

The EUR-Lex Retrieval Dataset: A Section-Aware and Human-Annotated Dataset for Long-Form Similar Case Retrieval

Christof Koolen

Visiting Professor

Centre for IT & IP Law, KU Leuven

christof.koolen@kuleuven.be

Abstract

This paper introduces the EUR-Lex Retrieval Dataset (EUR-LexRD): an evaluation benchmark for long-form similar case retrieval based on the case law handed down by the Court of Justice of the European Union (CJEU). The EUR-LexRD contains 20,697 English CJEU judgments published between 1954 and 2025. All cases are enriched with metadata and a case summary and have been segmented into six sections (e.g., legal context, facts, and reasoning). Each judgment also details outgoing citations to other European Case Law Identifiers.

The evaluation set features 100 query cases and a seeded candidate pool created with seven different retrieval strategies. The resulting benchmark consists of 12,365 query-candidate pairs, which were human-annotated by nine legal experts on a four-tiered relevance scale in order to establish a high-quality gold standard. The dataset’s features, namely the long document length and section segmentation, present new hurdles for retrieval and reranker models.

The EUR-LexRD is intended as a resource to advance the field of legal information retrieval by (i) making it possible to compare lexical, dense, hybrid and long-document retrieval strategies and (ii) supporting the fine-tuning of existing encoder or reranker models. The EUR-LexRD is released under a CC BY 4.0 license.

Dataset: <https://doi.org/10.48804/TK4IER>

Code: <https://github.com/christofkoolen/eurlx-retrieval-dataset>

1 Introduction

One of the more clearly defined tasks in the landscape of computational legal research is legal information retrieval (van Opijnen and Santos, 2017; Sansone and Sperlí, 2022). From lawyers looking for legal precedent to support their arguments, to judges combing through court databases to ver-

ify how a court has previously ruled on a particular topic, to legal scholars trawling journal article databases in search of high-quality contributions, all these situations have in common that they revolve around the retrieval of legal information. A particular subdomain is similar case retrieval or legal case retrieval, which revolves around finding judgments within a larger corpus of case law that are most similar to a given judgment (Wilson, 1962; Ma et al., 2021; Feng et al., 2024).

Developing and improving information retrieval systems hinges on the availability of suitable datasets and this is no different in the legal domain. In this respect, there are several factors that set legal information retrieval apart from information retrieval in a more general sense. For one, legal information retrieval has a more pronounced discrepancy between semantic similarity and legal relevance (Li et al., 2023a). A document could use wording that is slightly different from that of a search query and yet still be legally significant. Vice versa, a text that appears topically related at first glance might not be relevant at all because of a factual mismatch. Aside from this, legal documents also tend to be remarkably longer compared to more generalist texts and this poses challenges in its own right (Hou et al., 2025).

To further support scholarly research in this area, this paper introduces the **EUR-Lex Retrieval Dataset** (EUR-LexRD): a dataset created for the purpose of *long-form similar case retrieval*. The dataset contains the full text of 20,697 judgments handed down by the Court of Justice of the European Union (CJEU), accounting for quasi all English CJEU case law from the inception of the EU until 31 December 2025. It includes 100 structured queries and 12,365 human-annotated query-candidate pairs that are rated on a four-tiered relevance scale. All judgments are also segmented into different sections (e.g., legal context, facts and background, reasoning) and are further en-

riched with metadata and a case summary. As such, the dataset supports different “modes” of evaluation. For example, case summaries can be used as queries and full cases as candidates, as can any other combination of case sections. Additionally, since outgoing citations from each case are catalogued, retrieval analysis based on citation data is also possible.

The paper’s primary contribution is the public release of the EUR-LexRD as a new retrieval benchmark for EU case law. Its two further contributions are methodological. First, it supports the comparison of different interpretations of “legal relevance” through human-based and citation-based relevance judgments. Second, it sets out a framework to examine how the granularity of query and candidate representations can affect the performance of retrieval models. Taken altogether, these features set the EUR-LexRD apart from existing legal retrieval resources and support the evaluation of lexical, dense, hybrid, and long-document retrieval methods.

The paper is divided into six sections. Following the introduction, *Section 2* reviews the existing work on similar case retrieval datasets to situate the EUR-LexRD. *Section 3* is concerned with the methods and takes a closer look at the dataset itself. Here, particular attention is paid to how the data were collected (*Subsection 3.1*), how they were pre-processed (*Subsection 3.2*), how the candidate pool was seeded (*Subsection 3.3*), and how the annotation procedure was carried out (*Subsection 3.4*). *Section 4* then considers the results of these efforts and sheds light on the dataset’s main features and the inter-annotator agreement metrics. *Section 5* discusses potential use cases for the dataset. *Section 6* concludes the paper.

2 Related work: similar case retrieval datasets

2.1 Overview of existing benchmarks

Much of the initial development in case law retrieval was shaped by the Competition on Legal Information Extraction and Entailment (COLIEE). As a recurring annual competition since 2014, COLIEE datasets began to garner significant attention from 2020 onwards and, by now, form the basis of numerous conference proceedings and publications (Rabelo et al., 2021, 2022; Goebel et al., 2023, 2024, 2026). COLIEE remains one of the go-to evaluation sources for legal information retrieval

to this day, with its more recent datasets containing around 1,500 query cases and slightly fewer than 10,000 candidate cases, all of which are sourced from Canadian courts.

Another well-cited piece is the LeCaRD dataset. Released in 2021, LeCaRD is a legal case retrieval benchmark that consists of 107 query cases and a pool of approximately 43,000 candidate cases originating from China (Ma et al., 2021). Although relatively old by now, it is still featured regularly in scholarly works as a reference point (consider, for instance, Wang and Yu (2026)). In 2024, a new version of the dataset was released under the name LeCaRDv2. This newer version substantially expands the scope of the benchmark to include 800 query cases and 55,192 candidate cases (Li et al., 2024). Remaining jurisdictionally in China, several other notable datasets exist as well. For example, the annual China AI & Law Challenge has resulted in the publication of a number of datasets under the umbrella term “CAIL”, with the 2019-2022 editions being particularly popular (China AI & Law Challenge, n.d.; Xiao et al., 2019). Also worth citing is the MUSER dataset (Li et al., 2023b), which takes into account that different sections of a case can have a different impact on the retrieval results.

When turning our attention to Europe, emphasis should be placed on both ECtHR-PCR (Santosh et al., 2024) and GerDaLIR (Wrzalik and Krechel, 2021). While the former focuses on case law of the European Court of Human Rights, with an average of 5,283 candidate cases per query, the latter deals with German jurisprudence and contains approximately 123,000 queries across a collection of around 131,000 cases.

More recent datasets, such as the US case law-based LePaRD (Mahari et al., 2024) and CLERC (Hou et al., 2025), hint at a growing interest in so-called “passage-to-case retrieval”. In such a setting, short passages or quotes extracted from a judgment are used as queries, with the aim of retrieving the underlying precedent judgment to which the excerpt pertains. Notable too are several recent scholarly papers that focus on Korean case law (Kim et al., 2025; Cho et al., 2026).

2.2 Rationale for creating the EUR-LexRD

While the aforementioned datasets have been highly influential, it is difficult to transpose insights about evaluation benchmarks across different languages and, in a legal setting, also across differ-

ent jurisdictions. When considering the summary overview of datasets in *Appendix A*, it becomes apparent that the number of EU-based datasets is noticeably limited. Another point is that there have been interesting developments regarding the composition of retrieval datasets. For instance, as noted above, the MUSER dataset segments case law into different sections, as the strength of the retrieval signal can vary across these different sections (Li et al., 2023b).

Furthermore, a trend emerging in more recent datasets is an increase in average case length, which introduces input length constraints for contemporary retrieval models (Dan et al., 2026). For instance, Xiao et al. (2019) note that documents are often in excess of 512 tokens¹ and that this poses challenges. In MUSER, the average reported case length is 2,999 characters (Li et al., 2023b), whereas in LeCaRDv2 it is listed as 4,766 characters (Li et al., 2024). For the ECtHR-PCR dataset, an average word count of 5,530 is reported for the training set, although no character count is mentioned (Santosh et al., 2024). To date, input length constraints continue to pose an important hurdle when applying retrieval models to (lengthy) legal documents.

The aim of the EUR-Lex Retrieval Dataset is to complement these existing resources by making a new dataset available based on the case law of the CJEU that features: (i) cases segmented into different sections, (ii) human-annotated relevance labels, and (iii) relevance annotated on a graded scale.

3 Methods: data collection, pre-processing and annotation

3.1 Data collection

The source of input data for the EUR-Lex Retrieval Dataset is the case law published by Europe’s highest court, the CJEU. Its case law is made available via an online repository titled EUR-Lex.² Although the Court’s judgments are fully accessible via the website’s user interface, the options for accessing CJEU judgments in bulk are limited. For such purposes, researchers are directed to the EUR-Lex Webservice. After submitting a formal request that

specified the intended purpose and scope of the data collection activities, a username and a password were received. These were subsequently used to access EUR-Lex programmatically. All CJEU judgments published in English up to 31 December 2025 were targeted.

The data collection stage itself took place in two steps. First, the unique identifier of each case was obtained via the aforementioned EUR-Lex Webservice, together with the relevant case metadata (e.g., the case title, date of judgment, and document permalink). Second, the document permalinks were submitted to the EUR-Lex Cellar RESTful API in order to obtain the original full text of each judgment in HTML and plain-text formats.³ In total, 20,697 English CJEU judgments were aggregated, spanning from 21 December 1954 to 18 December 2025.⁴

3.2 Data pre-processing steps

Similar to the work of Shu et al. (2024), recourse was sought to a large language model (LLM) during the data pre-processing stage. In essence, the goal of the pre-processing pipeline is to ingest the raw judgment texts and convert them into structured JSON files. While the raw plain-text judgments could have been used, segmenting the judgments into smaller parts will ultimately allow for more granular experiments to be carried out with the dataset. Since only mild results could be obtained by segmenting the cases by means of regular expressions (*regex*) and since it would not have been possible for a human to produce the segmented case representations in a reasonable amount of time, an LLM was used instead.

The initial model selection was based on a series of trial runs with several AI providers while using a limited subset of cases. Aside from reasoning capabilities, the final model choice was influenced by the API pricing cost, the maximum context window length, and any applicable rate limits (i.e., input tokens per minute). Based on this assessment, *gemini-2.5-flash-lite* aligned best with our use case.

³As mentioned on the EUR-Lex website, the EUR-Lex Webservice only grants access to documents’ unique identifiers. These identifiers should then be used to obtain the actual source documents via the EUR-Lex Cellar RESTful API: <https://eur-lex.europa.eu/content/help/data-reuse/webservice.html>.

⁴Note that the EUR-Lex Webservice initially returned 20,726 unique cases. However, for 29 of these, no usable content could be retrieved via the case’s unique identifier.

¹Note that the paper mentions “characters” as opposed to tokens. However, since the paper involves experiments with BERT, it is assumed that the term “characters” was meant to refer to “tokens”, given that 512 tokens equates to the maximum context window of that model.

²See <https://eur-lex.europa.eu/>.

In a next step, additional pilot runs were carried out to optimize the model’s performance. At the outset, a temperature setting of 0.1 was chosen to get more deterministic output. The model was also afforded a thinking budget of 8,192 tokens. Generated responses were forced into a JSON format by explicitly specifying a response type in the API calls. At this stage, we also checked whether the best output could be achieved by using the HTML text or the plain-text version of each judgment. Based on this, the plain-text version of the judgments was used for the JSON conversion process.

Lastly, attention was devoted to optimizing the input prompt. Several iterations were drafted, tested, and then revised. In addition to enforcing a JSON output format, presenting the model with two examples of judgments and their correctly encoded JSON schemas led to the most notable quality uptick. The final version of the input prompt as well as an example of a filled-out JSON output schema are available in *Appendices B* and *C*.

The high-level instructions conveyed to the LLM focused on the following three components:

- (i) *Extraction of each case’s metadata.* Cases’ metadata were already obtained via the EUR-Lex Webservice, but some field values (e.g., case titles) were not always provided in the same format. The available case metadata were therefore passed on to the LLM to be rearranged in a consistent manner under the JSON parent key “case_metadata”.
- (ii) *Generation of a case summary.* For each judgment, an LLM-generated summary was appended to the JSON file. Generated output was kept separate from the actual text of the judgment by populating them into a separate key titled “case_summary”.
- (iii) *Segmentation of each case’s body text.* As its main task, the LLM was instructed to pinpoint the exact starting boundary of six predefined segments:
 - (a) The procedural introduction
 - (b) The legal context to the proceedings
 - (c) The facts and background to the proceedings
 - (d) The reasoning of the Court
 - (e) The decision on the costs
 - (f) The operative part

For each segment, heuristic cues were given to help the LLM identify the correct starting anchor. The anchor was then recorded by extracting the first 30 words of the segment verbatim. If a particular segment was not present in a judgment, the model was tasked to output {“start_anchor”: null}. All information was stored in the key “case_segments”.

The choice to split all judgments into six segments was largely influenced by the Rules of Procedure of the Court of Justice, which outline the different elements that need to be present in a CJEU judgment.⁵ To avoid over-segmentation, however, all information that pertained to the identification of the case (e.g., the case number, the parties’ names, the names of the parties’ representatives) was grouped in a single “procedural introduction” block. The other segments largely map back to one of the specific elements mentioned in the Rules of Procedure, with the only difference being that the facts are kept separate from the legal framework.

Once steps (i)-(iii) were completed, the full text of each judgment was reconstructed by means of a fully deterministic pipeline that used the positional indices of the start anchors together with the document start and end positions. Here, a strict forward-only ordering principle was enforced to guarantee that the full text was always rebuilt chronologically and in a non-overlapping manner. That said, the accuracy of the section boundaries still depended on the LLM correctly identifying the start anchors. To properly assess this, an audit was carried out on 100 cases to check the boundary selection. It was found that especially post-2000 case law was segmented well due to the more consistent document structure. For older CJEU judgments, however, the LLM failed at times to distinguish between “legal_context” and “facts_and_background” when there was no clear textual transition, when a segment was not clearly present, or when passages belonged to multiple segments. A good example in this respect is Case T-129/89, *Klaus Offermann v European Parliament* [1991] ECLI:EU:T:1991:55, in which one of the subheadings is titled “Facts and Procedure” and therefore relates to several sections at the same time. As a result, section boundaries in older judgments should be seen as less precise

⁵Article 87, Consolidated version of the Rules of Procedure of the Court of Justice of 25 September 2012

compared to those in more recent case law, in the sense that one section would occasionally cannibalize parts of an adjacent section. Note, however, that this only concerns the allocation of text between neighboring sections and not the reconstruction or ordering of the judgment as a whole.

As a last pre-processing step, a separate script was used to traverse all “case_segments” in search of any European Case Law Identifiers (ECLIs) cited in each judgment. All identifiers that matched the format ECLI:EU:[court]:[year]:[number] were collected and added under the JSON parent key “citations”. Three nested keys (“list”, “detailed”, and “edges”) contain, respectively, a list of unique cited ECLIs, a detailed list that indicates the section in which each citation occurred, and citation edges that link the citing case to each cited case. This structure supports the further use and analysis of citation information and the creation of network graphs.

3.3 Query and candidate selection

For similar case retrieval experiments to be carried out, an evaluation set of query-candidate pairs must first be created. Drawing inspiration from the work of Ma et al. (2021) and Li et al. (2023b), 100 judgments were randomly selected from the pool of 20,697 to serve as queries. A judgment was eligible to become a query case if its JSON file contained non-empty nested objects “facts_and_background” and “reasoning” within the parent key “case_segments”. Next, different sources were used to seed a pool of similar cases for each query; cases which were then subjected to human annotation:

- (i) *Citations*. All outgoing citations from the query case were included in the seed pool, since cited cases are likely to be relevant to the legal issues addressed in the query case.
- (ii) *Lexical retrieval*. The top- $k = 35$ candidates retrieved by both TF-IDF (Salton and Buckley, 1988) and BM25 (Robertson and Zaragoza, 2009) were added. These methods rely on lexical overlap, whereas BM25 also accounts for term importance and document length.
- (iii) *Dense embedding retrieval*. The top- $k = 35$ candidates retrieved by each of the embedding models all-MiniLM-L6-v2, Qwen3-Embedding-8B (Zhang et al., 2025),

and gemini-embedding-001 (Lee et al., 2025) were included. When creating embeddings, cases’ full-text representations were used on both the query and candidate side.

- (iv) *Random negatives*. To balance the pool and ensure the presence of clearly irrelevant examples, 10 randomly sampled negatives were also added per query.

For each query case, all results returned during the seeding step were unioned and deduplicated. This approach eventually yielded an average of 123.65 seeded candidate cases per query case.

The intent behind using this broad range of seeding methods is to minimize bias towards one particular retrieval method or the other (Buckley et al., 2006). Regarding the selection of the different embedding models, the rationale was to feature a range of models with different capabilities: a lightweight local embedding model, a state-of-the-art open-source embedding model, and a closed-source embedding model that can only be accessed via an external inference provider. As such, the selected models offer good coverage in terms of their size, maximum context window, and embedding dimensionality.

3.4 Data annotation

In the next phase, an annotation sheet was created based on the query-candidate collection. Nine legal experts subsequently went over this 12,365-row document and assigned a relevance score to each query-candidate pair based on cases’ summary and full text. The grading itself was done using a four-step labeling scale:

- (i) *Label grade 0 – Not relevant*. The candidate case has no meaningful legal usefulness in relation to the query case.
- (ii) *Label grade 1 – Marginally relevant*. The candidate case has some general or contextual connection to the query case, but offers little real value in resolving it.
- (iii) *Label grade 2 – Substantially relevant*. The candidate case is clearly relevant and would be useful for legal analysis in relation to the query case, despite not constituting a directly relevant case.
- (iv) *Label grade 3 – Directly relevant*. The candidate case directly addresses the same issue

and would be a key authority for resolving the underlying issue in the query case.

All nine annotators were selected from the KU Leuven’s Master of Laws program. This means that they had all completed at least a three-year Bachelor of Laws degree, which includes compulsory modules on European Union law. Furthermore, all annotators were paid for their data annotation work under the university’s standard student employment contract.

Prior to the annotation process, the group of annotators underwent two hours of annotation calibration training. First, all students received information about the project’s overall intent and objectives. As part of this, they were given a bundle of annotation guidelines (available in the [EUR-LexRD repository](#)). Next, they underwent a three-step training process. In the first round, they completed an exercise using a toy dataset. This was intended to familiarize the annotators with the annotation environment and explain how to submit responses. Initially, clear negative (label 0) and clear positive (label 3) query-candidate pairs were shown, before moving on to more nuanced pairs (labels 1 and 2). The second round involved a group calibration exercise with 10 query-candidate pairs, where the final labels were discussed collectively, along with the reasons for choosing or not choosing a particular label. The third round involved an individual calibration set of 15 query-candidate pairs. Once the answers to the individual calibration set had been submitted, the students received the worked solutions to all the annotation calibration exercises, along with individually tailored feedback based on their responses.

Following the completion of the training component, the nine annotators were granted access to their respective annotation sheets and were given one month to complete their annotations. Each annotator was tasked with annotating 1,548 or 1,547 rows of query-candidate pairs, of which:

- (i) 150 rows were shared among all annotators.
- (ii) 80 rows were double-annotated with one other annotator.
- (iii) 1318 or 1317 rows were single-annotated.

When looking at all 12,365 unique query-candidate pairs, this implies that 11,855 of those pairs (95.9%) were annotated by a single annotator, 360 pairs (2.9%) were annotated by two annotators,

and 150 pairs (1.2%) were annotated by all nine annotators.

The inclusion of (i) and (ii) was done to permit measuring inter-annotator agreement metrics. To that end, the shared and double-annotated rows were inserted randomly in the dataset, but they could only appear after a warm-up block of 150 single-annotated rows. The final relevance label was assigned by simple majority vote or, in the event of an *ex aequo*, the tie was resolved by the author. The goal of this human annotation process is to create a high-quality gold standard to evaluate the performance of different retrieval methods. This is because, although the seeding methods generate a pool of cases that are possibly relevant to the query cases, they do not establish a ground truth.

3.5 Benchmark representations and evaluation protocols

The dataset supports two evaluation settings. First, for retrieval systems that do not use the EUR-LexRD relevance judgments during training, hyperparameter selection, or any other stage of model development, the evaluation is to be carried out on the complete set of 100 queries (e.g., zero-shot and unsupervised methods). Second, for retrieval systems that do rely on the EUR-LexRD relevance judgments for the purpose of developing or fine-tuning a model, the benchmark contains a standardized query-level split that consists of 70 training, 10 development, and 20 test queries. In addition, five query-level cross-validation folds are also made available.

4 Results: the EUR-Lex Retrieval Dataset

The current section takes a closer look at the results of the data annotation efforts. To start with, [Table 1](#) presents an overview of the average character length and average word count for the different case segments. Full cases have 35,295 characters or 5,762 words on average and, for informational purposes, the longest document in the dataset⁶ has a total character length of 987,455 and spans 157,638 words. Turning to the citation insights, the average number of in-degree and out-degree citations per case is 4.90.

To quantify the human annotation efforts, a breakdown of the binary and graded judgment dis-

⁶The judgment in question is Joined cases T-191/98, T-212/98 to T-214/98, *Atlantic Container Line AB and Others v Commission of the European Communities* [2024] ECLI:EU:T:2003:245.

Section	Chars	Words	%
Full text	35,295	5,762	100.0
Procedural introduction	2,210	359	6.2
Legal context	4,867	785	13.6
Facts and background	6,422	1,049	18.2
Reasoning	20,641	3,377	58.6
Decision on costs	392	68	1.2
Operative part	755	123	2.1
Summary	1,645	242	—

Table 1: Overview of the average text length (in character count and in word count) for the different case segments. Note that “summary” is not part of the judgment’s full text. The percentage counts are relative to the average full text length.

Judgment	Label	Count	%
Binary	0 – Not relevant	9,215	74.5
	1 – Relevant	3,150	25.5
Graded	0 – Not relevant	5,223	42.2
	1 – Marginally relevant	3,992	32.3
	2 – Substantially relevant	2,345	19.0
	3 – Directly relevant	805	6.5

Table 2: A breakdown of the distribution of relevance labels as assigned by the annotators.

tribution is presented in *Table 2*. On the binary relevance scale, obtained by binning labels 0 and 1 to “not relevant” and labels 2 and 3 to “relevant”, the dataset is dominated by non-relevant query-candidate pairs (74.5%) whereas the relevant pairs account for 25.5%. Turning to the graded relevance scale, “not relevant” (42.2%) and “marginally relevant” (32.3%) are found to be the most frequent classes, followed by “substantially relevant” (19.0%) and then “directly relevant” (6.5%). Hence, while directly relevant cases are relatively rare, over half of the judgments (57.8%) display some degree of relevance.

Table 4 also outlines the dataset’s composition and shares insight into the relationship between the total document volume and the average number of candidate cases that were identified as “positive” for the query cases in particular.

4.1 Inter-annotator agreement metrics

As a next step, two different types of inter-annotator agreement metrics were taken into account: agreement among all annotators and agreement between pairs of annotators. *Table 3* presents these metrics.

Apropos the percentage of exact agreement, across all shared and double-annotated query-candidate pairs, annotators assigned the same label in 60.3% of cases. The highest boundary-specific disagreement occurred between labels 1 and 2. In those instances, annotators disagreed in 40.26% of

Inter-annotator agreement metric	Score
Exact agreement (graded)	60.3%
Within-one-grade agreement	93.3%
Exact agreement (binary)	77.0%
Krippendorff’s alpha (ordinal)	0.71
Average pairwise Cohen’s kappa	0.43

Table 3: Inter-annotator agreement metrics for the shared and double-annotated query-candidate pairs.

the comparisons that involved these two labels. In all cases of disagreement, however, there was a within-one-grade agreement of 93.3%. When reverting to the binary setting, the exact agreement is 77.0%. Because the dataset involved working with multiple annotators, Krippendorff’s alpha is considered. When treating the four relevance labels as ordinal, Krippendorff’s alpha was 0.71. Accordingly, this indicates a fairly good degree of agreement among the nine annotators.

Turning to the pairwise annotator agreement metrics, the average pairwise Cohen’s kappa was 0.43, thus falling in the range that indicates “moderate agreement” between pairs of annotators. For all intents and purposes, the inter-annotator agreement metrics point towards the finding that the data annotation stage yielded moderately reliable relevance labels.

4.2 Performance baselines

It goes beyond the scope of this dataset paper to present a full range of experiments with the dataset’s different representations, across different retrieval and reranker models, and while using different chunking strategies. That said, it is nonetheless instructive to establish a baseline level of retrieval performance with a strong lexical retriever and a dense retriever model. To that end, BM25 is used for the lexical baseline while GTE-ModernBERT-Base (Zhang et al., 2024) serves as the dense retriever.

4.2.1 Human-derived relevance assessment

For the human-derived relevance assessments (*Table 7*), and without drawing any firm conclusions, the strongest retrieval performance was obtained by pairing the “summary-full” query-candidate representation for BM25 ($P@10 = 0.580$; $nDCG@10 = 0.665$) and “full-full” for GTE-ModernBERT ($P@10 = 0.525$; $nDCG@10 = 0.613$).⁷ For both methods, us-

⁷Note that not too much weight should be given to BM25 outscoring GTE-ModernBERT since BM25 is one of the retrieval methods used during the seeding step, whereas GTE-

Set	Total documents	Total queries	Candidate cases per query	Avg. positive cases per query
Human annotation (graded, ≥ 2)	20,697	100	20,696	31.50
Citations	20,697	100	20,696	7.38

Table 4: Comparison of the human-annotated and citation-based evaluation sets. The human-annotated set considers relevance labels of 2 or higher as positive.

ing the full representations of the candidate cases consistently yielded the best scores, regardless of which “view” was used on the query side. Conversely, using only the facts section as either the query or candidate representation also consistently produced the lowest retrieval results. What is interesting for the held-out method is that the “summary-full” combination is quite competitive and achieves scores that sit slightly below those of “full-full”, despite being considerably shorter and thus quicker to compute. Overall, while these experiments only establish a rudimentary baseline, they do seem to illustrate that different document representations have a different influence on the retrieval results.

4.2.2 Citation-derived relevance assessment

As for the citation-derived relevance assessments (Table 8), the hypothesis is that cases cited in a query case might serve as a proxy for case-to-case relevance. To examine how closely this proxy aligns with the human relevance judgments, the citation-derived qrels were compared with the binary human-derived relevance labels (Table 5). Of the 738 query-candidate pairs that are identified as “relevant” on the basis of citations, only 308 of those were also deemed relevant by the human annotators. When using the human relevance judgments as the reference point, this corresponds to a citation precision of 41.7%. Conversely, of the 3,150 query-candidate pairs deemed “relevant” by the human annotators, only 308 were also present in the citation-derived qrels, which points to a citation recall of 9.8%. In other words, this means that the remaining 2,842 relevant pairs (about 90.2%) would be overlooked if citation links were used as the only source of relevance.

The retrieval experiments also seem to indicate that citations are, at best, an imperfect proxy. When citation-derived relevance is used for the evaluation, the highest $P@10$ scores are 0.165 for BM25 and 0.113 for GTE-ModernBERT. Taken together, these two findings (tentatively) suggest that citations are a subpar predictor of relevance within the body of CJEU case law. This aligns with other scholarly

ModernBERT is a held-out retrieval model.

	Human relevant	Human not relevant
Citation relevant	308	430
Citation not relevant	2,842	8,785

Table 5: Citation-derived versus binary human-derived relevance judgments within the annotated query-candidate pool. Human relevance treats relevance labels of 2 or higher as positive.

work, which points out that citations are often included to introduce a general principle of law or to quote a procedural rule, even if they do not have a clear bearing on the actual case (Tinarrage et al., 2025).

5 Discussion: features and use cases of the dataset

The EUR-LexRD is designed for section-aware long-form similar case retrieval and its core function is to answer the following question: given a particular query case, what is the most similar candidate case? In that respect, there are several features that set this dataset apart from existing ones.

First, the ability to test different representations of the query and candidate cases is worth pointing out. For example, queries can be encoded by means of their summary, their factual section, or the judgment’s full text. The same can also be done on the candidate side. As such, this means that it is possible to conduct experiments to determine whether particular segment types convey a stronger retrieval signal. By extension, due consideration should also be given to the length of the segments as this can inform the design of different chunking strategies.

Second, three different types of qrels are released: (i) binary citation relevance, (ii) binary human relevance, and (iii) graded human relevance. This means that different evaluation approaches are possible. Binary relevance can be used to establish a baseline level of retrieval accuracy (i.e., “does the model return any relevant cases?”), whereas accounting for the degree of relevance poses a more pronounced semantic challenge. While evaluation

based on citation relevance is probably less pertinent for CJEU case law, it can nonetheless shed interesting light on the data.

Third, with full judgments going well above 30,000 characters on average, one of the challenges relates to handling such lengthy documents. More specifically, during retrieval, it will not always be possible to fit a judgment's full text into the context window of a contemporary embedding model. Thus, different chunking and truncation strategies will need to be considered. During reranking, there will also be a significant computational hurdle when working with cases' full-text representations.

Ultimately, the utility of this dataset lies in its capacity to serve as an evaluation benchmark for diverse retrieval strategies. It can be used to compare methods across the full spectrum of information retrieval: from lexical approaches, to dense vector similarity search, and hybrid retrieval strategies. Beyond serving as a benchmark, the EUR-LexRD also offers pathways for model development and fine-tuning. One strategy here is to use the relevance labels to (post-)train a specialized model for case-to-case retrieval or reranking purposes.

6 Conclusion

The EUR-LexRD is made available under a CC BY 4.0 license and, thus, permits maximum freedom to the public to work with the data under an attribution requirement. In closing, it is hoped that the EUR-LexRD will serve as a useful resource to advance research on long-form similar case retrieval. By pinpointing some of the challenges inherent in handling complex and lengthy case law documents, the dataset is meant to pose an evaluation challenge intended to encourage the legal information retrieval community to develop more performant similar case retrieval systems.

Limitations

When working with the EUR-LexRD, several limitations should be kept in mind. First of all, the benchmark is limited to only 100 query cases, which is only a small sample compared to the full case law corpus. Furthermore, even though these 100 query cases were randomly selected from almost all English CJEU judgments, there is no detailed analysis available on how they are distributed across different subdomains of EU law. This could be relevant because retrieval performance might differ across subject areas.

During the data pre-processing step, prompt-based methods were applied with a closed-source LLM, which introduces reproducibility concerns. In addition, the quality of the LLM-generated case summaries was not validated against the original judgments, although such a validation would have provided an additional layer of quality assurance. The segmentation audit was likewise limited to 100 cases.

As concerns the data annotation stage, human relevance labels are only provided for a set of seeded candidate cases, since it is not feasible to annotate every possible query-candidate pair. In this respect, three reservations apply: (i) candidate pooling likely results in a bias towards those systems that contributed to the pool, (ii) notwithstanding acceptable inter-annotator agreement metrics, most query-candidate pairs were judged by a single annotator, and (iii) ties among annotators were resolved by the author alone, which means that these labels ultimately depend on a single adjudicator.

A further remark can be made about the fact that the corpus is exclusively composed of English CJEU judgments, which implies that evaluation results will likely not generalize to other languages or jurisdictions.

Acknowledgments

The author is very grateful to the KU Leuven Master of Laws (LL.M.) students Chaimae Harim, Chloë De Somviele, Emma Rens, Lies Boons, Marie Verstraete, Markus Hermans, Matthias De Broeck, Sylvan Nelissen, and Victor Reynaert for their efforts in annotating more than 12,000 query-candidate pairs. This work would also not have been possible without the support of the Research Foundation Flanders (FWO) [grant number 1258525N].

References

- Chris Buckley, Darrin Dimmick, Ian Soboroff, and Ellen Voorhees. 2006. [Bias and the limits of pooling](#). In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pages 619–620, New York, NY, USA. Association for Computing Machinery.
- China AI & Law Challenge. n.d. China AI & Law Challenge. <https://github.com/china-ai-law-challenge>. GitHub organization page, accessed 21 July 2026.

- Minhan Cho, Soyoung Park, S. Shyam Sundar, Daejin Choi, and Jinyoung Han. 2026. [KCC: Korean Civil Case Dataset for Legal Information Retrieval](#). In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '26, pages 3117–3123, New York, NY, USA. Association for Computing Machinery.
- Jingpei Dan, Dongbo Yang, and Yuming Wang. 2026. [Enhancing Chinese similar case retrieval based on key-fact and multi-legal-element joint augmentation](#). *Artificial Intelligence and Law*, pages 1–31.
- Yi Feng, Chuanyi Li, and Vincent Ng. 2024. [Legal Case Retrieval: A Survey of the State of the Art](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6472–6485, Bangkok, Thailand. Association for Computational Linguistics.
- Randy Goebel, Yoshinobu Kano, Mi-Young Kim, Calum Kwan, Ken Satoh, Hiroaki Yamada, and Masaharu Yoshioka. 2026. [An Overview of the COLIEE 2025 Competition: Legal Case Law and Statute Law Information Retrieval and Entailment](#). In *Proceedings of the Twentieth International Conference on Artificial Intelligence and Law*, ICAIL '25, pages 506–515, New York, NY, USA. Association for Computing Machinery.
- Randy Goebel, Yoshinobu Kano, Mi-Young Kim, Juliano Rabelo, Ken Satoh, and Masaharu Yoshioka. 2023. [Summary of the Competition on Legal Information, Extraction/Entailment \(COLIEE\) 2023](#). In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, ICAIL '23, pages 472–480, New York, NY, USA. Association for Computing Machinery.
- Randy Goebel, Yoshinobu Kano, Mi-Young Kim, Juliano Rabelo, Ken Satoh, and Masaharu Yoshioka. 2024. [Overview of Benchmark Datasets and Methods for the Legal Information Extraction/Entailment Competition \(COLIEE\) 2024](#). In *New Frontiers in Artificial Intelligence: JSAI International Symposium on Artificial Intelligence, JSAI-isAI 2024, Hamamatsu, Japan, May 28–29, 2024, Proceedings*, pages 109–124, Berlin, Heidelberg. Springer-Verlag.
- Abe Bohan Hou, Orion Weller, Guanghui Qin, Eugene Yang, Dawn Lawrie, Nils Holzenberger, Andrew Blair-Stanek, and Benjamin Van Durme. 2025. [CLERC: A Dataset for U. S. Legal Case Retrieval and Retrieval-Augmented Analysis Generation](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7913–7928, Albuquerque, New Mexico. Association for Computational Linguistics.
- Chaeun Kim, Jinu Lee, and Wonseok Hwang. 2025. [LegalSearchLM: Rethinking Legal Case Retrieval as Legal Elements Generation](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4521–4554, Suzhou, China. Association for Computational Linguistics.
- Jinhyuk Lee, Feiyang Chen, Sahil Dua, Daniel Cer, Madhuri Shanbhogue, Iftekhhar Naim, Gustavo Hernández Ábrego, Zhe Li, Kaifeng Chen, Henrique Schechter Vera, Xiaoqi Ren, Shanfeng Zhang, Daniel Salz, Michael Boratko, Jay Han, Blair Chen, Shuo Huang, Vikram Rao, Paul Suganthan, and 28 others. 2025. [Gemini Embedding: Generalizable Embeddings from Gemini](#). *arXiv preprint*. ArXiv:2503.07891 [cs.CL].
- Haitao Li, Qingyao Ai, Jia Chen, Qian Dong, Yueyue Wu, Yiqun Liu, Chong Chen, and Qi Tian. 2023a. [Sailer: Structure-aware pre-trained language model for legal case retrieval](#). In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 1035–1044, New York, NY, USA. Association for Computing Machinery.
- Haitao Li, Yunqiu Shao, Yueyue Wu, Qingyao Ai, Yixiao Ma, and Yiqun Liu. 2024. [LeCaRDv2: A Large-Scale Chinese Legal Case Retrieval Dataset](#). In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, pages 2251–2260, New York, United States.
- Qingquan Li, Yiran Hu, Feng Yao, Chaojun Xiao, Zhiyuan Liu, Maosong Sun, and Weixing Shen. 2023b. [MUSER: A Multi-View Similar Case Retrieval Dataset](#). In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 5336–5340. ArXiv:2310.15602 [cs.CL].
- Yixiao Ma, Yunqiu Shao, Yueyue Wu, Yiqun Liu, Ruizhe Zhang, Min Zhang, and Shaoping Ma. 2021. [LeCaRD: A Legal Case Retrieval Dataset for Chinese Law System](#). In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, pages 2342–2348, New York, NY, USA. Association for Computing Machinery.
- Robert Mahari, Dominik Stammach, Elliott Ash, and Alex Pentland. 2024. [LePaRD: A Large-Scale Dataset of Judicial Citations to Precedent](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9863–9877, Bangkok, Thailand. Association for Computational Linguistics.
- Juliano Rabelo, Randy Goebel, Mi-Young Kim, Yoshinobu Kano, Masaharu Yoshioka, and Ken Satoh. 2022. [Overview and Discussion of the Competition on Legal Information Extraction/Entailment \(COLIEE\) 2021](#). *The Review of Socionetwork Strategies*, 16(1):111–133.
- Juliano Rabelo, Mi-Young Kim, Randy Goebel, Masaharu Yoshioka, Yoshinobu Kano, and Ken Satoh. 2021. [COLIEE 2020: Methods for Legal Document Retrieval and Entailment](#). In Naoaki Okazaki, Katsutoshi Yada, Ken Satoh, and Koji Mineshima,

- editors, *New Frontiers in Artificial Intelligence*, volume 12758, pages 196–210. Springer International Publishing, Cham. Series Title: Lecture Notes in Computer Science.
- Stephen Robertson and Hugo Zaragoza. 2009. [The Probabilistic Relevance Framework: BM25 and Beyond](#). *Foundations and Trends in Information Retrieval*, 3(4):333–389.
- Gerard Salton and Christopher Buckley. 1988. [Term-weighting approaches in automatic text retrieval](#). *Information Processing & Management*, 24(5):513–523.
- Carlo Sansone and Giancarlo Sperli. 2022. [Legal Information Retrieval systems: State-of-the-art and open issues](#). *Information Systems*, 106:1–14.
- Tokala Yaswanth Sri Sai Santosh, Rashid Haddad, and Matthias Grabmair. 2024. [ECTHR-PCR: A Dataset for Precedent Understanding and Prior Case Retrieval in the European Court of Human Rights](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5473–5483, Torino, Italia. ELRA and ICCL.
- Dong Shu, Haoran Zhao, Xukun Liu, David Demeter, Mengnan Du, and Yongfeng Zhang. 2024. [Lawllm: Law large language model for the us legal system](#). In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, CIKM ’24*, page 4882–4889, New York, NY, USA. Association for Computing Machinery.
- Raphaël Tinarrage, Henrique Ennes, Lucas Resck, Lucas T Gomes, Jean R Ponciano, and Jorge Poco. 2025. [Empirical analysis of binding precedent efficiency in brazilian supreme court via case classification: R. tinarrage et al.](#) *Artificial Intelligence and Law*, pages 1–67.
- Marc van Opijnen and Cristiana Santos. 2017. [On the concept of relevance in legal information retrieval](#). *Artificial Intelligence and Law*, 25(1):65–87.
- Zhengying Wang and Jiong Yu. 2026. [Alter: a lightweight topic-aware representation legal case retrieval system](#). *Artificial Intelligence and Law*, pages 1–30.
- Robert Wilson. 1962. Computer Retrieval of Case Law. *SMU Law Review*, 16(3).
- Marco Wrzalik and Dirk Krechel. 2021. [GerDaLIR: A German Dataset for Legal Information Retrieval](#). In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 123–128, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Tianyang Zhang, Xianpei Han, Zhen Hu, Heng Wang, and Jianfeng Xu. 2019. [CAIL2019-SCM: A Dataset of Similar Case Matching in Legal Domain](#). *arXiv preprint. ArXiv:1911.08962 [cs.CL]*.
- Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, Meishan Zhang, Wenjie Li, and Min Zhang. 2024. [mGTE: Generalized long-context text representation and reranking models for multilingual text retrieval](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1393–1412, Miami, Florida, US. Association for Computational Linguistics.
- Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. [Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models](#). *arXiv preprint. ArXiv:2506.05176 [cs.CL]*.

Appendix

A Comparison of select similar case retrieval datasets

Year	Dataset	Task format	Jurisdiction (language)	Evaluation queries	Retrieval corpus / candidate universe	Judged candidates per query	Relevance signal
2017	FIRE 2017 IRLed	Case-to-prior-case retrieval	India (English)	200 cases	2,000 prior cases	2,000 candidates (shared across queries)	Binary citation-based relevance; 1,000 cited positives across 200 queries; 5 relevant/query
2019	CAIL2019-SCM	Triplet ranking	China (Mandarin)	8,964 triplets	3 cases per triplet	2 alternative candidates per anchor case	Pairwise preference with ground-truth label; no corpus ranking
2021	COLIEE 2020	Case-to-case retrieval	Canada (English)	650 total (520 train)	Separate pool of 200 cases for each query	200 candidates/query	Binary relevance; 5.15 relevant/query in the training set
2021	GerDaLIR	Passage-to-case retrieval	Germany (German)	122,975 total (98,380 train)	131,446 cases	131,446 candidates (shared across queries)	At least 1 cited or relevant case per passage query
2021	LeCaRD	Case-to-case retrieval	China (Mandarin)	107	43,823 cases	100 pooled and human-judged candidates/query	Four-level human relevance; 10.33 relevant/query
2023	MUSER	Multi-view case-to-case retrieval	China (Mandarin)	100	4,024 cases	100 pooled and human-judged candidates/query	Multi-view human relevance; 10.38 relevant/query
2024	LeCaRDv2	Case-to-case retrieval	China (Mandarin)	800	55,192 cases	100 pooled candidates; 30 human-judged candidates/query	Human-graded relevance; 20.89 relevant/query
2024	ECtHR-PCR	Facts-to-prior-case retrieval	Council of Europe (English)	15,204 total (9,787 train)	5,283 eligible cases/query on average in the training set	Citation-linked positives with sampled negatives	Binary citation-based relevance; 9.61 relevant/query in the training set
2026	EUR-LexRD (this work)	Long-form and multi-view case-to-case retrieval	European Union (English)	100	20,697 cases	123.65 pooled and human-judged candidates/query	Four-level human relevance (31.50 grade ≥ 2 relevant/query); citation relevance (7.38 relevant/query)

Table 6: An overview of select legal case retrieval datasets. “Retrieval corpus / candidate universe” refers to the collection of documents that is searched while evaluating, whereas “Judged candidates per query” refers to the subset of documents for which relevance labels are available. For datasets that feature a train/dev/test split, totals are reported.

B Data pre-processing input prompt

The following text was used as the input prompt supplied to Google's gemini-2.5-flash-lite during the data pre-processing stage. The input prompt was iteratively refined by the author as well as the LLM itself. The data to be processed were extracted from a Pandas DataFrame and then inserted into the prompt template by means of Python f-strings.

```
1 # CJEU judgment information extraction engine
2 You are a specialist legal information extraction engine for case law.
3 Your task is to parse the provided raw text of a Court of Justice of the European Union (CJEU) judgment and return one valid
   JSON object that conforms exactly to the schema and rules outlined below.
4 Do not explain your steps.
5 Output JSON only.
6
7 ---
8
9 ## Global output rules
10 - Output exactly one JSON object (no markdown, no comments, no code fences)
11 - All JSON strings must be valid JSON string literals. Do not include literal control characters (newline, tab, carriage
   return, etc.) inside any string value. If the copied text contains them, escape them using JSON escapes (\n, \t, \r)
   or \u000A etc.
12 - Do not add any keys that are not defined in the Final JSON Schema
13 - Use the following rules for missing values:
14   - For required string fields in the metadata: output the literal string unknown if information is missing/unavailable
15   - For segmentation anchors: output JSON null (not the string null) only when that entire segment is genuinely not present
16 - Never invent text. All anchors must be copied verbatim from the raw text.
17
18 ---
19
20 ## High-level process
21 - Step 1: Case metadata extraction
22 - Step 2: Summary and keyword generation
23 - Step 3: Case segmentation anchors
24 Perform these three steps in sequence and build the final JSON object according to the phase-by-phase instructions outlined
   below.
25
26 ---
27
28 ## Step 1: Case metadata extraction phase-by-phase instructions
29 Map fields as follows (copy literal values unless infer is stated):
30 - celex: copy literally from the Case metadata block line celex
31 - document_link: copy literally from the Case metadata block line document_link
32 - eclli: copy literally from the Case metadata block line eclli
33 - case_number: infer this from the Case metadata block line title (e.g., T-XXX/XX or C-XXX/XX)
34 - date_of_judgment: copy literally from the Case metadata block line date_of_judgment (must be ISO YYYY-MM-DD)
35 - document_type: copy literally from the Case metadata block line document_type
36 - court: copy literally from the Case metadata block line court
37 - case_name: infer this from the Case metadata block line title
38
39 ---
40
41 ## Step 2: Summary and keyword generation phase-by-phase instructions
42 Generate subject_matter, keywords, and summary_text based on the provided raw text:
43 - subject_matter: one concise sentence naming the central legal issue (no citations)
44 - keywords: 3-5 objects: { term: ..., description: ... }.
45   - Keywords must be meaningful concepts (e.g., state aid, public procurement, consumer protection)
46   - Do not use a bare legal citation/provision number as a keyword term
47 - summary_text: a 200-250 word summary in a neutral tone that clearly covers:
48   1. The background of the case
49   2. The key legal questions
50   3. The Courts reasoning
51   4. The outcome
52
53 ---
54
55 ## Step 3: Case segmentation anchors phase-by-phase instructions
56 Goal: identify the starting boundary of each of the following 8 segments in the raw text, in this exact order:
57 1. document_start
58 2. procedural_introduction
59 3. legal_context
60 4. facts_and_background
61 5. reasoning
62 6. decision_on_costs
63 7. operative_part
64 8. document_end
65 For each segment, output an array containing exactly one object: { start_anchor: ... }
66
67 ### Step 3.1: start_anchor rules
68 - start_anchor = the first 30 words of that segment (verbatim, in order)
69 - If the segment contains fewer than 30 words, output all words available for that segment
70 - Do not add ellipses (...) or any extra words or characters
71 - Anchors must be copied exactly from the raw text except that control characters must be represented using JSON escapes (e.
   g., newline as \n).
72 - document_start.start_anchor must begin at the first non-whitespace character of the raw text.
```

73

74 **### Step 3.2: headings**

75 - Segments often (but not always) start at a title heading line (e.g., Costs, Legal framework, Background to the dispute, Assessment by the Court, On those grounds).

76 - When a clear heading exists for a segment, the segment start should normally be set at the heading line itself (so the heading words appear at the beginning of start_anchor).

77 - Prefer explicit headings where present, but do not require headings if the document does not contain them.

78

79 **### Step 3.3: strict forward-only order**

80 - The structure must be consecutive in the raw text

81 - Segment start anchors must respect the segment order and must not go back in the text

82 - Concretely: the start of each segment must occur at a position that comes after the start of the previous segment.

83 - procedural_introduction.start_anchor must appear at or after document_start.start_anchor

84 - legal_context.start_anchor must appear at or after procedural_introduction.start_anchor

85 - facts_and_background.start_anchor must appear at or after legal_context.start_anchor

86 - reasoning.start_anchor must appear at or after facts_and_background.start_anchor

87 - decision_on_costs.start_anchor must appear at or after reasoning.start_anchor

88 - operative_part.start_anchor must appear at or after decision_on_costs.start_anchor

89 - document_end.start_anchor must appear at or after operative_part.start_anchor

90 - Your chosen start anchors must make it possible to reconstruct non-overlapping, contiguous segments by treating each segment as running up to the next segments start

91

92 **### Step 3.4: missing segments (use of null)**

93 - Not every judgment contains every segment as a standalone section

94 - If a segment is genuinely absent as a distinct section in the document, output: { start_anchor: null }

95 - Use null only when you are absolutely certain the segment does not appear as a standalone section (e.g., no relevant heading and no identifiable transition).

96 - Do not invent or approximate an anchor for a missing segment.

97 - Even if a segment is null, the remaining non-null segment starts must still follow the strict forward-only order.

98

99 **### Step 3.5: heuristics for identifying the case segments (context cues)**

100 Use these context cues to identify where each segment starts. Prefer explicit headings when present, but headings are not required.

101

102 1. document_start:

103 - Begins at the first non-whitespace character of the raw text

104 - Often contains metadata, formation of the Court, parties/representatives, and sometimes a Table of contents and/or a Summary

105 - Context cues to consider when locating the boundary:

106 - Judgment of the Court

107 - Case C-..., Case T-..., or similar

108 - Table of contents, Summary (when these appear near the top before the numbered paragraphs)

109

110 2. procedural_introduction:

111 - The procedural context of the case (how the case came before the Court / procedural history / what the request or action concerns)

112 - A strong cue is that the first numbered paragraph often begins here

113 - Context cues to consider when locating the boundary:

114 - Procedure before the Court

115 - This request for a preliminary ruling concerns...

116 - Often preceded by formulas such as: after hearing the Opinion of the Advocate General, gives the following

117 - Note: a top-of-document Summary usually belongs to document_start

118

119 3. legal_context:

120 - The relevant legal framework.

121 - Includes descriptions of EU law, national law, procedural rules, or applicable provisions that set the legal context for the analysis.

122 - Typically appears before the factual narrative is discussed in detail and before the Courts assessment

123 - Context cues to consider when locating the boundary:

124 - Legal framework

125 - Applicable law

126 - EU law

127 - National law

128 - Relevant provisions

129 - If no standalone legal framework section can be identified when scanning forward, set start_anchor to null (per the null rule).

130

131 4. facts_and_background:

132 - The factual narrative and background to the dispute.

133 - Includes the events leading to the case, the dispute in the main proceedings, and/or relevant factual findings.

134 - Context cues to consider when locating the boundary:

135 - Background to the dispute

136 - Background

137 - Facts

138 - Facts in the main proceedings

139 - The dispute in the main proceedings

140

141 5. reasoning:

142 - The Courts analysis applying the law to the facts.

143 - Includes assessment of pleas/grounds, answers to referred questions, interpretation, proportionality analysis, admissibility/substance, and the Courts reasoning leading to its conclusion.

144 - Context cues to consider when locating the boundary:

145 - Assessment by the Court

146 - Grounds

147 - Pleas in law

148 - The first/second ground of appeal

149 - Consideration of the questions referred

```

150 - Substance
151 - Admissibility
152
153 6. decision_on_costs:
154 - The part allocating costs (who bears costs, apportionment, interveners, etc.).
155 - Usually fairly short and located under a dedicated Costs heading.
156 - Context cues to consider when locating the boundary:
157 - Costs
158
159 7. operative_part:
160 - The dispositive section where the Court formally rules on the outcome.
161 - Contains the final orders (e.g., answers to questions, dismissal/upholding of an action/appeal, annulment, declarations).
162 - Context cues to consider when locating the boundary:
163 - On those grounds, the Court hereby rules:
164 - For these reasons, the Court hereby:
165 - The Court hereby declares and rules:
166 - Often formatted as numbered points/orders.
167
168 8. document_end:
169 - Any formal closing statements after the operative part
170 - Includes signatures, authenticity lines, language of the case, or closing formalities
171 - Ends at the final character of the raw text.
172 - Context cues to consider when locating the boundary:
173 - Signatures, name blocks, registrar/president lines, (Signatures)
174 - Language of the case: ...
175 - If the operative part ends at the end of the raw text (that is, no closing statement follows), set start_anchor to null
(per the null rule).
176
177 ### Step 3.6: boundary selection rule (tie-breaker)
178 - If boundaries are ambiguous:
179 - Prefer starting a segment at an explicit heading line (if present), so the heading appears in start_anchor.
180 - Otherwise, start at the first line/paragraph that clearly belongs to that segment by content.
181 - If selecting an anchor would violate the strict forward-only order, set that segment to null (only if you are certain it
is not present as a standalone section).
182
183 ---
184
185 ## Validation checklist
186 - Each non-null start_anchor contains at most 30 words (or fewer if the segment has fewer words).
187 - document_start.start_anchor begins with the first word of the raw text.
188 - start_anchor words are copied verbatim from the raw text (no edits, no extra words, no ellipses).
189 - All non-null segment starts respect the strict forward-only order (no backtracking).
190 - Output only the JSON object required by the schema; do not output anything else.
191
192 ---
193
194 ## Final JSON schema
195 <--- Blank JSON schema structure is inserted --->
196
197 ---
198
199 ## Few-shot examples (For guidance only. Do not parse this as the task input)
200
201 ### Example 1
202
203 ##### Raw text of Example 1
204 <--- Raw text is inserted --->
205
206 ##### Filled-out final JSON schema of Example 1
207 <--- JSON schema is inserted --->
208
209 ### Example 2
210
211 ##### Raw text of Example 2
212 <--- Raw text is inserted --->
213
214 ##### Filled-out final JSON Schema of Example 2
215 <--- JSON schema is inserted --->
216
217 ---
218 ---
219
220 ## Actual task input (parse ONLY the text below)
221 <--- Case metadata is inserted --->
222 <--- Raw text is inserted --->
223
224 ---

```

C JSON-formatted output

Below is an example of a correctly structured JSON schema for a single case. The case shown is Case C-750/21 P, *Pilatus Bank plc v European Central Bank* [2024] ECLI:EU:C:2024:124. As explained under

Subsection 3.2, the JSON-formatted output is subsequently used to reconstruct each case in its entirety by means of the positional index of the identified starting anchor of each section.

```

1 {
2   case_metadata: {
3     celex: 62021CJ0750,
4     document_link: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=cellar:8677f8ed-c665-11ee-95d9-01aa75ed71a1,
5     ecli: ECLI:EU:C:2024:124,
6     case_number: C-750/21 P,
7     date_of_judgment: 2024-02-08,
8     document_type: Judgment,
9     court: Court of Justice,
10    case_name: Pilatus Bank plc v European Central Bank
11  },
12  case_summary: {
13    subject_matter: Appeal concerning the ECBs prudential supervision competence over a credit institution after its
14    authorisation has been withdrawn and the lawfulness of legal representation for such an institution.,
15    keywords: [
16      {
17        term: Prudential supervision,
18        description: The monitoring of credit institutions by the ECB and national authorities to ensure financial stability
19        and soundness.
20      },
21      {
22        term: Legal representation,
23        description: The requirement for legal persons to be represented by a lawyer and the necessity of a valid authority
24        to act.
25      },
26      {
27        term: Withdrawal of authorisation,
28        description: The formal removal of a credit institutions license to operate, affecting the ECBs ongoing supervisory
29        competence.
30      },
31      {
32        term: Effective judicial protection,
33        description: The right of a party to have access to a court and be represented, particularly in cases of potential
34        conflicts of interest.
35      }
36    ],
37    summary_text: Pilatus Bank plc appealed an order of the General Court which dismissed its action for annulment against
38    an ECB email stating the ECB was no longer competent for its supervision following the withdrawal of its banking
39    license. The General Court had dismissed the action as manifestly lacking foundation. On appeal, the Court of Justice
40    raised a public policy issue regarding the banks legal representation. It noted that the Maltese Financial Services
41    Authority had appointed a competent person to assume all powers of the bank, including legal representation, to the
42    exclusion of the Board of Directors. The Court found that the General Court erred in law by failing to verify of its
43    own motion whether the lawyers authority to act, granted by the Board of Directors rather than the competent person,
44    was lawful. Setting aside the order, the Court of Justice gave final judgment on the admissibility. It determined that
45    under Maltese law, the Board of Directors no longer had the power to represent the bank or authorize a lawyer once the
46    competent person was appointed. Unlike the Trasta Komercbanka case, no conflict of interest was proven to exist with
47    the competent person that would necessitate allowing the Board to act to ensure effective judicial protection.
48    Consequently, the original action was declared inadmissible because the lawyer lacked a valid authority to act from a
49    body empowered to represent the bank.
50  },
51  case_segments: {
52    document_start: [
53      {
54        start_anchor: 62021CJ0750\n\n \nJUDGMENT OF THE COURT (Second Chamber)\n\n8 February 2024 (\n\n*1\n\n)\n\n(Appeal
55        Economic and monetary policy Prudential supervision of credit institutions Tasks of the European Central Bank (ECB)
56        Regulation
57      }
58    ],
59    procedural_introduction: [
60      {
61        start_anchor: Judgment\n\n1\n\nBy its appeal, Pilatus Bank plc asks the Court of Justice to set aside the order of
62        the General Court of the European Union of 24 September 2021,
63      }
64    ],
65    legal_context: [
66      {
67        start_anchor: Legal context\n\nRegulation (EU) No 1024/2013\n\n2\n\nArticle 1 of Council Regulation (EU) No
68        1024/2013 of 15 October 2013 conferring specific tasks on the European Central Bank concerning policies relating to
69      }
70    ],
71    facts_and_background: [
72      {
73        start_anchor: Background to the dispute\n\n7\n\nPilatus Bank is a credit institution established in Malta. As a less
74        significant credit institution, within the meaning of Article 6(4) of Regulation No 1024/2013,
75      }
76    ],
77    reasoning: [
78      {
79        start_anchor: The appeal\n\n32\n\nIt should be noted at the outset that, according to settled case-law, any fact
80        which relates to the admissibility of the action for annulment brought before the
81      }
82    ],
83    decision_on_costs: [

```

```

61     {
62         start_anchor: Costs\n72\n\nUnder Article 184(2) of the Rules of Procedure of the Court, where an appeal is well
63     }
64 ],
65 operative_part: [
66     {
67         start_anchor: On those grounds, the Court (Second Chamber) hereby:\n\n \n1.\n\nSets aside the order of the General
68     }
69 ],
70 document_end: [
71     {
72         start_anchor: Prechal\n\nBiltgen\n\nWahl\n\nPasser\n\nArastey Sahn\n\nDelivered in open court in Luxembourg on 8
73         February 2024.\n\nA. Calot Escobar\n\nRegistrar\n\nA. Prechal\n\nPresident of the Chamber\n\n(\n\n*\n\n)\tLanguage of
74         the case: English.
75     }
76 ]

```

D Human-derived relevance assessment

Method	Query view	Candidate view	P @10	R @10	MAP @10	nDCG @10	Judged @10
BM25	summary	summary	0.550	0.204	0.473	0.624	0.778
		facts	0.444	0.155	0.362	0.535	0.721
		reasoning	0.549	0.199	0.466	0.637	0.874
		full	0.580	0.215	0.498	0.665	0.915
	facts	summary	0.419	0.149	0.336	0.495	0.623
		facts	0.398	0.132	0.328	0.497	0.725
		reasoning	0.465	0.160	0.370	0.550	0.855
		full	0.494	0.174	0.406	0.586	0.903
	full	summary	0.494	0.181	0.409	0.563	0.713
		facts	0.271	0.087	0.208	0.350	0.524
		reasoning	0.475	0.156	0.399	0.566	<u>0.965</u>
		full	<u>0.561</u>	<u>0.207</u>	<u>0.492</u>	<u>0.655</u>	1.000
GTE-ModernBERT	summary	summary	<u>0.493</u>	<u>0.186</u>	<u>0.413</u>	<u>0.574</u>	0.729
		facts	0.340	0.126	0.254	<u>0.433</u>	0.600
		reasoning	0.411	0.153	0.337	0.505	0.666
		full	0.446	0.167	0.366	0.539	0.699
	facts	summary	0.395	0.142	0.306	0.466	0.612
		facts	0.341	0.109	0.257	0.426	0.616
		reasoning	0.392	0.137	0.294	0.473	0.668
		full	0.409	0.143	0.328	0.491	0.669
	full	summary	0.479	0.181	0.389	0.564	0.730
		facts	0.390	0.144	0.297	0.481	0.686
		reasoning	0.477	0.181	0.386	0.570	<u>0.794</u>
		full	0.525	0.198	0.430	0.613	0.830

Table 7: An overview of the retrieval performance for BM25 and GTE-ModernBERT for different query and candidate views based on the relevance labels assigned by the human annotators. The best result within each method block is typeset in bold and the second-best result is underlined. The following abbreviations are used: P (precision), R (recall), MAP (mean average precision), and nDCG (normalized discounted cumulative gain). Judged@ k refers to the proportion of the top- k retrieved documents that underwent a human relevance judgment.

E Citation-derived relevance assessment

Method	Query view	Candidate view	P @10	R @10	MAP @10	nDCG @10	Judged @10
BM25	summary	summary	0.133	0.266	0.164	0.258	0.133
		facts	0.104	0.199	0.111	0.188	0.104
		reasoning	0.121	0.243	0.147	0.233	0.121
		full	0.126	0.245	0.152	0.237	0.126
	facts	summary	0.094	0.180	0.111	0.181	0.094
		facts	0.092	0.181	0.102	0.172	0.092
		reasoning	0.098	0.192	0.105	0.181	0.098
		full	0.101	0.200	0.110	0.187	0.101
	full	summary	0.165	0.315	0.203	0.310	0.165
		facts	0.066	0.141	0.072	0.125	0.066
		reasoning	0.107	0.216	0.129	0.212	0.107
		full	<u>0.135</u>	<u>0.273</u>	<u>0.166</u>	<u>0.261</u>	<u>0.135</u>
GTE-ModernBERT	summary	summary	<u>0.109</u>	0.239	0.139	0.219	<u>0.109</u>
		facts	0.070	0.149	0.087	0.148	0.070
		reasoning	0.090	0.192	0.105	0.178	0.090
		full	0.100	0.203	0.126	0.201	0.100
	facts	summary	0.071	0.150	0.077	0.135	0.071
		facts	0.066	0.137	0.076	0.131	0.066
		reasoning	0.076	0.151	0.076	0.138	0.076
		full	0.075	0.142	0.086	0.147	0.075
	full	summary	0.091	0.177	0.114	0.183	0.091
		facts	0.082	0.157	0.101	0.167	0.082
		reasoning	0.101	0.203	0.123	0.199	0.101
		full	0.113	<u>0.222</u>	<u>0.136</u>	<u>0.219</u>	0.113

Table 8: An overview of the retrieval performance for BM25 and GTE-ModernBERT for different query and candidate views based on the citation-derived relevance labels. The best result within each method block is typeset in bold and the second-best result is underlined. The following abbreviations are used: P (precision), R (recall), MAP (mean average precision), and nDCG (normalized discounted cumulative gain). Judged@ k refers to the proportion of the top- k retrieved documents that underwent a human relevance judgment.

F Costs

In the spirit of transparency, the overview in *Table 9* details the costs associated with creating the EUR-Lex Retrieval Dataset as well as the hardware used (where this could be identified). The total cost was 5,382.36 EUR, most of which was spent on paying the data annotators.

Task	Cost type	Model	Cost (EUR)
Case segmentation	API endpoint	gemini-2.5-flash-lite	111.86
Candidate seeding	GPU rental (Nvidia RTX 4090)	all-MiniLM-L6-v2	3.00
Candidate seeding	GPU rental (Nvidia RTX 6000 Pro)	Qwen3-Embedding-8B	47.56
Candidate seeding	API endpoint	gemini-embedding-001	56.10
Data annotation	Labor costs (198 hours)	n/a	5,163.84

Table 9: Costs incurred while creating the EUR-LexRD. Candidate seeding costs are reported separately for the different embedding models that were used.