. Avoid generic overlaps that don't convey a clear legal interpretation.
e. Return at least one extraction.

4)Output structure

Return a **JSON object** with the following structure:
{
"para id": "paragraph identifier from input",
"section id": "section identifier from input",
"extracted phrases": [{
"case law excerpt": "exact phrase/excerpt from the case law (no rewording)",
"legislation excerpt": "corresponding phrase/excerpt from the section text (no rewording)",
"confidence": "High/Low High: any of the section provision is applied or textual overlap or clear semantic equivalence or interpretation of section; Low: Weaker connection",
"reasoning": "brief explanation of how the legislation section is applied or how legislation excerpt and case law excerpt are connected"}],
}
Rules:
- Extract only exact phrases from source texts.
- There should be nothing produced from your own side—Just extract from the
given sections of case law and legislation.
- No rephrasing or inference.
- Include only paired matches with clear legal interpretation.
- Return raw JSON without formatting or explanation.
- ALWAYS include the para id and section id from the input in your response.
- ALWAYS RETURN SOME RESULT!!!

The LLMs were then provided with a few few-shot examples, one of which was:

"para id": "ukftt grc 2025 289#para 14",
"para content": "14. The Applicant misunderstands the remit the application...",
"section id": "id/ukpga/2018/12 section-166",
"section text": "166. Orders to progress complaints (1) This section applies...",
"extracted phrases": [{
"case law excerpt": "The Applicant misunderstands the remit the application...",
"legislation excerpt": "The Tribunal may, on an application by the data...",
"confidence": "High",
"reasoning": "The judge applies section 166 of the Data Protection Act 2018..."
}]

The complete list of few-shot examples used in Phase 2 is available on GitHub. An example of a result returned by one of the LLMs (GPT-4o-mini) is the output produced by the model for the example shown above in (3). GPT-4omini identifies, with "High" confidence, that paragraph 13 of [2018] EWCA Civ 162 specifically applies the excerpt from the Proceeds of Crime Act 2002, Section 20, reported in the legislation excerpt JSON field. The case law excerpt JSON field instead contains the relevant passage from the case law paragraph that the LLM considered

in making its determination, while the reasoning provides the rationale for that decision.

(7) "para id": "ewca civ 162#para 13",

"para content": "13. We refused permission to adduce ...",
"section id": "id/ukpga/2002/29 section-170",
"section text": "170 No order made: reconsideration ... ",
"extracted phrases": [{

"case law excerpt": "The benefits that the Appellant obtained through his criminal tipping on the Land are irrelevant to that claim or its quantification. Moreover, the argument to be made on the strength of this evidence would appear to involve a claim to share in the Appellant's proceeds of crime, which might well be open to objection in principle.",
"legislation excerpt": "(a) the defendant has a criminal lifestyle but has not benefited from his general criminal conduct, or (b) the defendant does not have a criminal lifestyle and has not benefited from his particular criminal conduct.",
"confidence": "High",
"reasoning": "The case law reflects upon the irrelevance of the Appellant's benefits from illegal activities concerning the claim for damages, which aligns with the legislative context about assessing benefits from criminal conduct under section 170."

}]

For the same example, DeepSeek-R1 did not identify any excerpt from the Proceeds of Crime Act 2002, Section 20, whereas LLaMA 3.3 70B did identify one, but with "Low" confidence.

4.2.3. Step 3: adjudication and re-extraction of the excerpts

As shown in Figure 3, after the ensemble extraction of excerpts, only those with "High" confidence are retained, while all excerpts labeled "Low" confidence are discarded. Furthermore, since the ultimate goal of the process is to identify excerpts of legislative texts explicitly interpreted in the <paragraph>s, any results in which the legislation excerpt does not appear verbatim in the <section> are automatically assigned a "Low" confidence level and therefore discarded.

For the 18,004 <paragraph>–<section> pairs, the distribution of "High" and "Low" confidence is shown in Table 6. Specifically, the table indicates how many of the 18,004 pairs were assigned "High" confidence by all three models, by only two, by only one, or by none, for excerpts extracted from the <section> that apply to the <paragraph>.

Table 6
Confidence levels for <paragraph>–<section> pairs

Confidence category	Count	%
All three models "High"	4936	27.4
Two models "High," one "Low"	5346	29.7
One model "High," two "Low"	4389	24.4
All three models "Low"	3333	18.5

Following the flow chart in Figure 3, the 10,282 pairs (4936 + 5346) that received two or three "High" judgments were sent to Claude Sonnet 4.0, which acted as an independent LLM-as-a-judge, that is, responsible for selecting, among the two or three candidate excerpts, the one that best applied to the <paragraph>.

The 4389 pairs that received a single "High" confidence vote were directly included in the output. This was the case, for example, for the pair discussed above in (7), for which only GPT-4o-mini identified an excerpt of legislative text with "High" confidence.

The 3333 pairs for which all three models extracted excerpts but only with "Low" confidence were considered for regeneration. Indeed, we chose not to regenerate all of them, but only those where the corresponding pair, featuring the same <paragraph> but a different <section>, was also associated by the voting LLMs with three "Low" excerpts. This condition held for 1076 pairs, meaning that for 538 <paragraph>s (1076/2) all three LLMs failed to identify any excerpt of applied text from either of the associated <section>s.

Therefore, for 8464 (9002–538) <paragraph>s, that is, about 94% of the 9002 initial <paragraph>s, at least one of the three LLMs was able to identify an excerpt of legislative text applied to the <paragraph> from at least one of the two associated <section>s. Increasing the number of <section>s associated with each <paragraph> beyond two would likely raise this number further, since it would expand the pool of candidate <section>s from which the LLMs could extract applied text. However, we do not expect the proportion to increase substantially beyond 94% and therefore consider two candidate <section>s per <paragraph> to be sufficiently satisfactory.

As indicated by the flow chart in Figure 3, the 1076 pairs that received "Low" confidence from all three voting models were sent to Claude Sonnet 4.0 for regeneration using the same prompt as for the voting models. Claude Sonnet 4.0 largely confirmed the absence of any legal application in the corresponding <paragraph>s: only 32 pairs yielded excerpts that the model identified as applied to the <paragraph> with "High" confidence, 10 of which involved both <section>s linked to the same (5) <paragraph>s. These 32 pairs were added to the output, while the remaining 1044 (1076–32) were discarded.

In contrast, for the 10,282 pairs sent to Claude Sonnet 4.0 for adjudication (acting as an LLM-as-a-judge), the model was provided with the original <paragraph> and <section> texts, along with the complete outputs from the two or three LLMs. We remind that each output included (i) the candidate case law excerpt, (ii) the candidate legislative excerpt, and (iii) the reasoning supporting that extraction. The adjudicator's task was to select the output(s) most appropriate for the given <paragraph>–<section> pair, ensuring at least one selection per pair. The adjudication prompt was:

You are an expert legal analyst tasked with selecting excerpts of legislation most relevant to a given case law paragraph, based on results produced by AI models.

TASK: Analyze the provided section, paragraph, and list of excerpts. Each excerpt was extracted verbatim from the section and was identified by an AI model as applicable to the paragraph with high confidence. Each model also provided a relevant excerpt from the paragraph along with an explanation to justify its choice. From the list of provided excerpts, you must select the one that best applies to the paragraph. ALWAYS SELECT ONE EXCERPT (NEVER LEAVE IT EMPTY).

1) Evaluation criteria:

a. Relevance to paragraph context
b. Quality and specificity of legal terms
c. Clear connection between case law and legislation

d. Avoid redundant/duplicate content
e. Do not report explanations in the output

2) Response format:
Return only a valid JSON dictionary:
{
"para id": "original para id",
"model": "model name",
"section id": "section id",
"legislation excerpt": "selected legislation excerpt"
}

The excerpt selected by Claude Sonnet 4.0 from the two or three candidates is retained in the output, while the others are discarded. An interesting example is provided by the results for paragraph 66 of [2005] EWHC 402 (Fam)[19]:

(8) "para id": "ewhc fam 2005 402#para 66",
"para content": "66. Miss Theis further submitted that, on the information available now to the court and taking into account the statutory criteria, the target for saved excess income and surplus capital of the wife should be a fund sufficient to secure her through the minority of the children, i.e. after the term of four years has ended in 2009 until 2020 when S will be 21 years old. Etc."

All three models identified an excerpt of applied legislative text from section 29 of the Matrimonial Causes Act 1973[20] with "High" confidence, as follows:

(9) GPT-4o-mini:

"legislation excerpt": "no financial provision order and no order for a transfer of property under section 24(1)(a) above shall be made in favour of a child who has attained the age of eighteen",
"case law excerpt": "the target for saved excess income and surplus capital of the wife should be a fund sufficient to secure her through the minority of the children, i.e. after the term of four years has ended in 2009 until 2020 when S will be 21 years old",
"reasoning": "The paragraph discusses financial provision aimed at providing for the wife's needs during the minority of the children, reflecting the provisions under section 29 that restrict financial orders once the child reaches eighteen years of age, thereby indicating the court's considerations of timing and duration of support."

DeepSeek-R1:

"legislation excerpt": "shall not in the first instance extend beyond the date of the birthday of the child next following his attaining the upper limit of the compulsory school age (construed in accordance with section 8 of the Education Act 1996) unless the court considers that in the circumstances of the case the welfare of the child requires that it should extend to a later date; and (b) shall not in any event, subject to subsection (3) below, extend beyond the date of the child's eighteenth birthday.",
"case law excerpt": "the target for saved excess income and surplus capital of the wife should be a fund sufficient to secure her through the minority of the children, i.e. after the term of four years has ended in 2009 until 2020 when S will be 21 years old. She submitted that this is not a Duxbury case where the wife is entitled to fixed lifetime maintenance.",
"reasoning": "The case discusses the duration of financial provision orders in favour of children, specifically addressing the age limit and the possibility of extending the order beyond the child's 18th birthday if the child is in education or training. This directly relates to section 29 of the legislation, which sets out the duration and age limits for such orders."

LLaMA 3.3 70B:

"legislation excerpt": "shall not in any event, subject to subsection (3) below, extend beyond the date of the child's eighteenth birthday",
"case law excerpt": "the term of four years has ended in 2009 until 2020 when S will be 21 years old",
"reasoning": "The court is considering the duration of financial provision orders in favour of children, and the case law excerpt mentions a specific time frame that aligns with the legislative provision regarding the upper age limit for such orders."

Claude Sonnet 4.0 chose the result produced by GPT-4o-mini as the most representative applied excerpt from the section.

Interestingly, two of the LLMs (DeepSeek-R1 and LLaMA 3.3 70B) also identified, with "High" confidence, an excerpt of applied legislative text from section 25 of the same act[21], which is the second section associated with this paragraph by the FAISS library, whereas GPT-4o-mini did not identify any excerpt for this pair:

(10) DeepSeek-R1:

"legislation excerpt": "It shall be the duty of the court in deciding whether to exercise its powers under section 23, 24, 24A, 24B or 24E above and, if so, in what manner, to have regard to all the circumstances of the case, first consideration being given to the welfare while a minor of any child of the family who has not attained the age of eighteen.",
"case law excerpt": "Miss Theis further submitted that, on the information available now to the court and taking into account the statutory criteria, the target for saved excess income and surplus capital of the wife should be a fund sufficient to secure her through the minority of the children, i.e. after the term of four years has ended in 2009 until 2020 when S will be 21 years old.",
"reasoning": "The case law discusses the application of statutory criteria in determining financial arrangements for the wife, specifically considering the welfare of the children until they reach majority. This directly relates to the legislative requirement for the court to prioritize the welfare of minors when exercising its powers."

LLaMA 3.3 70B:

"legislation excerpt": "first consideration being given to the welfare while a minor of any child of the family who has not attained the age of eighteen",
"case law excerpt": "the target for saved excess income and surplus capital of the wife should be a fund sufficient to secure her through the minority of the children, i.e. after the term

[19] https://caselaw.nationalarchives.gov.uk/ewhc/fam/2005/402#para_66

[20] https://www.legislation.gov.uk/ukpga/1973/18#section-29

[21] https://www.legislation.gov.uk/ukpga/1973/18#section-25

Table 7
Distribution of excerpts selected by Claude Sonnet 4.0

Model	Count	%
DeepSeek-R1	6845	66.6
GPT-4o-mini	1935	18.8
Llama 3.3 70B	1502	14.6
Total	10,282	100.0

of four years has ended in 2009 until 2020 when S will be 21 years old",

"reasoning": "The court considers the welfare of the child as the first consideration in deciding how to exercise its powers, applying section 25(1) of the Matrimonial Causes Act 1973, to determine the financial arrangements for the wife until the children reach majority."

Claude Sonnet 4.0 chose the excerpt selected by DeepSeek-R1 for this pair.

Table 7 shows the distribution of model votes selected by Claude Sonnet 4.0.

The results show that Claude Sonnet 4.0 mostly "trusts" DeepSeek-R1, which accounted for the majority of final extractions (66.6%), followed by GPT-4omini (18.8%) and LLaMA 3.3 70B (14.6%).

These results suggest that, although DeepSeek-R1 was consistently favored during adjudication, both GPT-4o-mini and LLaMA 3.3 70B still contributed a substantial proportion (>30%), often capturing secondary yet valid applications of the law. The combination of ensemble voting and adjudication therefore appears to ensure that nuanced cases are not overlooked.

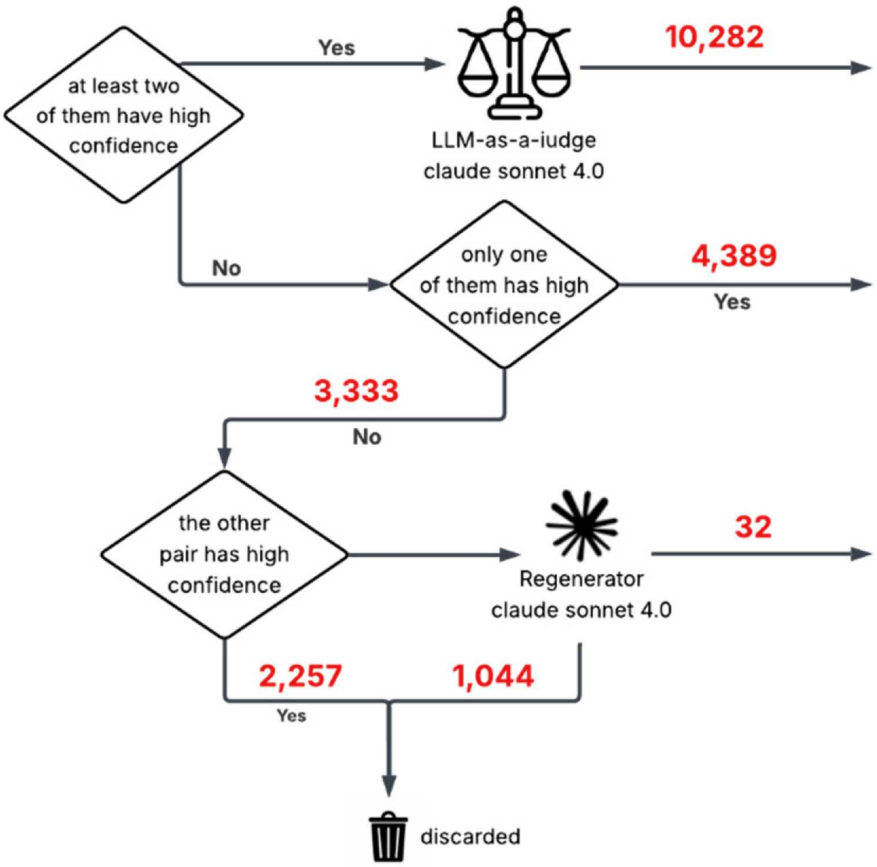
4.2.4. Results and evaluation

Figure 4, which "zooms in" on the rightmost part of Figure 3, summarizes the results of Phase 2, that is, the final outcomes of the proposed two-phase methodology. A total of 18,004 <paragraph>–<section> pairs were selected using the FAISS library.

For 10,282 of these pairs, at least two LLMs identified an excerpt of legislative text in the <section> that applies to the associated <paragraph> with high confidence; Claude Sonnet 4.0, acting as an LLM-as-judge, then selected one of the two or three identified excerpts and added it to the output. For 4389 pairs, a single LLM identified such an excerpt with high confidence. Of the remaining 3333 pairs, 2257 were discarded because, for the other pair associated with the same paragraph, at least one LLM successfully extracted an excerpt with high confidence. The remaining 1076 pairs, corresponding to 538 paragraphs each associated with two sections, were submitted to Claude Sonnet 4.0 for regeneration; however, it identified a relevant excerpt with high confidence for only 32 of these pairs.

For the evaluation of these results, we did not pool the outputs into a single set and randomly select a subset for evaluation from this single set. Instead, we divided the outputs into four categories reflecting the pipeline's decision process, selected subsets from each category, and asked four legal experts to validate them under controlled conditions.

Figure 4
Summary of the final results



(11)

1) Directly accepted outputs, where at least one model produced a "High" confidence excerpt and no adjudication was needed;
2) Adjudicated outputs, where the two or more models extracted the excerpts and Claude Sonnet 4 was tasked with selecting the most plausible excerpt;
3) Regenerated outputs, where no satisfactory excerpt was produced and Claude Sonnet 4 regenerated a new candidate;
4) Discarded outputs, where the system decided that no reliable excerpt could be extracted.

Given that adjudicated outputs formed by far the largest subset of the data, we selected 800 pairs from category (11.b) for evaluation, ensuring sufficient statistical coverage of the most complex and contested cases. From the directly accepted outputs, which were numerous but more homogeneous, we sampled 200 pairs to confirm that high-confidence extractions were consistently correct. Since only 32 outputs were regenerated, we evaluated all of them exhaustively. Finally, we included a targeted sample of 200 discarded outputs to analyze failure modes, that is, category (11.d). This distribution provided a balance between breadth (covering the majority classes) and depth (examining all rare or difficult cases), while keeping the annotation workload manageable for the expert.

The four legal experts who conducted the evaluation each hold a law degree. Each evaluator was assigned a distinct, non-overlapping subset of pairs to annotate, using structured annotation files with written guidelines defining the evaluation criteria.

Evaluation of the positive outputs. For the evaluation of the positive outputs, we prepared structured text files to guide the legal experts in a consistent way across the different categories. Each record included the following fields:

1) paragraph text
2) section text
3) case law excerpt
4) legislation excerpt
5) model
6) reasoning

For the directly accepted, adjudicated, and regenerated outputs, two Yes/No questions were posed to the evaluators:

1) Does the legislative excerpt extracted by the system correctly apply to the paragraph text?
2) If the answer to (1) is "Yes," is the reasoning given by the model coherent and consistent with the principles of statutory interpretation?

By separating correctness of the legislative excerpt from correctness of the reasoning and by explicitly categorizing failure cases, the evaluation provided a comprehensive picture of both the accuracy and the explainability of Phase 2.

The results in Table 8 highlight the relative strengths of the three decision paths.

Table 8
Accuracy metrics for different decision paths

Decision path	Extraction accuracy %	Reasoning accuracy %
Directly accepted	97.00	78.40
Adjudicated	97.11	96.84
Regenerated	90.63	100.0

Adjudication achieves the highest performance (97.11% extraction, 96.84% reasoning). This outcome is expected, since the results are drawn only from cases where at least two voting models agree that a paragraph is relevant, with Claude Sonnet 4 then selecting the best result from the agreed candidates. The results suggest that this strict consensus mechanism effectively reduces false positives and ensures consistent reasoning quality.

The directly accepted outputs, where only one of the three voting models identifies a legislative excerpt applied to the <paragraph>, achieve extraction accuracy only 0.11% lower than that of the adjudicated outputs. However, their reasoning accuracy is substantially lower, suggesting that directly accepted outputs should perhaps be discarded in LegalTech applications where explainability and transparency in this task are particularly important.

With respect to regeneration, the results suggest that even the extracted excerpts should be discarded in LegalTech applications that prioritize precision over recall, since their accuracy is more than 6% lower than that of the directly accepted counterparts. The reasoning of the pairs where the identified excerpt was correctly recognized as an application of the law to the case law <paragraph> was always evaluated as correct. However, as noted above, the evaluation considered all 32 regenerated pairs, meaning that in only 3 of these was the identified excerpt found to be incorrect. It therefore remains questionable whether 32 pairs provide a statistically meaningful basis for evaluation.

Taken together, the three decision paths illustrate how consensus, recovery, and acceptance each contribute to a balanced pipeline that combines high accuracy with broad coverage. As noted above, a "cautious" LegalTech application that prioritizes precision over recall might choose to rely only on the adjudicated outputs, that is, (11.b). On the other hand, if the legal service provided by the application can tolerate occasional errors (i.e., when recall is prioritized over precision), it could also safely include pairs from categories (11.a) and (11.c).

Error analysis. For discarded outputs, where the system did not identify any legislative excerpt from the pair <paragraph>–<section>, the evaluation aimed not only to assess the correctness of the non-result but also to provide additional specific insights into why no legislative excerpt was extracted from the pair. The annotation file contained only the paragraph text and the associated section text. Rather than validating an extraction, annotators were asked to assign one of three reasons for the system's failure:

1) Reason A: the paragraph does not contain any "application of law."
2) Reason B: the provided section is not applied to the paragraph or is not relevant to it.
3) Reason C: the system failed to extract the case term or legislative term with high confidence despite its presence.

This design allowed us to evaluate not only whether the system produced correct outputs when it attempted an extraction but also whether its decisions to reject paragraphs were justified.

Table 9
Distribution of failure reasons

Reason of failure	%
Reason A	20.50
Reason B	Reason
Reason C	10.5%

After examining the failure reasons in Table 9, we concluded that cases in which the system failed to extract any excerpt (Reason C) are quite rare, estimated at 10.5% of 1044 pairs, that is, approximately 110 pairs, which corresponds to 0.6% of the 18,004 pairs retrieved by the FAISS library.

We also found that instances in which the <paragraph>s do not actually contain a legal interpretation (Reason A) are relatively rare, accounting for 20.5% of the cases examined. This indicates that our Phase 1 adjudication method performed well: even without many annotated examples and without a specially trained model, we were able to identify applications of the law with high confidence using only a few examples.

The major reason why Phase 2 discarded most of the pairs is the wrong linking of <paragraph>s with the <section>s (Reason B). Therefore, while the study overall demonstrates strong performance in identifying and linking statutory excerpts to their applications in case law, a closer examination reveals systematic limitations in the retrieval component, specifically with the FAISS library used in Phase 2 to generate the set of candidate pairs.

On the one hand, this issue arises from the nature of legal discourse itself: judicial paragraphs often engage with multiple legal concepts simultaneously. As a result, a paragraph may reference several statutory provisions, some central and others more peripheral. Even when only a single provision is actively applied, the retrieval mechanism may still return secondary sections based on superficial similarity or thematic overlap.

On the other hand, a second source of error stems from the fixed-k constraint used during retrieval. In our current setup, we retrieve the top two <section>s with the highest semantic similarity to each <paragraph>, based on dense vector representations. While this approach ensures computational efficiency and preserves interpretability, it carries the risk of omitting other relevant provisions that might better align with the legal reasoning in the paragraph. This trade-off is particularly pronounced in cases involving complex or layered legal argumentation, where multiple provisions may be applied implicitly or sequentially.

To address these shortcomings, more adaptive and feedback-driven retrieval mechanisms could be employed. One promising direction is the use of *Corrective Retrieval-Augmented Generation* (Corrective RAG), which integrates a feedback loop into the retrieval pipeline. In Corrective RAG, an initial retrieval is followed by an evaluation phase where the relevance of the retrieved content is scored. If the confidence in the top-ranked passages is low, the system triggers a corrective step, adjusting either the query formulation or expanding the candidate pool dynamically [45]. This could help recover missed matches and reduce retrieval noise in borderline cases.

Another complementary technique is *Speculative RAG*, which mitigates retrieval ambiguity by generating multiple candidate outputs based on different retrieval subsets and then uses a verifier model to select the most accurate final result [46]. This strategy enables the system to explore a broader interpretative space without committing prematurely to a single retrieval path, making it particularly well-suited for legal reasoning, where multiple provisions can be in conceptual play.

Incorporating these advanced RAG architectures into our pipeline would not only improve retrieval recall and precision but also provide more robust grounding for downstream legal phrase extraction. This would be especially impactful in ambiguous scenarios where legal reasoning depends on context-sensitive or implicit applications of statutory language.

5. Conclusions and Future Works

This paper presented a novel approach to bridging UK statutory law with its judicial application by combining LLMs with structured LegalDocML data.

The key contributions of this study are threefold. First, we introduce a replicable methodology for identifying phrase-level applications of UK statutory law in case law, combining ensemble LLM voting with structured LegalDocML data. Second, we demonstrate that multi-model adjudication achieves 97.11% extraction accuracy and 96.84% reasoning accuracy, establishing a strong baseline for this task. Third, we show that LegalDocML's machine-readable structure substantially simplifies preprocessing and enables reliable alignment of LLM outputs with formal legal structures, avoiding the errors commonly introduced by processing PDF or Word formats.

These results advance legal AI beyond existing approaches, which have largely operated at the document level [22] or focused on identifying holdings and factors [24, 25]. Our phrase-level extraction demonstrates that combining LLMs with structured LegalDocML data can address a granularity of legal analysis that prior methods have not yet targeted, while the ensemble adjudication mechanism ensures both accuracy and explainability.

We adapt a voting–adjudication framework to identify and explain how short excerpts from UK legislation are applied within UK case law paragraphs. The system integrates state-of-the-art LLMs, official LegalDocML documents from The National Archives, and advanced prompting strategies to support legal practitioners in tracing abstract legal concepts to real-world factual scenarios.

By focusing on short excerpts rather than entire sections, our approach avoids textual dispersion, enabling more precise matches, while retaining entire paragraphs to preserve contextual richness. This balance allows practitioners to efficiently scan results while consulting the full case when needed.

The system operates in two phases, combining two LLM methodologies in each to improve accuracy and explainability. At its core is a voting–adjudication methodology, in which several models independently assess the same input and cast "votes." An adjudicator then reviews these outputs to determine the most appropriate result, enhancing robustness, reducing model-specific bias, and highlighting disagreements for resolution, which is essential in legal domains with inherent ambiguity.

Moreover, the system uses chain-of-thought prompting to strengthen explainability. Models articulate reasoning step by step and self-assess confidence, making each decision interpretable. This structured reasoning provides traceable justifications, yielding results that are accurate and suitable for high-stakes legal workflows demanding clarity, consistency, and accountability.

Together, the dataset, hybrid LLM–LegalDocML framework, and dynamic prompting strategies lay the groundwork for next-generation legal AI applications, illustrating how data-driven and knowledge-driven approaches can achieve both accuracy and interpretability.

Several directions for future research emerge from this work. First, the methodology can be extended from application detection to nuanced treatment classification, distinguishing positive, negative, and scope-adjusting applications of legislation. Second, the approach generalizes to interpretive soft law sources such as codes of practice and regulatory guidance. Third, retrieval can be enhanced through advanced RAG architectures such as Corrective RAG, which could improve recall for low-confidence

candidates. Finally, stratified analyses by court jurisdiction, legal application type, and phrase length, as well as formal baseline comparisons and ablation studies, would further strengthen the evaluation. These directions would extend the current framework into a holistic knowledge base representing how legal concepts evolve across time, jurisdictions, and contexts.

Funding Support

This paper was supported by Innovate UK project "Odyssey—Opening the National Archives Legal Data to AI for Access to Justice (A2J)," Innovate UK 10106412 (https://gtr.ukri.org/projects?ref=10106412).

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

Livio Robaldo is an Editorial Board Member for *Journal of Computational Law and Legal Technology* and was not involved in the editorial review or the decision to publish this article. The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available at https://github.com/SafiaK/BridgingCaseLawAndLegislation.

Author Contribution Statement

Safia Kanwal: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing. **Livio Robaldo:** Conceptualization, Methodology, Validation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition. **Stergios Aidinlis:** Conceptualization, Validation, Writing – original draft, Writing – review & editing. **Joseph Anim:** Conceptualization, Validation, Writing – original draft, Writing – review & editing. **Davide Liga**: Writing – original draft, Writing – review & editing.

References

[1] Ajani, G., Boella, G., di Caro, L., Robaldo, L., Humphreys, L., Praduroux, S., . . . , & Violato, A. (2017). The European legal taxonomy syllabus: A multi-lingual, multi-level ontology framework to untangle the web of European legal terminology. *Applied Ontology*, *11*(4), 325–375.

[2] Beltrán, J. F., & Ratti, G. B. (2012). *The logic of legal requirements: Essays on defeasibility*. Oxford University Press.

[3] Duxbury, N. (1995). *Patterns of American* jurisprudence. Clarendon Press.

[4] Merryman, J., & Pérez-Perdomo, R. (2018). *The civil law tradition: An introduction to the legal systems of Europe and Latin America*. Stanford University Press.

[5] Peoples, L. F. (2025). Artificial intelligence and legal analysis: Implications for legal education and the profession. *Law Library Journal*, *117*(1), 52–85.

[6] Santosh, T. Y. S. S., Ashley, K. D., Atkinson, K., & Grabmair, M. (2024). Towards supporting legal argumentation with NLP: Is more data really all you need? In *Proceedings of the Natural Legal Language Processing Workshop 2024*, 404–421. https://doi.org/10.18653/v1/2024.nllp-1.36

[7] Shen, J., Xu, J., Hu, H., Lin, L., Ma, G., Zheng, F., . . . , & Han, W. (2025). A law reasoning benchmark for LLM with tree-organized structures including factum probandum, evidence and experiences. In *Findings of the Association for Computational Linguistics: ACL 2025*, 17252–17274. https://doi.org/10.18653/v1/2025.findings-acl.887

[8] Zhang, K., Yu, W., Sun, Z., & Xu, J. (2025). SyLeR: A framework for explicit syllogistic legal reasoning in large language models. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, 4117–4127. https://doi.org/10.1145/3746252.3761120

[9] Zheng, L., Guha, N., Arifov, J., Zhang, S., Skreta, M., Manning, C. D., . . . , & Ho, D. E. (2025). A reasoning-focused legal retrieval benchmark. In *Proceedings of the 2025 Symposium on Computer Science and Law*, 169–193. https://doi.org/10.1145/3709025.3712219

[10] Prakken, H. (2024). On evaluating legal-reasoning capabilities of generative AI. In *Proceedings of the 24th Workshop on Computational Models of Natural Argument*, 100–112.

[11] Kant, M., Nabi, S., Kant, M., Scharrer, R., Ma, M., & Nabi, M. (2025). *Towards robust legal reasoning: Harnessing logical LLMs in law.* arXiv. https://doi.org/10.48550/arXiv.2502.17638

[12] Dahl, M., Magesh, V., Suzgun, M., & Ho, D. E. (2024). Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis*, *16*(1), 64–93. https://doi.org/10.1093/jla/laae003

[13] Liga, D., & Robaldo, L. (2023). Fine-tuning GPT-3 for legal rule classification. *Computer Law & Security Review*, *51*, 105864. https://doi.org/10.1016/j.clsr.2023.105864

[14] Xie, Y., Li, Z., Yin, Y., Wei, Z., Xu, G., & Luo, Y. (2024). Advancing legal citation text classification A Conv1D-based approach for multi-class classification. *Journal of Theory and Practice of Engineering Science*, *4*(02), 15–22.

[15] Liga, D., & Palmirani, M. (2020). Transfer learning with sentence embeddings for argumentative evidence classification. In *20th Workshop on Computational Models of Natural Argument*.

[16] Zhang, G., Nulty, P., & Lillis, D. (2022). Enhancing legal argument mining with domain pre-training and neural networks. *Journal of Data Mining and Digital Humanities, NLP4DH,* 1–15. https://doi.org/10.46298/jdmdh.9147

[17] Grundler, G., Galassi, A., Santin, P., Fidelangeli, A., Galli, F., Palmieri, E., . . . , & Torroni, P. (2024). AMELIA - Argument mining evaluation on legal documents in Italian: A CALAMITA challenge. In *Proceedings of the 10th Italian Conference on Computational Linguistics*, 1125–1134.

[18] Sleimi, A., Sannier, N., Sabetzadeh, M., Briand, L., Ceci, M., & Dann, J. (2021). An automated framework for the extraction of semantic legal metadata from legal texts. *Empirical Software Engineering*, *26*(3), 43. https://doi.org/10.1007/s10664-020-09933-5

[19] Kanwal, S., Robaldo, L., Anim, J., & Liga, D. (2025). Leveraging LLMs and LegalDocML to extract legal interpretations: A case study on UK legislation and case law. In *Local Proceedings of the 19th International Workshop on Juris-Informatics,* 33-48.

[20] Greco, C. M., & Tagarelli, A. (2024). Bringing order into the realm of transformer-based language models for artificial intelligence and law. *Artificial Intelligence and Law*, *32*, 863–1010. https://doi.org/10.1007/s10506-023-09374-7

[21] Dehghani, F., Dehghani, R., Naderzadeh Ardebili, Y., & Rahnamayan, S. (2025). Large language models in legal systems: A survey. *Humanities & Social Sciences Communications*, *12*(1), 1977. https://doi.org/10.1057/s41599-025-05924-3

[22] Corazza, M., Zilli, L., & Palmirani, M. (2024). Topic similarity of heterogeneous legal sources supporting the legislative process. In *Proceedings of the 10th Italian Conference on Computational Linguistics*, 244–250.

[23] Li, J., Qian, L., Liu, P., & Liu, T. (2024). Construction of legal knowledge graph based on knowledge-enhanced large language models. *Information*, *15*(11), 666. https://doi.org/10.3390/info15110666

[24] Arvin, C. (2025). Identifying legal holdings with LLMs: A systematic study of performance, scale, and memorization. In *Proceedings of the Twentieth International Conference on Artificial Intelligence and Law*, 404–408. https://doi.org/10.1145/3769126.3769128

[25] Gray, M., Savelka, J., Oliver, W., & Ashley, K. (2024). Using LLMs to discover legal factors. In *Legal Knowledge and Information Systems: JURIX 2024: The Thirty-Seventh Annual Conference,* 60–71. https://doi.org/10.3233/FAIA241234

[26] Fink, M., & Finck, M. (2022). Reasoned A (I) dministration: Explanation requirements in EU law and the automation of public administration. *European Law Review*, *47*(3), 376–392.

[27] Liga, D. (2023). *The interplay between lawfulness and explainability in the automated decision-making of EU administration.* SSRN. http://doi.org/10.2139/ssrn.4561012

[28] Trajano, G., Engelmann, D. C., Bordini, R. H., Sarkadi, S., Mumford, J., & Panisson, A. R. (2024). Translating natural language arguments to computational arguments using LLMs. In *Computational Models of Argument: Proceedings of COMMA 2024*, 289–300. https://doi.org/10.3233/FAIA240329

[29] Guha, N., Nyarko, J., Ho, D., Ré, C., Chilton, A., Narayana, A., . . . , & Li, Z. (2023). LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 44123–44279.

[30] Palmirani, M., Sperberg, R., Vergottini, G., & Vitali, F. (2018). *Akoma Ntoso version 1.0 part 1: XML vocabulary*. OASIS Standard. https://docs.oasis-open.org/legaldocml/akn-core/v1.0/os/part1-vocabulary/akn-core-v1.0-os-part1-vocabulary.html

[31] Palmirani, M., & Vitali, F. (2011). Akoma-Ntoso for legal documents. In G. Sartor, M. Palmirani, E. Francesconi, & M. A. Biasiotti (Eds.), *Legislative XML for the semantic web: Principles, models, standards for document management* (pp. 75–100). Springer. https://doi.org/10.1007/978-94-007-1887-6_6

[32] Hoadley, D., Conroy, A., & Nemsic, E. (2023). Mission possible! Free access to case law and the national archives. *Legal Information Management*, *23*(1), 16–20. https://doi.org/10.1017/S1472669623000063

[33] Cochran, W. G. (1977). *Sampling techniques*. John Wiley & Sons.

[34] Lohr, S. L. (2019). *Sampling: Design and analysis* (3rd ed.). CRC Press.

[35] Israel, G. D. (1992). *Determining sample size*. University of Florida, Institute of Food and Agricultural Sciences. https://doi.org/10.32473/EDIS-PD006-1992

[36] Dietterich, T. G. (2000). Ensemble methods in machine learning. In *Multiple Classifier Systems: First International Workshop,* 1–15. https://doi.org/10.1007/3-540-45014-9_1

[37] Dong, X., Yu, Z., Cao, W., Shi, Y., & Ma, Q. (2020). A survey on ensemble learning. *Frontiers of Computer Science*, *14*(2), 241–258. https://doi.org/10.1007/s11704-019-8208-z

[38] Qiao, S., Ou, Y., Zhang, N., Chen, X., Yao, Y., Deng, S., . . . , & Chen, H. (2023). Reasoning with language model prompting: A survey. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, *1*, 5368–5393. https://doi.org/10.18653/v1/2023.acl-long.294

[39] Uma, A. N., Fornaciari, T., Hovy, D., Paun, S., Plank, B., & Poesio, M. (2021). Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, *72*, 1385–1470. https://doi.org/10.1613/jair.1.12752

[40] Davani, A. M., Díaz, M., & Prabhakaran, V. (2022). Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, *10*, 92–110. https://doi.org/10.1162/tacl_a_00449

[41] Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., ..., & Stoica, I. (2023). Judging LLM-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, 46595-46623.*

[42] Lu, J., Ma, K., Wang, K., Xiao, K., Lee, R. K.-W., Xu, B., . . . , & Lin, H. (2025). Is LLM an overconfident judge? Unveiling the capabilities of LLMs in detecting offensive language with annotation disagreement. In *Findings of the Association for Computational Linguistics: ACL 2025*, 5609-5626. https://doi.org/10.18653/v1/2025.findings-acl.293

[43] Schick, T., & Schütze, H. (2021). Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 255–269. https://doi.org/10.18653/v1/2021.eacl-main.20

[44] Braun, D. (2024). I beg to differ: How disagreement is handled in the annotation of legal machine learning data sets. *Artificial Intelligence and Law*, *32*, 839–862. https://doi.org/10.1007/s10506-023-09369-4

[45] Yan, S.-Q., Gu, J.-C., Zhu, Y., & Ling, Z.-H. (2024).*Corrective retrieval augmented generation.* arXiv. https://doi.org/10.48550/arXiv.2401.15884

[46] Wang, Z. R., Wang, Z., Le, L., Zheng, H. S., Mishra, S., Perot, V., . . . , & Pfister, T. (2025). Speculative RAG: Enhancing retrieval augmented generation through drafting. In *International Conference on Learning Representations,*18483–18505.

How to Cite: Kanwal, S., Robaldo, L., Aidinlis, S., Anim, J., & Liga, D. (2026). Identifying How UK Legislation Is Applied in Case Law: An Ensemble LLM Approach Using LegalDocML. *Journal of Computational Law and Legal Technology*, *1*, A1. https://doi.org/10.47852/bonviewJCLLT62029448