Detecting Legal Citations in United Kingdom Court Judgments

Holli Sargeant[*]
University of Cambridge
Cambridge, UK
hs775@cam.ac.uk

Andreas Östling[*]
Uppsala University
Uppsala, Sweden
andreas.ostling@statistik.uu.se

Måns Magnusson
Uppsala University
Uppsala, Sweden
mans.magnusson@statistik.uu.se

Abstract

Legal citation detection in court judgments underpins reliable precedent mapping, citation analytics, and document retrieval. Extracting references to legislation and case law in the United Kingdom is especially challenging: citation styles have evolved over centuries, and judgments routinely cite foreign or historical authorities. We conduct the first systematic comparison of three modelling paradigms on this task using the Cambridge Law Corpus: (i) rule-based regular expressions; (ii) transformer-based encoders (BERT, RoBERTa, LEGAL-BERT, ModernBERT); and (iii) large language models (GPT-4.1). We produced a gold-standard high-quality corpus of 190 court judgments containing 45,179 fine-grained annotations for UK and non-UK legislation and case references. ModernBERT achieves a macro-averaged F1 of 93.3%, only marginally ahead of the other encoder-only models, yet significantly outperforming the strongest regular-expression baseline (35.42% F1) and GPT-4.1 (76.57% F1).

1 Introduction

The vast volume of legislative documents, court judgments, and other legal texts produced globally represents a significant repository of structured knowledge, logical reasoning, and historical jurisprudence. However, the inherent complexity of legal language, with its specialised terminology, intricate cross-referencing conventions, and nuanced semantics, poses unique challenges for computational methods (Dale, 2019; Ganguly et al., 2023; Trancoso et al., 2023; Glogar, 2023; Koenecke et al., 2025). Accurately identifying references to specific legal entities, such as legislation and case law within these texts, is critical in many legal applications (Cross and Harris, 1991; Duxbury, 2008; Lewis, 2021; Koenecke et al., 2025). Precise extraction and linking of legal references enables various practical applications, including facilitating efficient legal NLP research supporting automated legal question-answering systems (Martinez-Gil, 2023; Siino et al., 2025), enhancing the consistency and transparency of legal reasoning (Zheng et al., 2025), and enabling data-driven legal scholarship (Carmichael et al., 2017; Schmid et al., 2021; Sargeant et al., 2025). Robust legal reference identification is indispensable for researchers and policymakers who investigate court efficiency (Posner, 2000; Fowler et al., 2007; Dalton and Singer, 2014; Bellutta and Carley, 2024), evaluate the real-world impact of judicial reforms (Blackwell, 2020; Tinarrage et al., 2024; Garoupa and Rao, 2025), examine judicial behaviours and biases (Choi and Gulati, 2008; Ash et al., 2024; Lindholm et al., 2025), and construct citation networks to visualise and analyse how laws evolve over time (Mones et al., 2021; Schmid et al., 2021; Mohammadi et al., 2024).

Our work aims to advance legal entity recognition (LER) and legal citation extraction for legislation and case law in the United Kingdom (UK). It is particularly critical to UK law due to the jurisdiction's reliance on both legislation and case law precedents; it is also particularly challenging due to the novel and complex referencing conventions of the UK and frequent references to international and foreign laws. While progress has been made in LER (Section 1.2), no other study on UK law has been conducted in this area. We address this gap by focusing on extracting and linking legal references–legislation and case law–from the UK's Cambridge Law Corpus (CLC, Östling et al., 2023).

1.1 Legislation and Case Law in UK Law

In the UK's common law jurisdiction, there are two major sources of authoritative law: *legislation* (Acts of Parliament, also known as statutes, and statutory instruments, also known as regulations) and *case law* (court decisions or judgments).

[*]Equal contribution.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 26810–26836
November 4-9, 2025 ©2025 Association for Computational Linguistics

Case law from previous court judgments form a body of precedent law that guide judges future decisions on the same or similar legal issues (Raz, 2009). It underscores the doctrine of *stare decisis* ("to stand by things decided"), which is applied relatively strictly in the UK (Duxbury, 2008). Higher court decisions can bind lower courts, while judgments from lower or parallel courts may retain persuasive authority (Cross and Harris, 1991; Lewis, 2021). Consequently, accurate citation to both legislation and case law is crucial in shaping arguments and clarifying points of law. For computational approaches to law, extracting correct legal references is crucial and one of the most important annotations for legal AI research (Sannier et al., 2017; Frankenreiter and Livermore, 2020).

UK Legal References Despite the importance of accurate and precise legal referencing, UK case law lacks uniformity in citation format. While referencing guidelines exist for legal academic writing (Meredith and Nolan, 2012), judges write legal references in court judgments in natural prose, rather than being restricted to standardised templates.

Legislation is generally referenced by its full title and year, such as the "Equality Act 2010". It may also be abbreviated, such as the "EA 2010" or "the 2010 Act". Case law references are defined by the name of the case (normally the opposing parties), the year, and a unique identifier of the specific judgment. There is no uniform identification, and the same court decision can be cited differently. In 2001, the neutral citation system was introduced to improve case identification; it refers to the court judgment by the court abbreviation and case number (MOJ, 2001, 2002), such as [2022] UKSC 25 (i.e., UK Supreme Court 2022, Case 25), compared to [2023] 2 All ER 303 (i.e., All England Law Reports 2023, volume 2, page 303). Pinpoint citations direct readers to specific parts of legislation or cases and can appear in various forms, see detailed examples in Appendix A and B.

Widespread references to other jurisdictions amplify the diversity of citation forms encountered in UK court judgments. European Union (EU) law, with distinct citation conventions, is frequently referred to in UK case law due to its EU membership from 1973 to 2020. UK courts often consider foreign laws for international business transactions and disputes that often elect to be governed by the law of England and Wales.

The historical influence of British colonisation means UK law remains in force, was in force at periods of the dataset, or remains a dominant influence in over 80 legal systems worldwide (Daniels et al., 2011; CIA, 2025). Importantly, the UK Privy Council, among other roles, was formerly the most superior court for the entire British Empire and still allows appeals from several countries.

1.2 Legal Entity Reference Extraction

Efforts to derive insights and extract information from legal texts date back to the 1990s (Turtle, 1995). Since then, natural language processing (NLP) has been employed for various related tasks, including legal information extraction and retrieval (Goebel et al., 2024; Joshi et al., 2023). Multiple studies have focused on named entity recognition (NER) for legal data. Most focus on the identification of individuals (e.g., judges, lawyers), locations, and organisations (Çetindağ et al., 2023). Current work draws on legal data from various jurisdictions, such as the United States (US) (Trias et al., 2021; Dozier et al., 2010), European Court of Human Rights (Cardellino et al., 2017), Germany (Leitner et al., 2020, 2019), Italy (Bellandi et al., 2024; Pozzi et al., 2023), France (Mathis, 2022), Brazil (Luz de Araujo et al., 2018), and Turkey (Çetindağ et al., 2023).

In addition to general NER, some studies specifically address the extraction of legal references. For example, Neale (2013) employs pattern matching and context-free grammars to parse Canadian case citations, while Agnoloni et al. (2017) introduces the BO-ECLI Parser Engine for automatically extracting legal references from European case law. Further, Harašta and Šavelka (2017) and Harašta et al. (2018) adopt Conditional Random Fields (CRF) to extract legal references and annotate Czech case law automatically. Although these papers touch on LER, they do not provide methods and results specifically for reference extraction.

Several studies employ regular expressions (RegEx) to address the extraction of legal references. Sadeghian et al. (2018) utilises a complex RegEx pattern-matching schema to extract legal citations. However, the authors provide only brief examples of US legislation patterns and do not discuss citation extraction performance in depth. Similarly, Milz et al. (2021) develops a German legal citation network by matching text that includes the symbol §, which refers to an article. This approach only identifies legislation references that specify a particular article number and, notably,

captures only about 7.34% of all citations. Milz et al. (2024) applies a comparable method to build a New Zealand legal citation network, using multiple RegEx patterns on a small dataset but offering limited details on the diverse citation conventions or performance. In a related effort, Sartor et al. (2022) parses XML text from the EU Court of Justice and combines pre-structured legal references with RegEx to locate pinpoint references more efficiently. Gheewala et al. (2019) builds an adaptable model using the RegEx-based Java Annotation Patterns Engine (JAPE) to extract legal citations within international law automatically. However, there is a notable absence of papers that provide evaluations of RegEx performance for legal entity extraction.

Supervised machine learning (ML) approaches have also been explored. Tran et al. (2014) employs maximum entropy and support vector machines to detect references in Japanese legal texts, achieving an F1 score of 80.06% for reference detection and 85.61% accuracy for reference resolution. Since the introduction of the transformer (Vaswani et al., 2017), these architectures are increasingly used for more accurate extraction. Correia et al. (2022) collects and manually annotates Brazilian Supreme Court rulings for multiple nested entity types. They compare CRF and BiLSTM-CRF models for entity annotation, reporting an F1 score of 0.91 for both models. Similarly, Bach et al. (2019) examines CRF and BiLSTM-CRF models for reference extraction on Vietnamese legal documents, obtaining F1 scores of 94.72 with CRF, 94.66 with BiLSTM, and 95.35 with BiLSTM-CRF. The same team also adopts a transformer-based encoder architecture to extract Vietnamese legal references, with a notably high F1 score of 99.4 for joint reference and relation extraction (Thuy et al., 2023).

To the best of our knowledge, there is no existing work evaluating LER methods for UK law.

1.3 Contributions

In this work, we provide a systematic study of legal entity and reference extraction for UK law. Our main contributions are:

1. *Methodological Evaluation:* We compare the performance of rule-based approaches (RegEx), pre-trained transformer encoders (BERT, RoBERTa, LEGAL-BERT, ModernBERT), and a large language model (GPT-4.1) on extracting legislation and case law references in UK court cases.
2. *Annotated Legal Entity Dataset:* We curate and release `CLC-Citation`, an expert-annotated dataset with detailed labels for UK and non-UK legal references.We also provide a detailed annotation schema for annotating legal entities.
3. *Insights for Jurisdiction-Specific LER:* We provide an in-depth analysis of model performance across various challenges in UK LER, yielding practical guidance and highlighting where further research could improve performance.

2 The CLC-Citation Corpus

We base our work on the Cambridge Law Corpus (CLC, Östling et al., 2023). Each case in the CLC includes both the textual content and relevant metadata. We use the same sample of annotated data used in Östling et al. (2023), i.e., a stratified random sample by court. However, due to technical issues with the annotation tool INCEpTION, not all annotations could be retrieved, leaving us with 190 cases from the original sample. At least one case from each court was included in the sample.

Three different legal scholars made the initial annotations. However, the initial annotations revealed inconsistencies among the three annotators, likely due to the task's complexity and the limitations of the initial schema. To address these issues, maximise annotation quality and minimise annotation errors, our first author[1] undertook a systematic re-annotation in three steps:

1. *Schema Refinement:* Introduced sub-labels that differentiate UK entities from non-UK entities, and established detailed guidelines to ensure consistent interpretation and annotation of each new sub-label (see Appendix A).
2. *Systematic Re-annotation:* Carefully re-annotated all 190 cases, relying on the refined annotation guidelines (see Appendix A).
3. *Cross-Validation Corrections:* Cross-validated the training and validation sets using an initial RoBERTa model (Section 3) to identify inconsistencies. Detected misclassifications were reviewed, and any annotation errors were corrected in the training and validation sets.

As an additional quality-control step and to avoid introducing bias when estimating the generalisation error, our first author annotated the test data twice, with more than six months between the annotation rounds. Where discrepancies were iden-

[1]Holds a qualifying law degree, a graduate degree in law, and is a non-practising solicitor.

	Train	Val.	Test	Total
Cases	110	30	50	190
Tokens	511130	146475	324551	982156
UKCR	8347	3065	4482	15894
UKSR	9492	3170	4282	16944
CR	773	122	120	1015
SR	4059	289	375	4723
CNP	1525	430	525	2480
CNC	2149	601	868	3618
CLRR	255	118	132	505

Table 1: The `CLC-Citation` Corpus. UK Case Reference (**UKCR**), UK Statute Reference (**UKSR**), Case Reference (**CR**), Statute Reference (**SR**), Case Name Parties (**CNP**), Case Neutral Citation (**CNC**), Case Law Report Reference (**CLRR**).

tified, the differences were resolved and a final annotation made. The agreement between the two versions was very high. Only 46 out of 10,249 sentences contained any difference; a Cohen's κ (Cohen, 1960) of 0.99 based on 324,551 token labels in the test set.

Once the annotations had been finalised, we converted each case document into the BIO2 format (Ramshaw and Marcus, 1995). Under this scheme, tokens linked to legal references are tagged as 'B-REFERENCE' (begin) or 'I-REFERENCE' (inside), labelling remaining tokens as 'Other'. Table 1 details the legal entity corpus composition.

During evaluation, we merge all tokens labelled as either 'B-REFERENCE' or 'I-REFERENCE' into a single positive class, treating all other tokens as negative. More details on the experiments can be found in Appendix G.

3 Experiments

We evaluate three distinct approaches to LER on our annotated dataset: (1) RegEx, (2) pre-trained transformer encoder models, and (3) decoder-type large language models. Performance is evaluated using the standard F1-score, precision, recall, the Jaccard Index, and Seqeval F1 at the token level. The Jaccard Index, also known as the intersection over union, is calculated as $\text{TP}/(\text{TP} + \text{FP} + \text{FN})$ (Jaccard, 1901). Sequeval F1 calculates the predictions at the full entity level and will only count as a true positive if the entire sequence of tokens is correctly classified (Nakayama, 2018).

Due to the challenges of reliably detecting references using RegEx, we created a second dataset that focuses exclusively on UK legal references. In this dataset, we treat the CR and SR categories from Table 1 as non-references during both training and evaluation. To enable a more granular analysis of model performance, we additionally evaluate results specifically for statutes and cases separately, disregarding all other labels.

3.1 Regular Expressions

As a baseline, we developed RegEx patterns to identify legislation and case law references within UK court judgments. We leveraged legal expertise to capture typical citation patterns observed in court judgments. Guided by illustrative examples (Appendix B), we developed separate RegEx patterns for *legislation* references and *case law* references. The case law expressions were categorised into *overinclusive* and *underinclusive* variations to address inconsistencies in case names.

The RegEx patterns underwent iterative refinement utilising the training and validation subsets of the `CLC-Citation`. Refinements adjusted for punctuation, expanded coverage of naming variations, and enforcement of other citation components (e.g., law report abbreviations and court identifiers). Our final patterns focus on capturing common citations of UK legislation and case law, with partial coverage of non-UK references (Appendix D).

3.2 Pre-Trained Encoder Models

We fine-tune and evaluate three pre-trained encoder models on the training set: BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), LEGAL-BERT (Chalkidis et al., 2020), and ModernBERT (Warner et al., 2024). LEGAL-BERT is specifically trained on legal-domain corpora. Notably, LEGAL-BERT is an uncased model (no capitalisation), while RoBERTa is available only in a cased version. To isolate the impact of capitalisation and tokenisation, we fine-tune both cased and uncased versions of BERT and compare their performance against RoBERTa and LEGAL-BERT. ModernBERT is a recently proposed transformer architecture that builds upon its predecessors with several key enhancements (Warner et al., 2024). It incorporates engineering advances from the development of LLMs, features a 16-fold increase in context length, more efficient attention mechanisms, and a significantly larger pre-training corpus. Two different sets of models were trained, one for UK-only references and one for all legal references.

For fine-tuning, we adopt a consistent training protocol for all encoder models. Each model under-

goes 50 training epochs, with 100 warm-up steps and a weight decay of 0.01. The learning rate decreases linearly from 5e-5 to 1e-7 after warm-up. Hyperparameters were kept consistent during the evaluations. Model predictions are made at the sentence level, with token-level labels predicted individually for each sentence. After each epoch, the F1 score is computed on the validation set, and the model with the best validation F1 during training is used for further analysis.

3.3 Decoder-Type Large Language Models

We also evaluate a transformer-based decoder model for the LER task, which are typically employed for generative tasks. Specifically, we investigate GPT-4.1 (OpenAI, 2025), selected for its balance between computational cost and model performance. Given the substantial volume of legal text, applying a more expensive and resource-intensive model may be impractical in many academic or production contexts.

We employ a task-specific prompting strategy (detailed in Appendix E), which guided GPT-4.1 toward token-level reference extraction. The prompt was constructed in a few-shot learning approach, starting with a system message followed by example sentences from the training set with expected output, and concluding with the sentence to be classified along with instructions to respond in TANL format (Paolini et al., 2021). The TANL format is designed to be more natural for LLMs, uses fewer tokens, and avoids limitations of other structured files by annotating tagged words and labels within curly brackets, such as: "...which fall within the scope of {article 8 | REFERENCE}".

In addition to static few-shot learning with a fixed set of examples, we also explored a dynamic approach, framed as a lightweight retrieval-augmented generation (RAG) setup (Lewis et al., 2020). For each test sentence, we computed OpenAI text-embedding-3-small embeddings and used cosine similarity to select the eight most similar training examples (with at least four containing legal references). This method allows us to test whether tailoring the prompt to the specific context of each case improves the model's in-context learning and overall extraction accuracy by grounding predictions in contextually relevant training data. The final experiment used $45 ($28 input, $17 output, $0.02 embeddings) to perform predictions for the 50 full test cases using the OpenAI API. All predictions are generated using a temperature of 0.

4 Results and Discussion

Tables 2 summarise the performance of all evaluated methods. Since the corpus is focused on UK cases, we also present the results for UK legal references and all legal references. Across both the UK-only and all labels, transformer-based encoders decisively outperform the other approaches. BERT, RoBERTa, LEGAL-BERT, and ModernBERT all attain an F1 score exceeding 90%, with ModernBERT on the full label set achieving the highest score at 93.3%. The GPT-4.1 models achieves marginally higher results on the full label set than the set of only UK labels, exhibiting excellent recall (96.74%) but markedly lower precision (full 63.36%), resulting in a F1 of 76.57% for the better, dynamic model. However, our RegEx baselines perform substantially worse, the highest performing achieving 35.42% F1.

4.1 Regular Expressions

Despite RegEx being a popular approach for LER in several other studies (Section 1.2), this rule-based approach establishes the lower-bound for citation extraction on the `CLC-Citation` test set.

The core *legislation* pattern achieves very high precision (>91%) yet very low recall ($\approx$ 7%), yielding an overall F1 of only 13.15% on the UK label set (12.84% on all). The *underinclusive* variant, which includes legislation and adds conservative case law patterns, raises recall to around 18% while maintaining precision near 94%, improving F1 to 29.71% for the UK label set (29.59% on all). Conversely, the *overinclusive* pattern, including legislation and looser citation forms to maximise coverage of case names, more than doubles recall to around 24% at a significant loss to precision at around 71%, producing the strongest RegEx result at 35.53% for the UK label set (35.42% on all). These results confirm that exhaustive coverage of varied UK citation practice cannot be achieved without sacrificing precision when using RegEx patterns. Switching from the UK-only to the full label set has a < 0.2% F1 effect on every RegEx pattern, suggesting that the additional non-UK citation styles lie largely outside the pattern coverage.

Error analysis Table 3 illustrates some common errors in the test set. For example, pinpoint and abbreviated legislation references are frequently missed because the patterns cannot adequately cover abbreviations, parentheses, and pinpoint syntax to sections or articles. Specific RegEx rules

Model	Label set	F1	P	R	Jaccard	Seqeval F1
Regex (Leg.)	UK	13.15%	91.70%	7.08%	7.04%	3.42%
Regex (Over.)	UK	35.53%	70.55%	23.74%	21.60%	5.25%
Regex (Under.)	UK	29.71%	93.72%	17.66%	17.45%	5.55%
BERT-Cased	UK	91.12%	90.96%	91.28%	83.69%	73.54%
BERT-Uncased	UK	91.38%	91.08%	91.67%	84.12%	74.55%
Legal-BERT	UK	92.16%	88.42%	96.22%	85.46%	74.35%
RoBERTa	UK	90.50%	87.95%	93.20%	82.65%	71.88%
ModernBERT	UK	91.08%	90.52%	91.65%	83.62%	72.88%
GPT-4.1-Static	UK	68.43%	53.73%	94.17%	52.00%	38.38%
GPT-4.1-Dynamic	UK	74.23%	60.25%	96.64%	59.02%	44.65%
Regex (Leg.)	All	12.84%	93.74%	6.89%	6.86%	3.20%
Regex (Over.)	All	35.42%	72.98%	23.38%	21.52%	4.93%
Regex (Under.)	All	29.59%	97.28%	17.45%	17.36%	5.45%
BERT-Cased	All	92.50%	92.44%	92.56%	86.05%	75.91%
BERT-Uncased	All	91.86%	92.55%	91.19%	84.95%	75.96%
Legal-BERT	All	92.72%	89.74%	95.92%	86.43%	77.38%
RoBERTa	All	90.34%	88.78%	91.95%	82.38%	67.96%
ModernBERT	All	93.30%	92.73%	93.88%	87.44%	81.65%
GPT-4.1-Static	All	70.49%	56.37%	94.05%	54.43%	40.90%
GPT-4.1-Dynamic	All	76.57%	63.36%	96.74%	62.04%	47.45%

Table 2: Performance comparison across UK and all labels including F1, Precision, Recall, Jaccard Index, and Seqeval F1 metrics (Test Data). Encoder models are fine-tuned separately on the UK and all label set.

cannot cover the wide range of reference formats (legislation error). Relaxed expressions identify entire leading or trailing clauses around a citation, producing large, imprecise spans (overinclusive error). Strict patterns fail to identify the beginning of longer case names accurately, omitting key identifying details (underinclusive error).

Even the strongest performance on the overinclusive RegEx pattern is substantially below transformer models and GPT-4.1. These findings underscore the inherent limitations of rule-based methods in handling the complexity and variability of four centuries of heterogeneous citation practice in UK case law.

4.2 Pre-Trained Encoder Models

Five transformer encoders were fine-tuned on the training portion of `CLC-Citation`: BERT (cased and uncased), RoBERTa, LEGAL-BERT, and ModernBERT. Performance on the held-out test set shows high performance across all pre-trained encoder models.

All models exceed 90% F1 score on both UK and full labels in the test set. The best result is obtained with ModernBERT trained on the full label set (93.3% F1). All achieving slightly higher results on the full label set, Legal-BERT ranks second at 92.72% F1, followed by BERT-cased (92.5%), BERT-uncased (91.86%) and RoBERTa (90.34%). LEGAL-BERT yields the highest recall (95.92%) but sacrifices precision (89.74%), whereas ModernBERT attains the highest precision (92.73%) while maintaining competitive recall (93.88%). RoBERTa achieved a similar recall of 91.95% to the other models, although it had the lowest precision (88.78%).

Across all transformer models, performance on case law references consistently exceeds performance on statute references. This disparity is evident in the UK-only and complete (all) label sets, where case references achieve higher F1 scores, driven primarily by superior recall (91–96%) and more balanced precision (Table 4).

The modest performance difference observed between LEGAL-BERT (92.72% F1) and BERT-Cased (92.50% F1) raises questions regarding the incremental value provided by domain-specific encoders in legal text-based tasks. One reason for the marginal gap may be that the LEGAL-BERT pre-training corpus lacks coverage of UK court judgments. Specifically, LEGAL-BERT was pre-trained on US case law, EU case law, and UK legislation, but no UK case law (Chalkidis et al.,

RegEx Example Errors
Legislation Error
[Company] has appealed the Determination under section 192(2) of the Communications Act 2003 (“the 2003 Act”) ,
Overinclusive Error
In relation to this question, *[Surname] v [Company Name]* [1988] 1 WLR 116 is the governing authority.
Underinclusive Error
This point was very clearly made in *[Company Name] v Office of Communications* [2008] CAT 11 at paragraph [164] :

Table 3: RegEx example errors. True Labels; True Positives ; False Negatives ; False Positives .

2020). Consequently, citation formats and conventions common in UK judgments remained largely unseen during pre-training. However, BERT-cased still performed better on UK statutes than legal BERT, indicating that including statutes in the pre-training data will not necessarily improve identification of statute references.

Additionally, the fine-tuning of both models may have been sufficient for BERT-cased to acquire a robust representation of UK references from scratch, narrowing the gap with LEGAL-BERT. Thus, the benefit of domain pre-training is diluted by a jurisdictional mismatch. It may also explain why the newer ModernBERT achieves the slightly higher performance (93.30% F1) on the full dataset.

Although the F1 difference is marginal between LEGAL-BERT and BERT-cased, LEGAL-BERT demonstrated a notable recall advantage (+3.4%) driven by superior recognition of statutory acronyms and paragraph-level pinpoint citations. In contrast, it simultaneously incurred a precision penalty (-2.7%) due to over-labelling errors, such as capitalised acronyms. In contrast, ModernBERT achieves a more balanced precision-recall profile than Legal-BERT and BERT-cased, achieving the highest precision of all encoder models (92.73%) while maintaining strong recall (93.88%).

ModernBERT's superiority is most evident in the Seqeval F1 metric, where its score of 81.65% significantly surpasses the next best models (Legal-BERT 77.38%, BERT-cased 75.91%). This indicates that ModernBERT is more accurate at identifying the precise boundaries of legal citations. This advantage is reinforced by the Jaccard Index, where ModernBERT again leads with 87.44%, edging out Legal-BERT (86.43%) and BERT-cased (86.05%). The higher overlap underscores its consistency in capturing full citation spans while minimising incorrect additional tokens at the edges of the citation.

Error analysis The test set contains a total of 10,249 sentences. Of these, we identified 848 sentences with a disparity between at least one of LEGAL-BERT, RoBERTa, and ModernBERT predictions at the token level. Some examples are seen in Table 5.

Overall, there are very few errors and places where the models have diverging predictions, making it hard to draw thorough conclusions. However, when we examine the errors made by the models, we see some tendencies. All evaluated transformer models perform substantially worse on citation types with very low frequency in the training set, notably historical legal references and academic citations.

Historical legal references, primarily dating from the 16th and 17th centuries, exhibit formatting and linguistic patterns distinct from modern conventions. These references typically appear as abbreviated reporters and often involve lexical forms uncommon in contemporary legal writing (historical reference error in Table 5).

Academic citations, such as references to legal journals or scholarly monographs, are not considered legal references but often take a similar citation style. The encoder models sometimes incorrectly label academic citations as legal references, although such errors were more commonly made by RoBERTa (academic citation error in Table 5).

RoBERTa frequently produced false negatives on certain patterns; for example, it misses several unusual reference formats, such as unpublished references and one-word case name abbreviations. Table 5 shows RoBERTa's false negative on the unpublished reference, which BERT-cased, BERT-uncased, LEGAL-BERT, and ModernBERT all correctly labelled; and Roberta's and ModernBERT's false negative on the abbreviated case reference, which BERT-uncased and LEGAL-BERT correctly labelled. Additionally, LEGAL-BERT and ModernBERT frequently incorrectly labelled one word

Model	Label set	Statute			Case		
		F1	P	R	F1	P	R
BERT-Cased	UK	85.70%	80.67%	91.40%	84.80%	81.02%	88.96%
BERT-Uncased	UK	85.01%	80.59%	89.95%	85.83%	81.47%	90.69%
Legal-BERT	UK	84.34%	75.77%	95.10%	85.50%	76.89%	96.28%
RoBERTa	UK	81.85%	74.62%	90.64%	84.10%	76.18%	93.86%
ModernBERT	UK	84.31%	79.49%	89.76%	85.49%	80.52%	91.12%
GPT-4.1-Static	UK	46.29%	31.26%	89.18%	51.16%	34.58%	98.32%
GPT-4.1-Dynamic	UK	54.98%	38.46%	96.36%	56.12%	39.63%	96.09%
BERT-Cased	All	88.44%	84.11%	93.25%	87.94%	83.91%	92.37%
BERT-Uncased	All	87.80%	84.28%	91.63%	87.70%	84.19%	91.51%
Legal-BERT	All	87.74%	79.35%	98.12%	86.17%	78.75%	95.14%
RoBERTa	All	83.44%	77.07%	90.95%	84.50%	77.38%	93.08%
ModernBERT	All	88.34%	84.39%	92.68%	88.52%	84.38%	93.10%
GPT-4.1-Static	All	49.70%	34.45%	89.19%	53.36%	36.62%	98.31%
GPT-4.1-Dynamic	All	59.05%	42.54%	96.53%	58.82%	42.36%	96.19%

Table 4: Performance comparison across UK and all labels for statute and case categories for the transformer models and GPT-4.1. Encoder models are fine-tuned separately on the UK and all label set.

abbreviations that are capitalised in the case text but lowercase in LEGAL-BERT and split into multiple tokens. Both incorrectly label "NCCN 956" as a reference, which is an acronym for Network Charge Change Notice. While it may have been expected that capitalisation would inform the model accuracy on such references, in many cases it doesn't. The tokenisation of these abbreviated words may resemble other legal reference abbreviations.

Overall, these results confirm that transformer architectures are highly effective for our task, but also reveal that subtleties such as capitalisation, older citation formats, and jurisdiction-specific conventions may necessitate further domain adaptation, fine-tuning or specialised training.

4.3 Decoder-Type Large Language Models

GPT-4.1 was evaluated in a few-shot, zero-temperature setting using the prompt shown in Appendix E. On the UK-only label set GPT-4.1-Dynamic achieves the higher F1 of 74.23%, and on the all label set 76.57%. These scores place the decoder model above the RegEx baselines but well below the fine-tuned encoder models reported in Table 4. GPT-4.1 exhibits high recall (94.05–96.74%) but notably lower precision (53.73–63.36%), a pattern that persists across all citation types.

The precision shortfall is most acute for UK references. The prompt used for GPT-4.1 deliberately does not differentiate between UK and non-UK references to avoid the need for negative examples, which would make the prompt excessively long and less efficient. As a result, the model tends to over-generalise when the evaluation label requires fine-grained UK-specific patterns. This effect is muted on the full label set, where a wider variety of citation formats makes the model's broad predictions less likely to be counted as false positives. However, the more narrowly defined UK statute and case references reveal lower performance (Table 4). For UK statute references, for instance, GPT-4.1-Dynamic recalls 96.36% of true labels yet achieves only 38.46% precision, yielding a modest 54.98% F1. While UK case references has slightly higher performance, of 96.19% recall and 39.63% precision, with a 56.12% F1. GPT-4.1-Dynamic is best at all UK and non-UK case references.

Error analysis GPT-4.1 tends to over-predict, suggesting that few-shot prompting lacks the granularity to accurately identify token boundaries in complex legal texts. Since `CLC-Citation` is heavily imbalanced (mostly negative), the model is overcautious, predicting positive labels for many examples in an effort not to miss any actual positives.

GPT-4.1 frequently labels proper nouns, especially capitalised multi-word spans, as legal references. General publications and names are often misclassified as legal entities (see Table 6). It is likely that the prompt examples, which predominantly feature capitalised legal entities and references to publications, create these implicit associations. Additionally, GPT-4.1 frequently overextends predicted spans, including adjacent words,

Encoder Error Examples
Historical Reference Error
–Ersk . 2, 6, 15; Duke of Queensberry, Mor. 14, 251 .
Academic Citation Error
See David Vaver, Without Prejudice Communications Their Admissibility And Effect [1974] U Br Col LR 85, at 97-101 , an article commended in Phipson , para 24-14 , fn 47
Capitalised Abbreviations
Essentially the NCCN 956 charges imposed a sliding scale of payments...
Academic Citation Error
Subrogation, therefore, is a remedy, not a cause of action: see Goff & Jones Law of Restitution 4th .
Unpublished Reference
Lord Donaldson MR in the unreported case *[Surname] v [Surname]* (21st May 1990) :
Abbreviated Case Name
...it is not open to this court to depart from the *[Surname]* line of authority,

Table 5: Pre-trained encoder model example errors. True Labels; True Positives ; False Negatives ; False Positives .

GPT-4.1 Error
By then he had read the articles in The Sunday Times of 25th March,
Ordinarily I would quote at some length from the judgment of [Judge Surname] LJ .
although subsequently repealed by the Defamation Act 2013 .

Table 6: GPT-4.1 example errors.True Labels; True Positives ; False Negatives ; False Positives .

punctuation, or introductory phrases. The prompt request that the LLM responds in TANL format, an approach that was generally successful (see Appendix G). However, this came with an increased extraction cost of approximately $1 per case.

5 Conclusion

We study three main approaches to LER, and our results demonstrate that pre-trained encoder models are highly effective, achieving especially high F1 scores even with a training set as small as 110 cases. This underscores the effectiveness of data-driven methods for capturing complex, evolving citation formats. Nonetheless, both encoder-based and decoder-based models exhibit lower performance when encountering historical UK references that deviate from modern norms, highlighting the need for more tailored domain adaptation. Our results also clearly point to the problem of using regular expressions for LER. Although this approach has been popular in previous research, our results clearly indicate that LER is too complex for Regex, with very low F1 scores. While GPT-4.1 achieves high recall, its extremely low precision, resulting in a lower F1 and comparably high cost of a little less than $1 per case (excluding prompt adaptation), currently makes it less practical for this task. Our findings confirm that flexible, context-aware encoder architectures deliver superior results across varied legal citation styles.

Accurate and reliable citation extraction is foundational for a wide range of downstream legal NLP applications. These include constructing citation networks for analysing legal precedent, enhancing legal information retrieval and summarisation systems, and tracking case law evolution. Furthermore, the dataset we have curated provides a valuable test set for future research in domain adaptation and comparative legal studies. Although our work focuses on UK judgments, the citation patterns and challenges observed are relevant to other common law jurisdictions, opening pathways for developing more robust, cross-jurisdictional models.

A promising direction for improving performance is to address the limitations of standard tokenisers. While domain-specific transformer-based models such as LEGAL-BERT show considerable promise for this task, the lack of cased, domain-specific training tailored to UK legal texts seems to limit performance. This points to working more on domain adaptation of encoder models for UK law, such as a "CLC-ModernBERT". The development of robust and adaptable models for citation detection for UK law requires both domain-specific innovation and interdisciplinary collaboration.

Limitations

Our findings should be interpreted in light of two main limitations.

Dataset scope. The annotated dataset comprises only 190 UK court judgments from the CLC. While all courts are represented, it still covers a limited range of cases. Given the variability of legal references across UK case law, it is likely that there may be certain courts, periods of time, and types of cases that would need additional annotations to improve the quality of performance.

Compute cost. Experiments with commercial large language models are prohibitively expensive. While we were able to run experiments on GPT-4.1, these experiments cost $45 for the final version for the 50 test cases. GPT-4o was initially used for testing but was left behind once GPT-4.1, which showed better results in initial smaller testing and was cheaper, was released. We explored using GPT-o1, although it was prohibitively expensive. Rough estimates showed that the cost for GPT-o1 would have been 6 to 20 times the cost of GPT-4.1, depending on the distribution of input-tokens, cached-tokens and output tokens. Even if such experiments revealed superior results, it would not be feasible to use this method for identifying legal references across the entire CLC, which contains over 320,000 cases.

GPT-4.1 was used in chat completion mode. This has the obvious downside that all output is raw text that later has to be parsed into a workable format. It would significantly reduce cost and likely increase performance if we had direct access to the embeddings and could use them for direct token prediction, similar to the BERT-style models.

Research Ethics and Impact Statement

We believe there is very minimal ethical risk to our current research, although we discuss two relevant considerations.

Data. Our annotated dataset is based on the underlying Cambridge Law Corpus (CLC). The CLC is a corpus of publicly available UK court judgments released under the Open Government Licence, which grants worldwide, royalty-free, perpetual and non-exclusive licence. Access to the CLC, and to this annotated dataset, is restricted to researchers under certain terms and conditions. For details on the legal and ethical considerations concerning the CLC dataset, see Östling et al. (2023). UK court judgments are typically not anonymised, reflecting the principles of open justice. However, courts may order anonymisation in specific situations, such as asylum cases or where required by law (e.g., victims of sexual offences and children in family law). Additionally, case names and references contain the name of parties, including individuals, companies and other organisations, so it is not practical or feasible to anonymise legal references in the dataset but we have anonymised within the paper. Our dataset retains the original names in judgments, aligning with the legal framework for court data in the UK.

This project received approval from the University of Cambridge Research Ethics Committee on 4 April 2022 and the Swedish Ethical Review Authority in May 2022.

Impact. Automating legal entity recognition can accelerate legal scholarship and improve access to justice by enabling large-scale corpus studies. There is a potential risk that, in subsequent work, *incorrectly identified citations could propagate into downstream analyses*, skewing empirical findings or leading to flawed legal arguments. Court judgments contain several documented biases (Sargeant and Magnusson, 2024), but we do not expect these biases to arise in our specific task of citation extraction. We encourage researchers in the field of legal NLP to carefully consider citation accuracy before relying on automated outputs.

Acknowledgments

The authors made limited use of Claude (Sonnet 3.5, 3.7), GPT (3.5-turbo, 4o, o1, 4.1) and Gemini (2.5) to accelerate Python code development and debugging. All generated code was tested and approved solely by the authors.

References

Tommaso Agnoloni, Lorenzo Bacci, Ginevra Peruginelli, Marc van Opijnen, Jos van den Oever, Monica Palmirani, Luca Cervone, Octavian Bujor, Arantxa Arsuaga Lecuona, Boada García, Alberto A, Luigi Di Caro, and Giovanni Siragusa. 2017. Linking European Case Law: BO-ECLI Parser, an Open Framework for the Automatic Extraction of Legal Links. In *Legal Knowledge and Information Systems*, pages 113–118. IOS Press, Amsterdam, The Netherlands.

Elliott Ash, Daniel L. Chen, and Arianna Ornaghi. 2024. Gender Attitudes in the Judiciary: Evidence from US Circuit Courts. *American Economic Journal: Applied Economics*, 16(1):314–350.

Ngo Xuan Bach, Nguyen Thi Thanh Thuy, Dang Bao Chien, Trieu Khuong Duy, To Minh Hien, and Tu Minh Phuong. 2019. Reference Extraction from Vietnamese Legal Documents. In *Proceedings of the 10th International Symposium on Information and Communication Technology*, pages 486–493, New York, NY, USA. Association for Computing Machinery.

Valerio Bellandi, Christian Bernasconi, Fausto Lodi, Matteo Palmonari, Riccardo Pozzi, Marco Ripamonti, and Stefano Siccardi. 2024. An entity-centric approach to manage court judgments based on Natural Language Processing. *Computer Law & Security Review*, 52:105904.

Daniele Bellutta and Kathleen M. Carley. 2024. Indicators of the formation of precedent at the International Court of Justice. *Social Networks*, 79:1–13.

Michael Blackwell. 2020. Indeterminacy, Disagreement and the Human Rights Act: An Empirical Study of Litigation in the UK House of Lords and Supreme Court 1997–2017. *The Modern Law Review*, 83(2):285–320.

Cristian Cardellino, Milagro Teruel, Laura Alonso Alemany, and Serena Villata. 2017. A low-cost, high-coverage legal named entity recognizer, classifier and linker . In *Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law*, ICAIL '17, pages 9–18, New York, NY, USA. Association for Computing Machinery.

Ian Carmichael, James Wudel, Michael Kim, and James Jushchuk. 2017. Examining the Evolution of Legal Precedent Through Citation Network Analysis. *North Carolina Law Review*, 96:227.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert: The muppets straight out of law school. *Preprint*, arXiv:2010.02559.

Stephen J. Choi and G. Mitu Gulati. 2008. Bias in Judicial Citations: A Window into the Behavior of Judges? *The Journal of Legal Studies*, 37(1):87–130.

CIA. 2025. Legal System, The World Factbook.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Fernando A. Correia, Alexandre A. A. Almeida, José Luiz Nunes, Kaline G. Santos, Ivar A. Hartmann, Felipe A. Silva, and Hélio Lopes. 2022. Fine-Grained Legal Entity Annotation: A Case Study on the Brazilian Supreme Court. *Information Processing & Management*, 59(1):102794.

Rupert Cross and J. W. Harris. 1991. *Precedent in English Law*, 4 edition. Clarendon Press, Oxford.

Can Çetindağ, Berkay Yazıcıoğlu, and Aykut Koç. 2023. Named-entity recognition in Turkish legal texts. *Natural Language Engineering*, 29(3):615–642.

Robert Dale. 2019. Law and Word Order: NLP in Legal Tech. *Natural Language Engineering*, 25(1):211–217.

Teresa Dalton and Jordan M. Singer. 2014. Bigger Isn't Always Better: An Analysis of Court Efficiency Using Hierarchical Linear Modeling. *Pace Law Review*, 34(3):1169.

Ronald Daniels, Michael Trebilcock, and Lindsey Carson. 2011. The Legacy of Empire: The Common Law Inheritance and Commitments to Legality in Former British Colonies. *American Journal of Comparative Law*, 59(1):111–178.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Christopher Dozier, Ravikumar Kondadadi, Marc Light, Arun Vachher, Sriharsha Veeramachaneni, and Ramdev Wudali. 2010. Named Entity Recognition and Resolution in Legal Text. In Enrico Francesconi, Simonetta Montemagni, Wim Peters, and Daniela Tiscornia, editors, *Semantic Processing of Legal Texts: Where the Language of Law Meets the Law of Language*, pages 27–43. Springer, Berlin, Heidelberg.

Neil Duxbury. 2008. *The Nature and Authority of Precedent*. Cambridge University Press, Cambridge, UK.

James H. Fowler, Timothy R. Johnson, James F. Spriggs, Sangick Jeon, and Paul J. Wahlbeck. 2007. Network Analysis and the Law: Measuring the Legal Importance of Precedents at the U.S. Supreme Court. *Political Analysis*, 15(3):324–346.

Jens Frankenreiter and Michael A. Livermore. 2020. Computational Methods in Legal Analysis. *Annual Review of Law and Social Science*, 16(1):39–57.

Debasis Ganguly, Jack G. Conrad, Kripabandhu Ghosh, Saptarshi Ghosh, Pawan Goyal, Paheli Bhattacharya, Shubham Kumar Nigam, and Shounak Paul. 2023. *Legal IR and NLP: The History, Challenges, and State-of-the-Art*, page 331–340. Springer Nature Switzerland.

Nuno Garoupa and Weijia Rao. 2025. Foreign Judges and Foreign Case Citations: A Study of the Hong Kong Court of Final Appeal. *Journal of Law and Courts*, page 1–26.

Akshita Gheewala, Chris Turner, and Jean-Rémi de Maistre. 2019. Automatic Extraction of Legal Citations using Natural Language Processing. In *Proceedings of the 15th International Conference on Web Information Systems and Technologies*, page 202–209, Portugal. SciTePress.

Ondřej Glogar. 2023. The Concept of Legal Language: What Makes Legal Language 'Legal'? *International Journal for the Semiotics of Law - Revue internationale de Sémiotique juridique*, 36(3):1081–1107.

Randy Goebel, Yoshinobu Kano, Mi-Young Kim, Juliano Rabelo, Ken Satoh, and Masaharu Yoshioka. 2024. Overview and Discussion of the Competition on Legal Information, Extraction/Entailment (COLIEE) 2023. *The Review of Socionetwork Strategies*, 18(1):27–47.

Jakub Harašta and Jaromír Šavelka. 2017. *Toward Linking Heterogenous References in Czech Court Decisions to Content*, pages 177–182. IOS Press, Amsterdam, The Netherlands.

Jakub Harašta, Jaromír Šavelka, František Kasl, Adéla Kotková, Pavel Loutocký, Jakub Míšek, Daniela Procházková, Helena Pullmannová, Petr Semenišín, Tamara Šejnová, Nikola Šimková, Michal Vosinek, Lucie Zavadilová, and Jan Zibner. 2018. Annotated Corpus of Czech Case Law for Reference Recognition Tasks. In Petr Sojka, Aleš Horák, Ivan Kopeček, and Karel Pala, editors, *Text, Speech, and Dialogue*, volume 11107, pages 239–250. Springer International Publishing, Cham.

Paul Jaccard. 1901. Étude comparative de la distribution florale dans une portion des alpes et du jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37(142):547.

Abhinav Joshi, Akshat Sharma, Sai Kiran Tanikella, and Ashutosh Modi. 2023. U-creat: Unsupervised case retrieval using events extraction. *Preprint*, arXiv:2307.05260.

Allison Koenecke, Jed Stiglitz, David Mimno, and Matthew Wilkens. 2025. Tasks and Roles in Legal AI: Data Curation, Annotation, and Verification. *Preprint*, arXiv:2504.01349.

Elena Leitner, Georg Rehm, and Julian Moreno-Schneider. 2019. Fine-Grained Named Entity Recognition in Legal Documents. In *Semantic Systems. The Power of AI and Knowledge Graphs*, pages 272–287, Cham. Springer International Publishing.

Elena Leitner, Georg Rehm, and Julian Moreno-Schneider. 2020. A Dataset of German Legal Documents for Named Entity Recognition . In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4478–4485, Marseille, France. European Language Resources Association.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Sebastian Lewis. 2021. Precedent and the Rule of Law. *Oxford Journal of Legal Studies*, 41(4):873–898.

Johan Lindholm, Mattias Derlén, and Daniel Naurin. 2025. A Source-Based Theory of Variation in Judicial Reasoning: Evidence from Sweden. *Journal of Law & Empirical Analysis*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *Preprint*, arXiv:1907.11692.

Pedro Henrique Luz de Araujo, Teófilo E. de Campos, Renato R. R. de Oliveira, Matheus Stauffer, Samuel Couto, and Paulo Bermejo. 2018. LeNER-Br: A Dataset for Named Entity Recognition in Brazilian Legal Text . In *Computational Processing of the Portuguese Language*, pages 313–323, Cham. Springer International Publishing.

Jorge Martinez-Gil. 2023. A survey on legal question–answering systems. *Computer Science Review*, 48:100552.

Bruno Mathis. 2022. Extracting Proceedings Data from Court Cases with Machine Learning . *Stats*, 5(4):1305–1320.

Sandra Meredith and Donal Nolan. 2012. *OSCOLA: Oxford University Standard for Citation of Legal Authorities* , 4 edition. Hart, Oxford.

Tobias Milz, Michael Granitzer, and Jelena Mitrović. 2021. Analysis of a German Legal Citation Network. In *Proceedings of the 13th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2021)*, pages 147–154, Portugal. SciTePress.

Tobias Milz, Elizabeth Macpherson, and Varvara Vetrova. 2024. Law in Order: An Open Legal Citation Network for New Zealand. In *Data Science and Machine Learning*, pages 211–225, Singapore. Springer Nature.

M. Mohammadi, L. M. Bruijn, M. Wieling, and M. Vols. 2024. Combining topic modelling and citation network analysis to study case law from the European

Court on Human Rights on the right to respect for private and family life. *Preprint*, arXiv:2401.16429.

MOJ. 2001. Practice Direction (Judgments: Form and Citation) [2001] 1 WLR 194.

MOJ. 2002. Practice Direction (Judgments: Neutral Citations) [2002] 1 WLR 346.

Enys Mones, Piotr Sapieżyński, Simon Thordal, Henrik Palmer Olsen, and Sune Lehmann. 2021. Emergence of network effects and predictability in the judicial system. *Scientific Reports*, 11(1).

Hiroki Nakayama. 2018. seqeval: A python framework for sequence labeling evaluation. Software available from https://github.com/chakki-works/seqeval.

Thom Neale. 2013. Citation Analysis of Canadian Case Law. *Journal of Open Access to Law*, 1(1):1–60.

OpenAI. 2025. Introducing gpt-4.1 in the api. Accessed: 2025-05-15.

Andreas Östling, Holli Sargeant, Huiyuan Xie, Ludwig Bull, Alexander Terenin, Leif Jonsson, Måns Magnusson, and Felix Steffek. 2023. The Cambridge Law Corpus: A Dataset for Legal AI Research. *Advances in Neural Information Processing Systems*, 36:41355–41385.

Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cicero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. 2021. Structured prediction as translation between augmented natural languages. *Preprint*, arXiv:2101.05779.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Richard Posner. 2000. An Economic Analysis of the Use of Citations in the Law. *American Law and Economics Review*, 2(2):381–406.

Riccardo Pozzi, Riccardo Rubini, Christian Bernasconi, and Matteo Palmonari. 2023. Named Entity Recognition and Linking for Entity Extraction from Italian Civil Judgements. In *Advances in Artificial Intelligence*, pages 187–201. Springer, Cham.

Lance Ramshaw and Mitch Marcus. 1995. Text Chunking using Transformation-Based Learning. In *Third Workshop on Very Large Corpora*.

Joseph Raz. 2009. *The authority of law: essays on law and morality*, 2 edition. Oxford University Press, Oxford.

Ali Sadeghian, Laksshman Sundaram, Daisy Zhe Wang, William F. Hamilton, Karl Branting, and Craig Pfeifer. 2018. Automatic semantic edge labeling over legal citation graphs. *Artificial Intelligence and Law*, 26(2):127–144.

Nicolas Sannier, Morayo Adedjouma, Mehrdad Sabetzadeh, and Lionel Briand. 2017. An automated framework for detection and resolution of cross references in legal texts. *Requirements Engineering*, 22(2):215–237.

Holli Sargeant, Ahmed Izzidien, and Felix Steffek. 2025. Topic classification of case law using a large language model and a new taxonomy for UK law: AI insights into summary judgment. *Artificial Intelligence and Law*.

Holli Sargeant and Måns Magnusson. 2024. Bias in Legal Data for Generative AI. In *Generative AI and Law (GenLaw '24')*.

Galileo Sartor, Piera Santin, Davide Audrito, Emilio Sulis, and Luigi Di Caro. 2022. Automated Extraction and Representation of Citation Network: A CJEU Case-Study. In *Advances in Conceptual Modeling*, volume 13650, pages 102–111. Springer, Cham.

Christian S. Schmid, Ted Hsuan Yun Chen, and Bruce A. Desmarais. 2021. Generative Dynamics of Supreme Court Citations: Analysis with a New Statistical Network Model. *Political Analysis*, 30(4):515–534.

Marco Siino, Mariana Falco, Daniele Croce, and Paolo Rosso. 2025. Exploring LLMs Applications in Law: A Literature Review on Current Legal NLP Approaches. *IEEE Access*, 13:18253–18276.

Nguyen Thi Thanh Thuy, Nguyen Ngoc Diep, Ngo Xuan Bach, and Tu Minh Phuong. 2023. Joint Reference and Relation Extraction from Legal Documents with Enhanced Decoder Input . *Cybernetics and Information Technologies*, 23(2):72–86.

Raphaël Tinarrage, Henrique Ennes, Lucas E. Resck, Lucas T. Gomes, Jean R. Ponciano, and Jorge Poco. 2024. Empirical analysis of Binding Precedent efficiency in the Brazilian Supreme Court via Similar Case Retrieval. *Preprint*, arXiv:2407.07004.

Oanh Thi Tran, Bach Xuan Ngo, Minh Le Nguyen, and Akira Shimazu. 2014. Automated reference resolution in legal texts. *Artificial Intelligence and Law*, 22(1):29–60.

Isabel Trancoso, Nuno Mamede, Bruno Martins, H. Sofia Pinto, and Ricardo Ribeiro. 2023. *The Impact of Language Technologies in the Legal Domain*, page 25–46. Springer International Publishing.

Fernando Trias, Hongming Wang, Sylvain Jaume, and Stratos Idreos. 2021. Named entity recognition in historic legal text: A transformer and state machine ensemble method. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages

172–179, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Howard Turtle. 1995. Text retrieval in the legal world. *Artificial Intelligence and Law*, 3(1):5–54.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *Preprint*, arXiv:2412.13663.

Lucia Zheng, Neel Guha, Javokhir Arifov, Sarah Zhang, Michal Skreta, Christopher D. Manning, Peter Henderson, and Daniel E. Ho. 2025. A Reasoning-Focused Legal Retrieval Benchmark. In *Proceedings of the 2025 Symposium on Computer Science and Law*, CSLAW '25, page 169–193, New York, NY, USA. Association for Computing Machinery.

A Legal Reference Annotation Guidelines

A.1 UK Statute References [UK_STATUTE_REFERENCE]

Definition References to primary and secondary UK legislation. The annotation should, where possible, include the name of the statute, type, and year. It also contains Bills and other parliamentary documents, such as the Hansard. However, it does not include government documents, such as guidance notes or circulars.

Legislation type There are two main types of legislation in the UK. Primary legislation, known as Acts of Parliament or Statutes, are named Act [1]. Secondary legislation, known as Statutory Instruments, are named by their type, including regulations [2], orders [3], codes [4], and rules [5]. Secondary legislation is sometimes cited with the statutory instrument number [6]. Parliamentary documents should be annotated with all information [7].

Older statutes Sometimes older statutes are referenced with the regnal year and chapter number [8] or [9].

Abbreviations Legislation is often abbreviated, particularly in cases that repeatedly refer to the same legislation or have a long, full name. It can be abbreviated by its name [10] or by year [11]. If abbreviated only as "the Act" or "the Directive" it is not annotated because of the lack of context [12].

Pinpoints A pinpoint is a reference to a particular part of the legislation. Statutes are composed of elements such as chapters (ch/chs), parts (pt/pts), sections (s/ss), subsections (sub-s/sub-ss), paragraphs (para/paras), subparagraphs (subpara/subparas), and schedules (sch/schs). Statutory instruments include regulations (reg/regs), rules (r/rr), and articles (art/arts). Pinpoint citations can either be written in full [13] or using these abbreviations [14]; both are annotated. For example, Section 15 can be written as "s 15", Subsection 1 of Section 15 as "s 15(1)", and Paragraph b of Subsection 1 of Section 15 as "s 15(1)(b)". Certain references omit the abbreviations, such as Civil Procedure Rules [15].

Examples

1 Equality Act 2010

2 Equality Act 2010 (Amendment) Regulations 2023

3 Equality Act 2010 (Age Exceptions) Order 2012

4 Statutory Code of Practice on Equal Pay

5 Civil Procedure Rules 1998

6 Equality Act 2010 (Age Exceptions) Order 2012, SI 2012/2466

7 ... they relied upon the Official Report (Hansard) of the proceedings in Standing Committee A on the Finance (No 2) Bill on 24 June 1993 at Cols 590.

8 Crown Debts Act 1801 (41 Geo 3 c 90) was the 90th Act to receive royal assent in the 41st year of the reign of George III.

9 ...he did not enfeoff by the deed. Quod nota. 10 H. 6, 7.

10 The relevant law is the Equality Act 2010 ("the EA").

11 The issue falls under the 2010 Act.

12 It is clear that the Act instead of the Directive applies.

13 Section 19 defines ...

14 ..., see Equality Act 2010, s 19(2).

15 This is an application under CPR 24.

A.2 Other Statute References [STATUTE_REFERENCE]

Definition References to legislation from other jurisdictions. Cases may either refer to legislation from other jurisdictions in the format that they would be cited in the respective jurisdiction [1] or in a similar style to the UK, but putting the relevant jurisdiction in parentheses [2]. EU legislation, including treaties, regulations, directives, decisions, recommendations, and opinions, follows a drastically different reference style and should be annotated as Statute_Reference even if it was enforceable UK law during its EU membership [3]. The annotation should follow, where appropriate, the additional guidance for UK Statute References.

Examples

1 We now consider the French approach, see *loi* n° 2019-22 du 23 mars 2019 de programmation 2018-2022 et de réforme pour la justice.

2 Similar language is used in Human Rights Act 1993 (NZ).

3 See, e.g., Council Directive 2000/78/EC of 27 November 2000 establishing a general framework for equal treatment in employment and occupation [2000] OJ L 303/16.

A.3 UK Case References [UK_CASE_REFERENCE]

Definition References to previous decisions by a UK court or that follow the UK reference style. The annotation should, where possible, include the entire case name and year, then either the neutral citation [1] or the volume, law report abbreviation and first page [2], or both [3].

Pinpoints A pinpoint refers to a specific paragraph of a judgment or page of a report. If the judgment has numbered paragraphs, the pinpoint to a particular paragraph number is often shown in square brackets [4]. If the pinpoint reference is to a page number, it is usually inserted after a comma [5]. There are many instances where a judge will refer to a pinpoint reference by writing out the paragraph or page, annotate all relevant information that forms part of the citation [6], but not extraneous text that becomes explanation [7].

Judges' names A reference may be made to the judge who wrote the judgment text. In some cases, multiple judges will write separate opinions within the same case. The judge's name should only be annotated where it is part of the formal citation [8], but not when it is part of the natural text explaining the case [9].

Abbreviations Where cases are abbreviated, the annotation should include all the relevant abbreviated information, such as the case name [10], case name and year [11], or neutral citation [12].

Cases from other jurisdictions with UK citation While the UK was an EU member, several EU cases were published in UK law reports. While these cases may be from EU jurisdiction, when published in UK law reports, they will be annotated as UK case references [13].

The UK Privy Council, among other roles, was formerly the court of last resort for the entire British Empire. Canada abolished Privy Council appeals in 1949, India and South Africa in 1950, Australia in 1986, and New Zealand in 2003. Currently, eleven Commonwealth countries outside the United Kingdom retain Privy Council appeals, in addition to various British and New Zealand territories. Therefore, several cases with subject matter from different jurisdictions will be included in our UK dataset and annotated as UK case references [14].

Not annotated There is a common way of referencing in natural text that we have decided not to

include in our annotations. A pinpoint citation may be made separate from the initial case citation, often after case quotes, where in parenthesis the judge name and pinpoint is referenced [15]. However, the same style is also used to refer to another judge's judgment or transcript within the same or an earlier case, so we have excluded this from citations to avoid common use as internal cross-references.

Examples

1 *[Company Name] v [Company Name]* [1947] EWCA Civ 1
2 *[Company Name] v [Company Name]* (1948) 1 KB 223
3 *[Company Name] v [Company Name]* [1947] EWCA Civ 1, (1948) 1 KB 223.
4 . . . , see *[Company Name] v [Company Name] and others* [2022] UKSC 25 [21]-[22].
5 . . . , see *[Company Name] v [Company Name] and others* [2024] AC 211, 230.
6 In *[Company Name] v [Company Name] and others* [2022] UKSC 25 at paragraphs 21 and 22, . . .
7 In *[Company Name] v [Company Name] and others*, Lord [Judge Surname] discussed this issue, in particular at paragraph 22 they said . . .
8 . . . , see *[Company Name] v [Company Name] and others* [2022] UKSC 25 [21]-[22], per Lord [Judge Surname].
9 As Lady [Judge Surname] pointed out in *[Company Name] v [Company Name] and others* [2024] AC 211 . . .
10 The court considered the test for *[Company Name]* reasonableness.
11 In *[Company Name], supra*, the issue in dispute . . .
12 Therefore, the court must consider reasonableness: [1947] EWCA Civ 1.
13 The EU Case C-464/01, *[Surname] v [Company Name]* [2005] ECR I-439, was published in UK law reports: *[Surname] v [Company Name]* (2006) QB 204.
14 An Australian case was appealed to the UK Privy Council: *[Surname] v [Government]* [1925] AC 338.
15 The meaning "includes . . . " (per Lady [Judge Surname], at 110).

A.4 Other Case References [CASE_REFERENCE]

Cases from other jurisdictions References to previous decisions by courts of other jurisdictions. Cases from other jurisdictions are usually in the format that they would be cited in the respective jurisdiction. In several countries, primarily jurisdictions based on English Common Law, the citations follow the same format as UK Case References. They are distinguishable by the difference in law report abbreviations or neutral citations [1]. Other jurisdictions have more distinct reference styles, such as the US [2] and the EU [3]. The annotation should follow, where appropriate, the additional guidance for the UK Case Reference.

Examples

1 As held in the Australian case *[Surname] v [Government]* (1992) 175 CLR 1, [1992] HCA 23, . . .
2 Recently, the US Supreme Court overturned *[Surname] v [Surname]* 410 US 113, 163–64 (1973).
3 . . . , see Case C–176/03 *Commission v Council* [2005] ECR I–7879 at paragraph 51.

A.5 Header Text Legal References [CASE_NAME_PARTIES, CASE_NEUTRAL_CITATION, CASE_LAW_REPORT_REFERENCE]

Definition References to the current case are contained in the header of the case judgment text, shown in the XML sample below. There are often three legal references that should be annotated in this header, but are not true "citations", so are tagged with separate labels. The case title contains the party names separated by "v" [1]. The case neutral citation is identified in the header, the label excludes the date [2]. In some case headers, the official law report references associated with the current case is listed [3].

Examples

```
<?xml version="1.0" encoding="UTF-8"?>
<basic_case_document
    xmlns="https://github.com/anon">
  <CLC-ID>5f91 ... 044</CLC-ID>
  <case_text><![CDATA[
England and Wales Court of Appeal (Civil
↪  Division) Decisions
```

```
<NAME> v <NAME> & <NAME>[1,
↪ CASE_NAME_PARTIES] [2019] EWCA Civ
↪ 125[2, CASE_NEUTRAL_CITATION] (19
↪ February 2019)
[2019] ICR 1155[3,
↪ CASE_LAW_REPORT_REFERENCE],
[2019] EWCA Civ 125[2,
↪ CASE_NEUTRAL_CITATION],
[2019] IRLR 545[3,
↪ CASE_LAW_REPORT_REFERENCE]

Neutral Citation Number: [2019] EWCA Civ
↪ 125[2, CASE_NEUTRAL_CITATION]
...
]]></case_text>
</basic_case_document>
```

B Example Legal References

B.1 UK Legal References

Primary Legislation

Legislation
Template: `Title | Act | Year`
Example: *Equality Act 2010*

Pinpoint
Template: `Title | Act | Year | Pinpoint`
Example: *Equality Act 2010* s 19

Pinpoint
Template: `Pinpoint | Title | Act | Year`
Example: Section 19 of the *Equality Act 2010*

Abbreviation
Template: `Title | Act | Year (abbrev)`
Example: *Equality Act 2010* (the "Act")

Abbreviation
Template: `Title | Act | Year (abbrev)`
Example: *Equality Act 2010* (EqA 2010)

In-text abbreviation
Template: `In-text abbreviation`
Example: I shall refer to the *Equality Act 2010* as "the 2010 Act".

Secondary Legislation

Regulation
Template: `Title | Regulations | Year`
Example: *Equality Act 2010 (Amendment) Regulations 2023*

Order
Template: `Title | Order | Year`
Example: *Equality Act 2010 (Age Exceptions) Order 2012*

Pinpoint
Template: `Title | Type | Year | Pinpoint`
Example: *Equality Act 2010 (Amendment) Regulations 2023* s 19

Pinpoint
Template: `Phrase with pinpoint`
Example: Section 19 of the *Equality Act 2010 (Amendment) Regulations 2023*

Abbreviation
Template: `Title | Type | Year | (abbrev)`
Example: *Equality Act 2010 (Age Exceptions) Order 2012* (the "Order")

In-text abbreviation
Template: `In-text abbreviation`
Example: Hereinafter referred to as the "Regulations".

Case Law

Neutral citation
Template: `Case name | [year] | abbrev | [number]`
Example: *[Government], R (On the Application Of) v Police Appeals Tribunal* [2024] EWHC 2348 (Admin)

Law-report citation
Template: `Case name | (year) | vol | Law Report | [number]`
Example: *R (on the application of [Surname]) v [Government] (Rev 1)* (2020) 3 WLR 1298

Pinpoint
Template: `Case name | (year) | Law Report | [number], [page]`
Example: *[Surname] v [Government]* (1893) QB 256, 287

Pinpoint
Template: `Pinpoint | Case name | [year] | abbrev | [number]`
Example: At para 334 of *[Company Name] v [Company Name] & Ors* [2022] UKSC 25, [2023] 2 All ER 303, Lord [Judge Surname] said . . .

Abbreviation
Template: `Case name | (year) | Law Report`

| [number] | (abbrev)
Example: *Regina ([Surname] and others) v [Government]* [2021] 1 WLR 2326 (*[Surname]*)

Abbreviation
Template: "Quote" Abbrev (at pinpoint)
Example: "excessively long documents serve to conceal rather than illuminate", *[Surname]* (at 2348 per Lord [Judge Surname] CJ)

B.2 EU Legal References

Primary Legislation

EU Directive
Template: Legislation type and number | Legislation title | [year] | OJ series | issue/first page
Example: Directive 2006/54/EC of the European Parliament and of the Council of 5 July 2006 on the implementation of the principle of equal opportunities and equal treatment of men and women in matters of employment and occupation (recast) [2006] OJ L 204/23.

EU Regulation
Template: Legislation type and number | Legislation title (abbrev) | [year] | OJ series | issue/first page
Example: Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (GDPR) [2016] OJ L 119/1

Abbreviation pinpoint
Template: abbrev | article/recital number
Example: GDPR art 22(1)

Case Law

Case
Template: Case number | Case name | [year] | ECR | volume-| first page
Example: Case C-170/84 *[Company Name] v [Surname]* [1986] ECR I-01607

Pinpoint (case number)
Template: Case number | pinpoint
Example: Case C-170/84, paras 47–48

C Dataset Statistics of Cases per Court

Court	Train	Val.	Test	Total
UKAITUR	18	2	13	33
EWCA-Civ	8	4	4	16
UKEAT	9	3	2	14
EWHC-Admin	6	2	2	10
UKET	6	1	3	10
UKICO	5	0	3	8
UKPC	4	0	3	7
UKFTT-TC	2	1	2	5
EWCA-Crim	4	0	0	4
EWHC-Ch	2	1	1	4
EWLVT	1	2	1	4
UKHL	2	0	2	4
EWHC-Comm	2	0	1	3
EWHC-Fam	1	1	1	3
EWHC-QB	2	0	1	3
EWHC-TCC	0	3	0	3
UKSSCSC	3	0	0	3
CAT	0	0	2	2
EWCC-Fam	1	0	1	2
EWCOP	2	0	0	2
EWCST	2	0	0	2
EWFC-OJ	1	1	0	2
EWHC-CP	1	1	0	2
EWHC-Exch	1	0	1	2
EWHC-KB	1	0	1	2
EWHC-Mercantile	1	0	1	2
EWLandRA	1	1	0	2
Misc	1	0	1	2
UKFSM	2	0	0	2
UKFTT-HESC	2	0	0	2
UKIAT	1	1	0	2
UKIT	0	0	2	2
UKSC	2	0	0	2
UKSPC	1	1	0	2
UKUT-AAC	2	0	0	2
UKUT-IAC	2	0	0	2
UKUT-LC	0	2	0	2
UKVAT	1	1	0	2
UKVAT-Customs	2	0	0	2
UKVAT-Excise	0	0	2	2
EWHC-Admlty	1	0	0	1
EWHC-IPEC	1	0	0	1
EWHC-Patents	1	0	0	1
EWLands	1	0	0	1
EWMC-FPC	1	0	0	1
EWPCC	1	0	0	1
UKFTT-PC	0	1	0	1
UKIPTrib	1	0	0	1
UKUT-TCC	1	0	0	1
UKVAT-Landfill	0	1	0	1
Total	110	30	50	190

D Regular Expressions Patterns

Legislation

```
\b
(?:
    # Legislation Reference Pattern
    (?P<legislation_reference>
        (?:
            # Pattern A: Legislation Reference followed by Pinpoint
            (?:the\s+)?  # Optional "the" at the start, not captured
            (?P<legislation_title>
                (?P<act_title>
                    (?:[A-Z][a-zA-Z]*  # First word starting with uppercase letter
                        (?:\s+
                            (?:of|the|and|for|in|on|by|to|\([^\)]*\)|[A-Z][a-zA-Z]*)
                                # Conjunctions or more uppercase words
                        )*
                    )
                )
                \s+
                (?P<type>Act|Rules|Order|Regulations)  # Legislation type
                \s+
                (?P<year>\d{4})  # Year
            )
            (?P<pinpoint_legislation_a>
                \s*,?\s*
                (?:[Ss]|[Ss][Ss]|[Ss]ection|[Ss]ections|
                  [Pp]t|[Pp]ts|[Pp]art|[Pp]arts|
                  [Cc]h|[Cc]hs|[Cc]hapter|[Cc]hapters|
                  [Ss]ch|[Ss]chs|[Ss]chedule|[Ss]chedules)
                \s*
                [\d\s*\-\(\)\/]+
            )?
        )
        |
        # Pattern B: Pinpoint following the Legislation Reference
        (?:
            (?P<pinpoint_legislation_b>
            (?:[Ss]ection|[Ss]|[Ss][Ss]|
              [Ss]ch|[Ss]chs|
              [Pp]t|[Pp]ts|[Pp]art|[Pp]arts|
              [Cc]h|[Cc]hs|[Cc]hapter|[Cc]hapters)
            \s*
            [\d\-\(\)\/]+
        )
        \s+of\s+(?:the\s+)?  # "of the" pattern, with "the" optional and not
            captured
        (?P<legislation_title_b>
            (?P<act_title_b>
                (?:[A-Z][a-zA-Z]*  # First word starting with uppercase letter
                    (?:\s+
                        (?:of|the|and|for|in|on|by|to|\([^\)]*\)|[A-Z][a-zA-Z]*)  #
                            Conjunctions or more uppercase words
                        )*
                )
            )
            \s+
            (?P<type_b>Act|Rules|Order|Regulations)  # Legislation type
            \s+
            (?P<year_b>\d{4})  # Year
        )
    )
)
)
\b
```

Legislation and Overinclusive Case Law RegEx

```
\b
    # Overinclusive Case Citation Pattern
    (?P<case_citation>
        (?P<case_name>
            (?:[A-Z][a-zA-Z',.\s()-]*\s*(?:v|and|of|the|for|in|on|by|to|&)\s+)+
            [A-Z][a-zA-Z',.\s()-]*
        )
        \s+
        (?P<date>\[\s*\d{4}\s*\]|\(\s*\d{4}\s*\))
        \s+
        (?:
            (?P<court_abbrev>[A-Z]{2,}(?:\s+[A-Z][a-z]+)*)
            \s*
            (?P<case_number>\d+)
            (?:\s*\(\s*(?P<division_after>Ch|Fam|QB|KB|Admin|Admlty|Comm|Pat|TCC)\s
                *\))?
            |
            (?P<volume_number>\d+)
            \s+
            (?P<law_report_abbrev>[A-Z][A-Za-z\s()&]+)
            \s+
            (?P<page_number>\d+)
        )
        (?P<pinpoint_case>
            (?:
                \s*,?\s*
                (?:
                    (?:at\s+)?(?:paragraph|para|paras|p|pp)\.?\s*
                    (?:\d+(?:\s*-\s*\d+)?|\[\d+\](?:\s*-\s*\[\d+\])?)
                    |
                    \[\d+\](?:\s*-\s*\[\d+\])?
                    |
                    ,?\s*\d+(?:\s*-\s*\d+)?
                )
            )
        )?
    )
\b
|
\b
(?:
    # Legislation Reference Pattern
    (?P<legislation_reference>
        (?:
            # Pattern A: Legislation Reference followed by Pinpoint
            (?:the\s+)?  # Optional "the" at the start, not captured
            (?P<legislation_title>
                (?P<act_title>
                    (?:[A-Z][a-zA-Z]*  # First word starting with uppercase letter
                        (?:\s+
                            (?:of|the|and|for|in|on|by|to|\([^\)]*\)|[A-Z][a-zA-Z]*)
                                # Conjunctions or more uppercase words
                        )*
                    )
                )
                \s+
                (?P<type>Act|Rules|Order|Regulations)  # Legislation type
                \s+
                (?P<year>\d{4})  # Year
            )
            (?P<pinpoint_legislation_a>
                \s*,?\s*
                (?:[Ss]|[Ss][Ss]|[Ss]ection|[Ss]ections|
                  [Pp]t|[Pp]ts|[Pp]art|[Pp]arts|
                  [Cc]h|[Cc]hs|[Cc]hapter|[Cc]hapters|
                  [Ss]ch|[Ss]chs|[Ss]chedule|[Ss]chedules)
                \s*
                [\d\s*\-\(\)\/]+
```

```
            )?
        )
        |
        # Pattern B: Pinpoint following the Legislation Reference
        (?:
            (?P<pinpoint_legislation_b>
            (?:[Ss]ection|[Ss]|[Ss][Ss]|
              [Ss]ch|[Ss]chs|
              [Pp]t|[Pp]ts|[Pp]art|[Pp]arts|
              [Cc]h|[Cc]hs|[Cc]hapter|[Cc]hapters)
            \s*
            [\d\-\(\)\/]+
        )
        \s+of\s+(?:the\s+)?  # "of the" pattern, with "the" optional and not
            captured
        (?P<legislation_title_b>
            (?P<act_title_b>
                (?:[A-Z][a-zA-Z]*  # First word starting with uppercase letter
                    (?:\s+
                        (?:of|the|and|for|in|on|by|to|\([^\)]*\)|[A-Z][a-zA-Z]*)  #
                            Conjunctions or more uppercase words
                    )*
                )
            )
            \s+
            (?P<type_b>Act|Rules|Order|Regulations)  # Legislation type
            \s+
            (?P<year_b>\d{4})  # Year
        )
    )
)
)
\b
```

Underinclusive Case Law and Legislation Reference RegEx

```
\b
    # Underinclusive Case Citation Pattern
    (?P<case_name>
        (?:[A-Z][a-z]+\s*(?:['']\s[A-Z]?[a-z]+)*\s*)+  # Words in case name
        (?:,?\s*(?:and|&|v\.?|vs\.?)\s*            # Connectors between parties
        (?:[A-Z][a-z]+\s*(?:['']\s[A-Z]?[a-z]+)*\s*)+)*  # More words after
            connectors
    )
    \s*
    (?P<date>\[\s*\d{4}\s*\]|\(\s*\d{4}\s*\))  # Date in square brackets or
        parentheses
    \s+
    (?:
        (?P<court_abbrev>[A-Z]{2,}(?:\s+[A-Z][a-z]+)*)  # Court abbreviation
        \s+
        (?P<case_number>\d+)
        (?:\s*\(\s*(?P<division_after>Ch|Fam|QB|KB|Admin|Admlty|Comm|Pat|TCC)\s*\))?
             # Division after (optional)
        |
        (?P<volume_number>\d+)
        \s+
        (?P<law_report_abbrev>[A-Z][A-Za-z\s()&]+)
        \s+
        (?P<page_number>\d+)
    )
(?P<pinpoint_case>
    (?:\s*,?\s*
        (?:
            (?:at\s+)?(?:paragraph|para|paras|p|pp)\.?\s*
            (?:\d+(?:\s*-\s*\d+)?|\[\d+\](?:\s*-\s*\[\d+\])?)
            |
            \[\d+\](?:\s*-\s*\[\d+\])?
            |
```

```
            ,\s*\d+(?:\s*-\s*\d+)?
        )
    )?
)
\b
|
\b
(?:
    # Legislation Reference Pattern
    (?P<legislation_reference>
        (?:
            # Pattern A: Legislation Reference followed by Pinpoint
            (?:the\s+)?  # Optional "the" at the start, not captured
            (?P<legislation_title>
                (?P<act_title>
                    (?:[A-Z][a-zA-Z]*  # First word starting with uppercase letter
                        (?:\s+
                            (?:of|the|and|for|in|on|by|to|\([^\)]*\)|[A-Z][a-zA-Z]*)
                                # Conjunctions or more uppercase words
                        )*
                    )
                )
                \s+
                (?P<type>Act|Rules|Order|Regulations)  # Legislation type
                \s+
                (?P<year>\d{4})  # Year
            )
            (?P<pinpoint_legislation_a>
                \s*,?\s*
                (?:[Ss]|[Ss][Ss]|[Ss]ection|[Ss]ections|
                  [Pp]t|[Pp]ts|[Pp]art|[Pp]arts|
                  [Cc]h|[Cc]hs|[Cc]hapter|[Cc]hapters|
                  [Ss]ch|[Ss]chs|[Ss]chedule|[Ss]chedules)
                \s*
                [\d\s*\-\(\)\/]+
            )?
        )
        |
        # Pattern B: Pinpoint following the Legislation Reference
        (?:
            (?P<pinpoint_legislation_b>
            (?:[Ss]ection|[Ss]|[Ss][Ss]|
              [Ss]ch|[Ss]chs|
              [Pp]t|[Pp]ts|[Pp]art|[Pp]arts|
              [Cc]h|[Cc]hs|[Cc]hapter|[Cc]hapters)
            \s*
            [\d\-\(\)\/]+
        )
        \s+of\s+(?:the\s+)?  # "of the" pattern, with "the" optional and not
            captured
        (?P<legislation_title_b>
            (?P<act_title_b>
                (?:[A-Z][a-zA-Z]*  # First word starting with uppercase letter
                    (?:\s+
                        (?:of|the|and|for|in|on|by|to|\([^\)]*\)|[A-Z][a-zA-Z]*)  #
                            Conjunctions or more uppercase words
                    )*
                )
            )
            \s+
            (?P<type_b>Act|Rules|Order|Regulations)  # Legislation type
            \s+
            (?P<year_b>\d{4})  # Year
        )
    )
)
)
\b
```

E GPT-4.1 Prompt

The following Python code is used to generate the prompt for the OpenAI API call. The `few_shot_examples` are the closest eight sentences, with at least four containing an entity example among the training data, based on the OpenAI `text-embedding-3-small` embeddings. An example output is shown in Appendix E.1

```
def construct_prompt(tokens, use_iob=False, uk_only=False,
    few_shot_examples=None):
    """Return a TANL-based prompt for the given list of *tokens*.

    * The model must output the sentence back in TANL format (
       entities wrapped in
      braces with a pipe and their label).
    * `use_iob` is kept only so existing callers don't break, but it
       is ignored
      - TANL is always requested now.
    * `few_shot_examples` is a list of dicts each containing keys "
       text" and "tanl".
    """

    sentence = " ".join(tokens)

    # Introductory instruction
    prompt = (
        "Convert the following text from a UK legal judgment into
           TANL format.\n"
        "Wrap every legal reference (statute or case law) with braces
            and append its label.\n"
        "The only label you should use is REFERENCE.\n"
        "Return ONLY the transformed text - no commentary, no JSON.\n
           \n"
    )

    # Few-shot section ----------
    if few_shot_examples:
        prompt += "Examples:\n\n"
        for i, ex in enumerate(few_shot_examples, 1):
            prompt += (
                f"{i}. Original:  {ex['text']}\n"
                f"   Output:    {ex['tanl']}\n\n"
            )
    else:
        prompt += (
            "Examples:\n\n"
            "1. Original:  The issue falls under the Equality Act
               2010 (\"the EA\").\n"
            "   Output:    The issue falls under the { Equality Act
               2010 | REFERENCE } (\"the EA\").\n\n"
            "2. Original:  Indirect discrimination is relevant here;
               see Equality Act 2010, s 19(2).\n"
            "   Output:    Indirect discrimination is relevant here;
               see { Equality Act 2010 , s 19(2) | REFERENCE }.\n\n"
```

```
            "3. Original:  See, e.g., Council Directive 2000/78/EC of
                27 November 2000 establishing a general framework for
                equal treatment in employment and occupation [2000]
               OJ L 303/16.\n"
            "   Output:    See, e.g., { Council Directive 2000/78/EC
               of 27 November 2000 | REFERENCE } establishing a
               general framework for equal treatment in employment
               and occupation { [2000] OJ L 303/16 | REFERENCE }.\n\n
               "
            "4. Original:  This issue has been considered before, see
                BTI 2014 LLC v Sequana SA and others [2022] UKSC 25
               [21]-[22], per Lord Reed.\n"
            "   Output:    This issue has been considered before, see
                { BTI 2014 LLC v Sequana SA and others [2022] UKSC 25
                [21]-[22] , per Lord Reed | REFERENCE }.\n\n"
            "5. Original:  In BTI, supra, the legal issue is the same
               .\n"
            "   Output:    In { BTI , supra | REFERENCE }, the legal
               issue is the same.\n\n"
            "6. Original:  The EU Case C-464/01, Gruber v Bay Wa AG
               [2005] ECR I-439, was published in other law reports.\
               n"
            "   Output:    The EU { Case C-464/01 , Gruber v Bay Wa
               AG [2005] ECR I-439 | REFERENCE }, was published in
               other law reports.\n\n"
            "7. Original:  It is clear that the Act instead of the
               Directive applies.\n"
            "   Output:    It is clear that the Act instead of the
               Directive applies.\n\n"
            "8. Original:  The meaning \"includes this definition\" (
               per Lady Hale, at 110).\n"
            "   Output:    The meaning \"includes this definition\" (
               per Lady Hale, at 110).\n\n"
        )

    # Add the actual text to be labelled
    prompt += "Text: " + sentence + "\n\nReturn ONLY the transformed
       text."  # Keep last instruction crystal-clear

    return prompt
```

E.1 Prompt output example

The prompt for the sentence:

```
Those properties were registered in the names of 2 Gibraltar
   companies namely Rosork Holdings Ltd ( " Rosork " ) and Fairlann
   Trading Ltd ( " Fairlann " ) ( together the " vendor companies " )
    .
```

is fed to the API as follows:

```
Convert the following text from a UK legal judgment into TANL format.
Wrap every legal reference (statute or case law) with braces and
   append its label.
The only label you should use is REFERENCE.
Return ONLY the transformed text - no commentary.

Examples:
1. Original:  The issue falls under the Equality Act 2010 ("the EA").
   Output:    The issue falls under the { Equality Act 2010 |
      REFERENCE } ("the EA").
2. Original:  Indirect discrimination is relevant here; see Equality
   Act 2010, s 19(2).
   Output:    Indirect discrimination is relevant here; see {
      Equality Act 2010 , s 19(2) | REFERENCE }.
3. Original:  See, e.g., Council Directive 2000/78/EC of 27 November
   2000 establishing a general framework for equal treatment in
   employment and occupation [2000] OJ L 303/16.
   Output:    See, e.g., { Council Directive 2000/78/EC of 27
      November 2000 | REFERENCE } establishing a general framework
      for equal treatment in employment and occupation { [2000] OJ L
      303/16 | REFERENCE }.
4. Original:  This issue has been considered before, see BTI 2014 LLC
    v Sequana SA and others [2022] UKSC 25 [21]-[22], per Lord Reed.
   Output:    This issue has been considered before, see { BTI 2014
      LLC v Sequana SA and others [2022] UKSC 25 [21]-[22] , per Lord
       Reed | REFERENCE }.
5. Original:  In BTI, supra, the legal issue is the same.
   Output:    In { BTI , supra | REFERENCE }, the legal issue is the
      same.
6. Original:  The EU Case C-464/01, Gruber v Bay Wa AG [2005] ECR I
   -439, was published in other law reports.
   Output:    The EU { Case C-464/01 , Gruber v Bay Wa AG [2005] ECR
      I-439 | REFERENCE }, was published in other law reports.
7. Original:  It is clear that the Act instead of the Directive
   applies.
   Output:    It is clear that the Act instead of the Directive
      applies.
8. Original:  The meaning "includes this definition" (per Lady Hale,
   at 110).
   Output:    The meaning "includes this definition" (per Lady Hale,
      at 110).

Text: Those properties were registered in the names of 2 Gibraltar
   companies namely Rosork Holdings Ltd ( " Rosork " ) and Fairlann
   Trading Ltd ( " Fairlann " ) ( together the " vendor companies " )
    .

Return ONLY the transformed text.
```

F Additional Results

Below are supplementary metrics that were used for model development and data enhancement. Table 7 shows F1, precision, recall, Jaccard Index and Seqeval F1 for UK and all label sets, and performance across statute and case references.

Model	Label set	Reference	F1	P	R	Jaccard	Seqeval F1
Regex (Leg.)	All	Case	0.00%	0.00%	0.00%	0.00%	0.00%
Regex (Over.)	All	Case	43.15%	62.01%	33.08%	27.51%	3.00%
Regex (Under.)	All	Case	35.27%	94.99%	21.65%	21.41%	5.79%
BERT-Cased	All	Case	87.94%	83.91%	92.37%	78.47%	57.05%
BERT-Uncased	All	Case	87.70%	84.19%	91.51%	78.09%	60.39%
Legal-BERT	All	Case	86.17%	78.75%	95.14%	75.70%	56.97%
RoBERTa	All	Case	84.50%	77.38%	93.08%	73.16%	48.95%
ModernBERT	All	Case	88.52%	84.38%	93.10%	79.41%	64.05%
GPT-4.1-Static	All	Case	53.36%	36.62%	98.31%	36.39%	23.86%
GPT-4.1-Dynamic	All	Case	58.82%	42.36%	96.19%	41.66%	25.04%
Regex (Leg.)	All	Statute	27.40%	93.74%	16.04%	15.87%	5.40%
Regex (Over.)	All	Statute	23.62%	44.39%	16.09%	13.39%	4.96%
Regex (Under.)	All	Statute	27.67%	93.46%	16.24%	16.06%	5.38%
BERT-Cased	All	Statute	88.44%	84.11%	93.25%	79.28%	75.18%
BERT-Uncased	All	Statute	87.80%	84.28%	91.63%	78.25%	74.29%
Legal-BERT	All	Statute	87.74%	79.35%	98.12%	78.16%	75.36%
RoBERTa	All	Statute	83.44%	77.07%	90.95%	71.58%	59.19%
ModernBERT	All	Statute	88.34%	84.39%	92.68%	79.12%	79.46%
GPT-4.1-Static	All	Statute	49.70%	34.45%	89.19%	33.07%	31.10%
GPT-4.1-Dynamic	All	Statute	59.05%	42.54%	96.53%	41.90%	34.15%
Regex (Leg.)	UK	Case	0.00%	0.00%	0.00%	0.00%	0.00%
Regex (Over.)	UK	Case	41.99%	58.90%	32.63%	26.58%	3.11%
Regex (Under.)	UK	Case	34.14%	88.60%	21.14%	20.58%	5.36%
BERT-Cased	UK	Case	84.80%	81.02%	88.96%	73.62%	54.47%
BERT-Uncased	UK	Case	85.83%	81.47%	90.69%	75.18%	57.22%
Legal-BERT	UK	Case	85.50%	76.89%	96.28%	74.67%	53.88%
RoBERTa	UK	Case	84.10%	76.18%	93.86%	72.56%	55.56%
ModernBERT	UK	Case	85.49%	80.52%	91.12%	74.66%	54.94%
GPT-4.1-Static	UK	Case	51.16%	34.58%	98.32%	34.38%	22.46%
GPT-4.1-Dynamic	UK	Case	56.12%	39.63%	96.09%	39.00%	23.30%
Regex (Leg.)	UK	Statute	28.82%	91.70%	17.10%	16.83%	5.90%
Regex (Over.)	UK	Statute	24.24%	41.67%	17.10%	13.79%	5.36%
Regex (Under.)	UK	Statute	28.81%	85.83%	17.31%	16.83%	5.83%
BERT-Cased	UK	Statute	85.70%	80.67%	91.40%	74.98%	71.22%
BERT-Uncased	UK	Statute	85.01%	80.59%	89.95%	73.93%	71.19%
Legal-BERT	UK	Statute	84.34%	75.77%	95.10%	72.92%	69.81%
RoBERTa	UK	Statute	81.85%	74.62%	90.64%	69.28%	63.61%
ModernBERT	UK	Statute	84.31%	79.49%	89.76%	72.88%	69.50%
GPT-4.1-Static	UK	Statute	46.29%	31.26%	89.18%	30.12%	28.20%
GPT-4.1-Dynamic	UK	Statute	54.98%	38.46%	96.36%	37.91%	30.88%

Table 7: Performance comparison with statute and case reference types.

G Experiment details

Precision, recall and F1 were computed using the `scikit-learn` library (Pedregosa et al., 2011). The Jaccard Index, also known as the intersection over union, is calculated as $\mathrm{TP}/(\mathrm{TP} + \mathrm{FP} + \mathrm{FN})$ (Jaccard, 1901). Sequeval F1 calculates the predictions at the full entity level and will only count as a true positive if the entire sequence of tokens is correctly classified (Nakayama, 2018).

During analysis, one additional reference in the test set was found to be incorrectly classified and corrected before the final results were computed.

For the GPT 4.1, analysis revealed that six output sentences were truncated from all test sentences because they exceeded 15,000 tokens. These sentences were excluded from the GPT analysis.

Label	**All**			**UK**		
	Binary	Statute	Case	Binary	Statute	Case
O	0	0	0	0	0	0
SR	1	1	N/A	0	0	N/A
UKSR	1	1	N/A	1	1	N/A
CR	1	N/A	1	0	N/A	0
UKCR	1	N/A	1	1	N/A	1
CNC	1	N/A	N/A	1	N/A	N/A
CNP	1	N/A	N/A	1	N/A	N/A
CLRR	1	N/A	N/A	1	N/A	N/A

Table 8: Label usage (Binary) across Full and UK train/-val/test datasets and label configurations for detailed evaluation (Statute & Case). Other (**O**), Statute Reference (**SR**), UK Statute Reference (**UKSR**), Case Reference (**CR**), UK Case Reference (**UKCR**), Case Neutral Citation (**CNC**), Case Name Parties (**CNP**), Case Law Report Reference (**CLRR**). 0 and 1 are used to denote negative and positive labels. Tokens corresponding to labels marked with N/A are ignored for computing performance metrics for those subsets.