

# Neural Reranking for UK Statutory Retrieval: Provision-Level Evaluation and an Open Distilled Model

Amal Saad Alshehri<sup>1,3</sup>, Can Eken<sup>2</sup>, Nelly Bencomo<sup>1</sup>,  
Amir Atapour-Abarghouei<sup>1\*</sup>

<sup>1</sup>Department of Computer Science, Durham University, Durham, UK.

<sup>2</sup>Durham Law School, Durham University, Durham, UK.

<sup>3</sup>Department of Computer Science, Jazan University, Jazan, Saudi Arabia.

\*Corresponding author(s). E-mail(s):

[amir.atapour-abarghouei@durham.ac.uk](mailto:amir.atapour-abarghouei@durham.ac.uk);

Contributing authors: [ashahri@jazanu.edu.sa](mailto:ashahri@jazanu.edu.sa); [can.eken@durham.ac.uk](mailto:can.eken@durham.ac.uk);

[nelly.bencomo@durham.ac.uk](mailto:nelly.bencomo@durham.ac.uk);

## Abstract

This work explores provision-level retrieval and neural reranking for UK primary and secondary legislation. We introduce UK-STATUTECORPUS, a corpus of recent UK Acts and statutory instruments from *legislation.gov.uk*, together with a 100-query evaluation set of practitioner-style questions whose graded relevance judgements distinguish legally operative, supporting and contextual provisions. Using BM25 and an MPNet-based dense retriever to build candidate sets, we evaluate ten neural rerankers, including transformer cross-encoders, a late-interaction reranker, an LLM-based listwise reranker and proprietary APIs. Across both sparse and dense pools, neural reranking consistently improves normalized Discounted Cumulative Gain (nDCG) and Mean Reciprocal Rank (MRR) over first-stage retrieval. We further distil a proprietary Voyage reranker into a ModernBERT-based cross-encoder, Distilled-Voyage-ModernBERT, which approaches the teacher’s effectiveness and outperforms other open rerankers on our benchmark. Results are based on 100 expert-validated queries, each linked

to three graded provisions from a single UK instrument, so they characterise single-instrument, provision-level retrieval over recent UK legislation.

**Keywords:** Legal information retrieval, statutory retrieval, neural reranking, knowledge distillation, UK legislation

## 1 Introduction

Legislation is the primary vehicle through which modern legal systems create rights and impose duties. For many legal AI applications, the core retrieval task is therefore not simply to ‘find something about a topic’, but to locate the specific statutory provisions that govern the issue. Missing these provisions can render an answer doctrinally invalid, and empirical studies on statute-law retrieval (Nguyen et al. 2022) and retrieval-augmented legal question answering over Korean statutes and precedents (Kim et al. 2024) show that even specialised transformer-based systems and RAG-enhanced LLMs still struggle to reliably retrieve the relevant statutory provisions.

Statutory texts make this task particularly demanding. In contrast to judicial opinions, which organise narrative reasoning around issues and holdings, and to commercial contracts, which tend to follow regular clause and document templates, statutes encode legal meaning in modular, cross-referenced provisions that are amended and layered over time, contributing to a complex and difficult statute book (Office of the Parliamentary Counsel 2013). At the same time, elements such as titles, headings, preambles and general statements of purpose are used to organise and explain legislation but, as drafting guidance notes, they do not override the operative provisions that actually create rights, duties or remedies (Congressional Research Service 2020). From the perspective of a retrieval system, the passages with the strongest lexical or semantic overlap to a natural-language query are therefore not always those that answer it in a legal sense.

Standard retrieval models struggle with this mismatch. Sparse lexical retrievers such as BM25 (Robertson et al. 2009) and dense bi-encoder models such as MPNet (Song et al. 2020) reward term overlap or global semantic similarity, so on statutory corpora they often rank long expository or contextual sections above the short legally operative subsections (e.g., defining eligibility, specifying offences or conferring powers). In our experiments on UK primary and secondary legislation, BM25 and MPNet typically retrieve the operative provision somewhere in the candidate set but often below the top ranks, which is a material doctrinal error. Returning contextual discussion instead of the provision that changes the law yields a different doctrinal answer, which is not a harmless ranking discrepancy.

Despite the centrality of legislation to legal practice, most existing datasets and evaluation benchmarks in legal information retrieval focus on case law or contracts. In

the contract domain, datasets of commercial contracts and merger agreements such as the CUAD dataset introduced by Hendrycks et al. (2021) and the MAUD dataset proposed by Wang et al. (2023), together with retrieval-augmented evaluation benchmarks such as LegalBench-RAG Pipitone and Alami (2024) that build on these and related corpora, typically assess systems on their ability to locate specific contractual provisions. Recent large-scale retrieval work over judicial opinions instead focuses on case law collections, for example the CLERC dataset introduced by Hou et al. (2025). A small number of recent RAG benchmarks draw on statutory corpora in other jurisdictions, including HyPA-RAG (Kalra et al. 2024), LexRAG (Zhang et al. 2025) and the retrieve-then-read pipeline of Louis et al. (2023), but there is still no provision-level evaluation resource for UK primary and secondary legislation that evaluates individually addressable statutory provisions with graded relevance labels.

In this work, we study neural reranking for *provision-level* statutory retrieval over UK primary and secondary legislation. Our evaluation is at the level of individual statutory passages, each annotated with one of three relevance grades indicating whether it directly answers the query, strongly supports the answer, or is only contextually related. We evaluate several families of rerankers: transformer cross-encoders for passage and document ranking (Nogueira et al. 2020), late-interaction models such as ColBERT (Khattab and Zaharia 2020), LLM-based listwise rerankers (Sun et al. 2023; Pradeep et al. 2023b) and BGE cross-encoder rerankers (Xiao et al. 2024). However, their behaviour on statutory text, especially in distinguishing legally operative provisions from supporting or contextual material, remains underexplored.

Taken together, these considerations motivate a systematic evaluation of neural rerankers on provision-level statutory retrieval over UK legislation and an investigation into whether a compact open cross-encoder can approximate the behaviour of a strong proprietary legal reranker.

This study is guided by the following research questions:

- **RQ1:** How effectively do modern neural rerankers rank statutory passages in provision-level retrieval on our evaluation dataset?
- **RQ2:** To what extent does neural reranking improve over sparse (BM25) and dense (MPNet) first-stage retrieval for these statutory queries?
- **RQ3:** Can the behaviour of a proprietary legal reranker (Voyage) be distilled into a smaller ModernBERT-based student model without substantial loss of performance?
- **RQ4:** How do rerankers differ in their treatment of provisions at different relevance levels—grade 3 passages that directly answer the query, grade 2 passages that supply key powers, definitions or conditions needed to apply that answer, and grade 1 passages that are only contextually related?

Our contributions, organised around these questions, are threefold:

1. **An open distilled reranker competitive with proprietary models (RQ2, RQ3, RQ4).** We distil a proprietary Voyage legal reranker into a ModernBERT

based cross-encoder using 5,221 query–passage–score triples derived from statutory text. On our evaluation dataset, the resulting Distilled-Voyage-ModernBERT achieves nDCG@10 and MRR@10 that are not statistically distinguishable from its teacher under our query-level significance tests, while matching or exceeding all other open rerankers we evaluate (Sections 4.1, 5.5).

- 2. Empirical analysis of statutory retrieval and neural reranking (RQ1, RQ2, RQ4).** Using BM25 and MPNet as first stage retrievers, we compare ten neural rerankers, including transformer cross-encoders, a late interaction model, an LLM-based listwise reranker and two commercial API-based rerankers, on provision-level retrieval over UK legislation. We evaluate on both sparse and dense candidate pools, analyse performance by relevance grade and assess query-level statistical significance of model differences (Section 5).
- 3. UK-StatuteCorpus and a graded provision-level evaluation dataset (RQ1, RQ4).** We construct UK-STATUTECORPUS, a corpus of 12,604 UK primary and secondary legislative instruments from *legislation.gov.uk* segmented into 124,796 passages with stable identifiers, and a 100 query evaluation set of practitioner style statutory questions with three way graded relevance judgements for individually addressable provisions (Sections 3.1 and 3.2).

The remainder of the paper is organised as follows. Section 2 introduces background on statutory text, neural information retrieval and reranking, knowledge distillation for retrieval, and related legal-domain corpora and RAG benchmarks. Section 3 presents the corpus and evaluation dataset. Section 4 describes the reranking models and distillation setup. Section 5 reports empirical results, and Sections 6–8 conclude with discussion, limitations and directions for future work.

## 2 Background and Related Work

This section summarises properties of statutory text that matter for retrieval, then reviews neural retrieval and reranking (Section 2.2), knowledge distillation for information retrieval (Section 2.3) and legal-domain corpora and RAG benchmarks that are most closely related to our work (Section 2.4).

### 2.1 Statutory Text and Provision-Level Relevance

Statutes differ sharply in form and function from both case law and private agreements. Legislative instruments are modular and hierarchical, organised into parts, sections, regulations, schedules and sub-paragraphs, and are repeatedly amended over time (Office of the Parliamentary Counsel 2013; Congressional Research Service 2020). Legal effect is often concentrated in narrowly drafted provisions that create or extinguish rights or duties, define terms and eligibility conditions, specify offences and penalties or confer powers on public bodies, while surrounding material may be topically related but does not itself change the normative position of an actor.

For retrieval, this structure creates a persistent misalignment between textual similarity and legal relevance. Passages with the highest lexical or semantic overlap to a

natural-language query are not necessarily those that answer it in a legal sense. For example, preambles and other explanatory material can dominate the ranking even when an operative clause appears elsewhere in the instrument. In practice, relevance is graded at the level of individual provisions. A practitioner often needs (a) the clause that directly creates or alters the legal position of an actor, for instance by imposing an obligation, conferring a power or creating a right (legally operative), (b) passages that clarify how that clause applies in practice (supporting), and (c) further provisions that situate the overall scheme (contextual). Any evaluation of statutory retrieval therefore needs to distinguish provisions that effect legal change from those that merely discuss or frame the same topic, and our benchmark adopts this three-way distinction explicitly.

## 2.2 Neural Retrieval and Reranking

Modern text retrieval typically follows a two-stage design in which a first-stage retriever constructs an index and returns a candidate set, and a second-stage reranker then scores these candidates with a more expressive model. On the retrieval side, sparse lexical methods such as BM25 (Robertson et al. 2009) remain widely used in legal search because they are robust and interpretable, while dense retrievers based on bi-encoder architectures (Song et al. 2020) embed queries and passages into a shared vector space and retrieve by nearest neighbour search, which improves recall when semantically relevant passages lack term overlap. However, both families tend to favour longer, discursive passages over short clauses when judged purely by similarity, which is problematic for statutes in which short operative provisions carry most of the legal content.

Neural rerankers address this by using richer interaction patterns between query and passage. Cross-encoder rerankers jointly encode the concatenated query–passage pair and output a relevance score (Nogueira et al. 2020). Late-interaction models such as ColBERT and ColBERTv2 (Khattab and Zaharia 2020; Santhanam et al. 2022) encode queries and passages independently but retain token-level representations, aggregating token similarities at scoring time. More recent work uses large language models (LLMs) as listwise rerankers that reorder sets of candidates (Sun et al. 2023; Pradeep et al. 2023b), and develops strong multilingual cross-encoder rerankers such as the BGE family evaluated on MTEB (Muennighoff et al. 2023; Xiao et al. 2024). These architectures are typically trained on general-domain datasets such as MS MARCO (Bajaj et al. 2016), and results on BEIR (Thakur et al. 2021) shows that models tuned on web search can degrade substantially when evaluated zero-shot on specialised domains.

In legal settings, this domain shift is visible in clause-level contract retrieval (e.g., CUAD and the contract retrieval tasks in LegalBench-RAG (Hendrycks et al. 2021; Pipitone and Alami 2024)), where improvements over strong lexical baselines can be modest or inconsistent, and in case law retrieval over large collections such as CLERC (Hou et al. 2025), where domain-specific architectures and training regimes are needed to reach high recall. Work on answer retrieval in legal community question answering further indicates that cross-encoders which incorporate structured legal signals (for

example, question tags and other forum metadata) can outperform otherwise comparable text-only rerankers (Askari et al. 2024). Taken together, these studies suggest that applying neural reranking to statutes is promising but non-trivial as models must cope with long heavily structured instruments and relevance criteria centred on legally operative provisions rather than topical similarity alone. This motivates our focus on provision-level statutory retrieval over UK legislation and a comparative evaluation of general-purpose and legal-oriented rerankers in this setting (RQ1, RQ2, RQ4; cf. Sections 3 and 5).

### 2.3 Knowledge Distillation for Neural Retrieval

Knowledge distillation transfers the behaviour of a high-capacity “teacher” model to a smaller “student” model by training the student to match the teacher’s outputs. In information retrieval, distillation has been used to compress cross-encoder teachers into more efficient dense retrievers and rerankers. Margin-MSE (Hofstätter et al. 2021a) and TAS-B (Hofstätter et al. 2021b) show that score differences from cross-encoder or late-interaction teachers can supervise bi-encoder students, improving effectiveness without prohibitive inference costs. GPL (Wang et al. 2022) combines synthetic query generation, hard-negative mining and cross-encoder pseudo-labelling to adapt dense retrievers to new domains leading to substantial gains on BEIR (Thakur et al. 2021).

More recently, this distillation paradigm has been extended to settings where the teacher is an LLM-based reranker (Pradeep et al. 2023a; Baldelli et al. 2024). RankVicuna and RankZephyr use a powerful LLM to produce listwise rankings over candidate passages and then fine-tune smaller open LLM rerankers (e.g. Vicuna- and Zephyr-based models) to imitate these listwise preferences (Pradeep et al. 2023a). TWOLAR instead employs ChatGPT as a zero-shot teacher, generates synthetic queries from the MS MARCO corpus by combining cropped sentences with docT5query-generated queries, retrieves diverse candidate sets using BM25, SPLADE, DRAGON and BM25+monoT5, and distils the LLM’s listwise judgements into Flan-T5-based cross-encoder rerankers trained with a RankNet loss, achieving strong results on TREC DL and BEIR benchmarks (Baldelli et al. 2024). Both lines of work report experiments on general-domain web corpora such as MS MARCO and BEIR; they do not report results on statutory corpora or provision-level legal relevance labels.

In this work, we instantiate this teacher–student framework for statutory retrieval. As teacher, we use the API-based reranking model that achieved the strongest performance in our preliminary experiments on UK statutory queries; the student is a ModernBERT-based cross-encoder trained on teacher-scored statutory query–passage pairs (Section 4).

### 2.4 Legal-Domain Corpora, RAG Benchmarks and Statutory Resources

Legal AI has a long history of modelling legal reasoning, from early case-based systems such as HYPO (Ashley 1991) and CATO (Ashley 2002) to transformer-based language models trained on large legal corpora. LEGAL-BERT (Chalkidis et al. 2020),

its multilingual MultiEURLEX variant (Chalkidis et al. 2021) and the LexFiles-based models (Chalkidis et al. 2023) are pretrained on mixtures of statutes, case law, EU legislation and other legal texts and are evaluated mainly on legal text classification and entailment benchmarks. This line of work demonstrates the benefits of legal-domain pretraining, but does not consider statutory retrieval and reranking.

Retrieval-augmented generation (RAG) (Lewis et al. 2020) has recently been adopted in legal AI. LegalBench-RAG (Pipitone and Alami 2024) benchmarks the retrieval component of RAG pipelines on contract review and privacy-policy question answering, while CLERC (Hou et al. 2025) and CaseGPT (Yang 2024) study retrieval-augmented reasoning over large collections of case law and other case-based corpora. Other recent work brings RAG to statutory sources: HyPA-RAG evaluates question answering over an AI-specific statute and associated rules (New York City’s Local Law 144) (Kalra et al. 2024), LexRAG benchmarks multi-turn legal consultations with retrieval of relevant legal articles and grounded response generation (Zhang et al. 2025), and Louis et al. (2024) propose and assess a “retrieve-then-read” pipeline in which a dense retriever first selects Belgian statutory articles and an LLM reader then generates long-form answers conditioned on those articles. These resources highlight the importance of high-precision retrieval for downstream generation, but focus on contracts, privacy policies, case law or non-UK statutes.

To our knowledge, there is no corpus and evaluation set for UK primary and secondary legislation that assesses retrieval at the level of individual statutory provisions with graded relevance labels. We introduce such a corpus and evaluation to enable systematic comparison of neural rerankers for provision-level statutory retrieval over UK legislation.

## 3 Methodology

### 3.1 Corpus Construction

We construct UK-STATUTECORPUS, a large-scale corpus of UK primary and secondary legislation designed specifically for provision-level statutory information retrieval. Existing legal NLP benchmarks focus primarily on case law (Hou et al. 2025) and contracts (Hendrycks et al. 2021), and, to our knowledge, no prior resource has been available that supports provision-level evaluation of retrieval and reranking over UK statutes with graded relevance labels. The corpus is derived from *legislation.gov.uk*, the official public repository for UK legislation maintained by The National Archives.<sup>1</sup> This source provides authoritative versions of UK primary and secondary legislation, including devolved legislation for Scotland, Wales and Northern Ireland. We implemented a custom extraction pipeline to obtain legislative texts at scale while preserving structural and provenance information. Extracted content is normalised and stored in a hierarchical directory structure organised by jurisdiction, legislation type, year and instrument title.

---

<sup>1</sup>Content is used in accordance with the Open Government Licence v3.0 (OGL). See <https://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/>.

### 3.1.1 Segmentation Strategy

Each legislative instrument is segmented into passages that initially correspond to individually addressable pages on *legislation.gov.uk*. These units align with the structural elements presented to readers, typically numbered sections, regulations, articles, schedules or explanatory notes, rather than with arbitrary fixed-length text windows. After subsequent preprocessing (described below), some passages may span multiple such pages when short segments are merged.

This segmentation strategy yields several methodological advantages:

1. *Legal coherence*: Each passage corresponds to a meaningful unit of legislative drafting, preserving the internal logic and cross-referencing patterns of the instrument.
2. *Provenance*: Each passage is associated with a stable canonical URL on *legislation.gov.uk*, enabling direct verification and downstream integration into legal research tools.
3. *Realistic granularity*: Passage lengths reflect the natural heterogeneity of statutory drafting, from short commencement or extent provisions to lengthy schedules.

During preprocessing, very short segments (typically isolated headings or part markers with fewer than 15 tokens) are merged with the following passage (or, if there is no following passage, with the preceding passage) to ensure that each unit contains substantive legal content suitable for retrieval and reranking.

### 3.1.2 Corpus Statistics

The final corpus contains **12,604 legislative instruments** segmented into **124,796 passages**, with a total of approximately **61.3 million tokens**. The corpus covers legislation enacted or made between 2020 and 2024, capturing the post-Brexit UK legislative landscape, including major regulatory activity in public health, trade and devolved governance.

Table 1 summarises the composition by legislation type. UK Statutory Instruments constitute the largest category (45.2% of instruments, 43.7% of passages), followed by Scottish and Welsh secondary legislation. Primary legislation (UK Public General Acts and Acts of the devolved legislatures) comprises fewer instruments but substantially more passages per instrument, which reflects the greater structural complexity of Acts relative to secondary instruments.

### 3.1.3 Passage Length Distribution

Table 2 reports passage length statistics. The median passage contains 214 tokens, with an interquartile range of 116–436 tokens. The vast majority of passages (91.8%) fall within the 50–1,000 token range typical of individual sections or regulations. Only 1.0% of passages exceed 5,000 tokens, corresponding primarily to schedules and extensive tabular material, while fewer than 0.3% contain fewer than 50 tokens, typically short commencement or extent provisions that already form complete provisions and were therefore not merged.

**Table 1:** Corpus composition by legislation type (2020–2024).

Legislation type	Instruments	Passages
UK Statutory Instruments	5,696	54,567
UK Draft Statutory Instruments	1,119	14,497
UK Public General Acts	193	13,953
Scottish Statutory Instruments	2,111	11,501
Wales Statutory Instruments	1,476	10,263
Northern Ireland Statutory Rules	1,415	9,934
Scottish Draft Statutory Instruments	380	3,113
Acts of the Scottish Parliament	68	2,868
Acts of Senedd Cymru	18	2,103
Acts of the Northern Ireland Assembly	50	1,305
Northern Ireland Draft Statutory Rules	57	346
UK Church Measures	9	240
Acts of the National Assembly for Wales	3	67
UK Local Acts	3	33
UK Ministerial Orders	5	5
UK Ministerial Directions	1	1
<b>Total</b>	<b>12,604</b>	<b>124,796</b>

**Table 2:** Passage length distribution (tokens) in UK-STATUTECORPUS.

Statistic	Tokens
1st percentile	52
25th percentile	116
Median (50th)	214
75th percentile	436
99th percentile	4,959
Mean	491
Standard deviation	1,538

This distribution closely mirrors the structure of legislation as published and therefore provides a realistic evaluation setting for retrieval systems. In contrast to corpora constructed via fixed-length windowing, UK-STATUTECORPUS preserves the variable granularity that retrieval systems must handle in production legal applications.

### 3.1.4 Instrument Complexity

Due to the diversity of legislative forms, the number of passages per instrument varies substantially. Table 3 reports the distribution of passages per instrument. The median instrument contains 5 passages, typical of a short Statutory Instrument with a title, commencement provision, substantive regulations and signature block. The mean of 9.9 passages reflects the influence of larger instruments, with the most complex Act in the corpus containing 768 passages. This variability is characteristic of real legislative corpora and distinguishes UK-STATUTECORPUS from synthetic benchmarks with artificially uniform document structure.

**Table 3:** The distribution of passages per instrument

Statistic	Passages
Minimum	1
25th percentile	3
Median (50th)	5
75th percentile	10
Maximum	768
Mean	9.9

**Table 4:** Metadata schema for each passage in UK-STATUTECORPUS.

Field	Type	Description
<code>doc_id</code>	string	Unique identifier for the parent instrument
<code>chunk_id</code>	string	Unique identifier for the passage
<code>source_url</code>	string	Canonical URL on legislation.gov.uk
<code>chunk_position</code>	int	Position within the instrument (1-indexed)
<code>n_chunks_in_doc</code>	int	Total number of passages in the parent instrument
<code>year</code>	int	Year of enactment or making
<code>LegislationType</code>	string	Legislative category (see Table 1)
<code>file_name</code>	string	Title of the instrument
<code>token_count</code>	int	Number of tokens in the passage
<code>chunk_title</code>	string	Section/regulation heading (if present)
<code>chunk_summary</code>	string	Extractive summary of passage content
<code>subjects</code>	string	Semicolon-delimited subject keywords

### 3.1.5 Data Schema

Each passage is stored as a JSON object with two top-level fields: `content` (raw legislative text) and `metadata`. The metadata schema is summarised in Table 4. All fields are populated for every passage, though some, for example, `chunk_title`, may be empty when no heading is present.

Although the metadata includes derived fields such as `chunk_title`, `chunk_summary`, and `subjects`, these are provided solely for navigation and downstream applications. They are *not* used by any retrieval or reranking model in this study. All retrieval and reranking experiments operate exclusively on the `content` text so the reported performance reflects retrieval from unstructured statutory language rather than from curated annotations.

### 3.1.6 Retrieval Representations

To study reranking under different first-stage retrieval paradigms, we construct two separate candidate pools: one sparse (BM25) and one dense (MPNet bi-encoder).

### *Sparse retrieval (BM25):*

We build a standard lexical index using BM25. Each passage is tokenised and indexed, and documents are ranked according to BM25 term frequency, inverse document frequency and document-length normalisation. BM25 serves as a strong and interpretable baseline and reflects the retrieval approach widely used in production legal search systems.

### *Dense retrieval (MPNet bi-encoder).*

For dense retrieval, we use the `all-mpnet-base-v2` checkpoint, a general-purpose sentence encoder based on the MPNet architecture. A bi-encoder architecture is employed as the query and passage are encoded independently into 768-dimensional vectors and similarity is measured by cosine similarity. Passage embeddings are indexed with FAISS (Johnson et al. 2019) to enable efficient approximate nearest neighbour search over the corpus.

For each evaluation query, we retrieve the top- $k = 100$  candidates from each pool, yielding two candidate sets (one lexical, one dense) on which all reranking models are evaluated. We focus on BM25 and MPNet as widely-used first-stage baselines. More specialised hybrid or domain-adapted retrievers fall outside the scope of this work and can be explored in future work.

The corpus, evaluation queries, graded relevance judgements and distillation dataset are released together with the extraction and preprocessing code to support reproducibility and future research on statutory retrieval.<sup>2</sup>

## 3.2 Evaluation Dataset

To evaluate retrieval and reranking performance under practitioner-style statutory information needs, we construct a 100-query evaluation set with graded relevance judgements. Each query is anchored within a single instrument to isolate intra-instrument provision ranking. Cross-instrument retrieval is out of scope and listed as future work.

### *Candidate pool:*

We constructed a controlled candidate pool of 1,463 legislative instruments from UK-STATUTECORPUS by selecting documents containing between 4 and 8 passages, with each passage strictly bounded between 500 and 1,000 tokens. This selection criterion ensures that candidates possess sufficient semantic density to represent substantive statutory provisions. This avoids the trivial retrieval scenarios inherent in very short instruments but maintains a bounded context that facilitates consistent high-fidelity relevance annotation. By anchoring each query within a single instrument, this design isolates the challenge of intra-instrument retrieval, which requires models to distinguish legally operative clauses from textually similar but non-operative material within the same legislative framework.

---

<sup>2</sup>For the purpose of double-blind review, the repository link is omitted. It will be provided at a stable archival location upon acceptance.

### *Query generation pipeline:*

Queries and relevance judgements are constructed using a controlled three-stage pipeline based on locally served language models via Ollama, a local LLM inference framework (Ollama Inc. 2025). To reduce shared failure modes and provide independent checks, each stage uses a different model:

1. **Generator** (`qwen2.5:14b-instruct`): Given document metadata and up to 8 passages from a single instrument, the generator produces one statutory question beginning with the template prefix “Under the [LegislationType], ...” and selects exactly three passages with graded relevance labels. The labels follow a three-point scale: grade 3 (direct answer), grade 2 (strongly supporting), grade 1 (tangentially related but providing useful context). For each selection, the model also produces a brief rationale ( $\leq 30$  words) justifying the assigned grade.
2. **Verifier** (`llama3.1:8b-instruct`): The verifier checks the generator output for compliance with the expected JSON schema, valid passage identifiers, correct label distribution (exactly one passage at each grade) and internal consistency between the query and selected passages. Minor structural errors (for example, inconsistent passage identifier prefixes) are automatically repaired and instances with more substantive issues are discarded.
3. **Auditor** (`mistral:7b-instruct`): The auditor performs an independent validation pass. It verifies schema adherence and assesses whether the query is answerable using only the three selected passages, without requiring external legislative or factual knowledge.

An instance is accepted into the evaluation set only if it passes all three stages. Instances failing any stage, due to schema violations, invalid passage references, incorrect grade distributions or semantic inconsistencies, are discarded and the pipeline moves on to the next candidate instrument. We repeat this process over candidate instruments until we obtain 100 validated query instances.

### *Expert review:*

To ensure legal validity, a legal-domain expert reviewed the complete dataset. For each query, the expert inspects the query text, the full content of the three selected passages, the assigned relevance grades and the model-generated rationales. The expert assesses both whether the query corresponds to a plausible statutory information need and whether the ranking of the three passages is legally defensible given the query. Only one instance (Q0002) required correction: two passages had their relative ordering adjusted by swapping their grades to better reflect their legal relationship to the query. No queries were removed at this stage and after this correction the dataset was frozen for all subsequent experiments.

### *Dataset characteristics:*

The final evaluation set consists of 100 queries over 100 unique legislative instruments, with 300 passage-level relevance judgements. By construction, each query has exactly one passage at each relevance grade (3, 2 and 1), yielding a controlled setting in which

**Table 5:** Evaluation dataset statistics.

Statistic	Value
Queries	100
Unique instruments	100
Relevance judgements	300
Grade 3 (directly answers)	100
Grade 2 (strongly supporting)	100
Grade 1 (related context)	100
Year range	2020–2024

**Table 6:** Example query with graded relevance judgements from the evaluation set.

Query (Q0001):	
“Under the Scottish Statutory Instruments, what are the consequences for contravening a direction given by a local authority?”	
Grade	Passage (truncated)
3	<i>Offences and penalties</i> — “A person who contravenes a direction under regulation 5(1), 6(1) or 7(1) commits an offence. . . An offence under this regulation is punishable on summary conviction by a fine not exceeding the statutory maximum.”
2	<i>Directions relating to individual premises</i> — “A local authority may give a direction imposing prohibitions, requirements or restrictions in relation to the entry into, departure from, or location of persons in, specified premises. . .”
1	<i>Explanatory Note</i> — “These Regulations make provision for a local authority to give directions relating to specified premises, events and public outdoor places in its area. . .”

rerankers must distinguish provisions that directly answer the query (grade 3) from strongly supporting (grade 2) and contextual (grade 1) material. The evaluation set spans multiple legislation types, including UK Statutory Instruments, Scottish Statutory Instruments, Welsh Statutory Instruments, Northern Ireland Statutory Rules and draft instruments, covering the period 2020–2024. Summary statistics are reported in Table 5 and a representative example is shown in Table 6.

## 4 Models and Distillation Methodology

In this section, we describe the reranking models evaluated in our experiments and the teacher–student distillation setup used to obtain our open, locally deployable cross-encoder reranker, Distilled-Voyage-ModernBERT. We first outline the distillation data and teacher scores, then present the student architecture and training objective and finally summarise the inference-time reranking pipeline. Quantitative results, including loss curves and correlations between teacher and student scores, are reported in Section 5.

## 4.1 Knowledge Distillation to ModernBERT

We adopt a teacher–student distillation setup to obtain an open, locally deployable cross-encoder reranker that approximates the behaviour of a proprietary teacher model on statutory text. The student model, Distilled-Voyage-ModernBERT, uses a ModernBERT encoder as its backbone and is trained to regress the teacher’s continuous relevance scores.

### 4.1.1 Synthetic Query Generation

To obtain a training set of information-seeking queries grounded in statutory text, we generate synthetic questions from sampled documents in the training portion of the corpus. For each seed instrument, short excerpts (typically one or two passages) are provided to a locally hosted language model together with instructions to formulate plausible questions that a practitioner or researcher might ask about the provisions. A light normalisation step (removing incomplete questions and deduplicating near-identical queries) yields approximately 600 well-formed queries. The language model is used solely for query generation and does not participate in scoring.

### 4.1.2 Candidate Retrieval

For each synthetic query, we retrieve an initial pool of candidate passages using the BM25 index over the training corpus described in Section 3.1. The top-ranked segments for each query are retained, providing a lexically driven high-recall set of potentially relevant passages. At this stage, no supervision is applied. The role of BM25 is solely to define a tractable candidate space that the teacher model will subsequently score.

### 4.1.3 Teacher Scoring

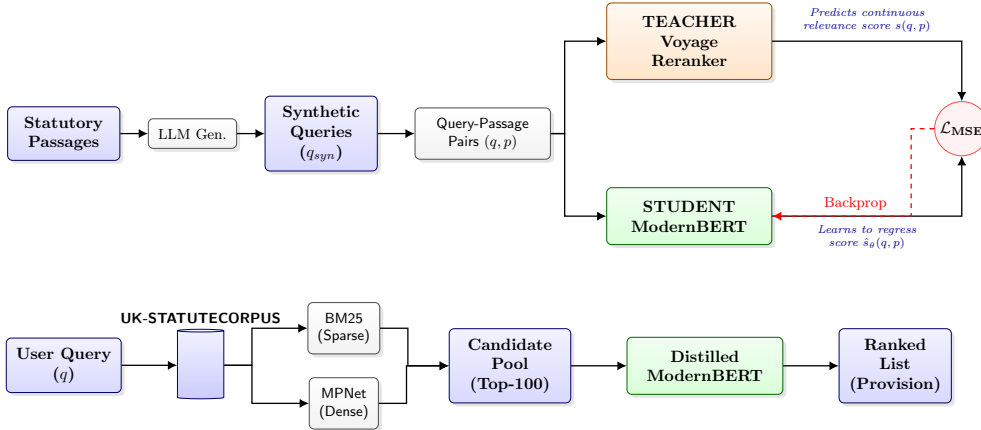
A proprietary Voyage reranker serves as the teacher model. For each query–candidate pair  $(q, p)$ , the teacher outputs a continuous relevance score

$$s(q, p) \in [0, 1],$$

interpretable as the model’s confidence that the passage addresses the query. These teacher scores are the *only* supervision signal used in distillation and no manual judgements or additional large language model scores are involved.

BM25 retrieval over the  $\approx 600$  synthetic queries initially produces a large pool of query–candidate pairs. After scoring all pairs with the teacher and removing duplicate query–passage combinations for a given query, we obtain a regression dataset of approximately 5,200 unique instances. A random sample of these query–passage–score triples is manually inspected to confirm that higher teacher scores correspond to more intuitively relevant passages. The final regression dataset is released alongside the corpus to facilitate further research on statutory reranking.

Let  $\mathcal{D} = \{(q_i, p_i, s_i)\}_{i=1}^N$  denote the scored dataset, where  $q_i$  is a query,  $p_i$  is a passage, and  $s_i = s(q_i, p_i)$  is the teacher’s relevance score, we obtain  $N = 5,221$  query–passage–score triples. The score distribution is broad (minimum  $\approx 1.4 \times 10^{-4}$ , maximum  $\approx$



**Fig. 1: Framework Architecture.** Offline distillation of a ModernBERT student from a Voyage teacher and its deployment for statutory reranking.

0.99992, mean  $\approx 0.577$ , standard deviation  $\approx 0.393$ ), providing a rich continuous supervision signal.

We split  $\mathcal{D}$  into a training subset  $\mathcal{D}_{\text{train}}$  (90%, 4,699 examples) and a validation subset  $\mathcal{D}_{\text{val}}$  (10%, 522 examples).

#### 4.1.4 Student Model Architecture and Objective

The student model is implemented as a ModernBERT-based cross-encoder reranker, initialised from the `nomic-ai/modernbert-embed-base` (Nussbaum et al. 2024). The full architecture is illustrated in Figure 1, which depicts how the student model is trained offline to mimic the teacher’s scoring behaviour and subsequently deployed to rerank statutory candidates at inference time. Given a query  $q$  and passage  $p$ , we construct a single input sequence

$$x = [\text{CLS}] q [\text{SEP}] p [\text{SEP}],$$

tokenise it, and feed it through the ModernBERT encoder. Let  $h_{\text{CLS}}(q, p) \in \mathbb{R}^d$  denote the final hidden representation at the [CLS] position. A linear regression head maps this representation to a scalar prediction

$$\hat{s}_{\theta}(q, p) = w^{\top} h_{\text{CLS}}(q, p) + b,$$

where  $\theta$  denotes all trainable parameters (encoder and head).

We train the student to approximate the teacher’s relevance scores using a pointwise mean squared error (MSE) objective. For a minibatch  $\mathcal{B} \subset \mathcal{D}_{\text{train}}$ , the loss is

$$\mathcal{L}_{\text{MSE}}(\theta; \mathcal{B}) = \frac{1}{|\mathcal{B}|} \sum_{(q_i, p_i, s_i) \in \mathcal{B}} (\hat{s}_{\theta}(q_i, p_i) - s_i)^2.$$

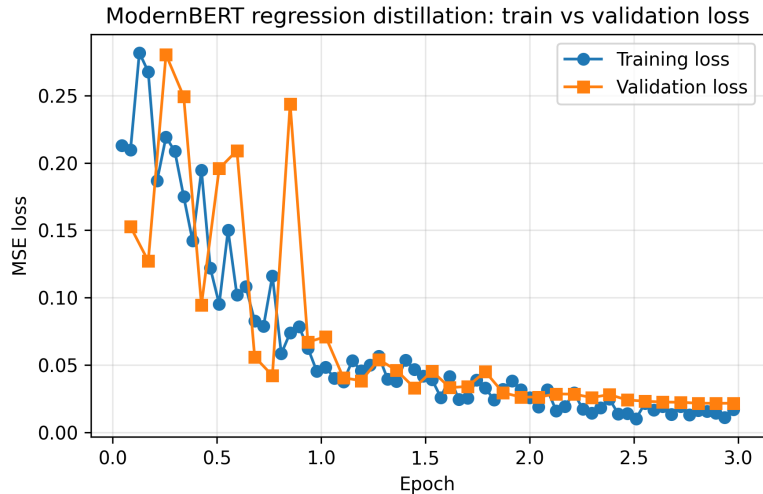
Fine-tuning is performed with a maximum input length of 2,048 tokens (joint over query and passage), batch size 4, learning rate  $2 \times 10^{-5}$ , mixed-precision training and three epochs of optimisation. The validation set  $\mathcal{D}_{\text{val}}$  is used solely for monitoring regression quality and early stopping.

#### 4.1.5 Corpus and Evaluation Split

All distillation experiments are conducted on the statutory corpus described in Section 3.1. To avoid contamination of the evaluation benchmark, we reserve a disjoint subset of instruments solely for evaluation. This held-out portion is fixed in advance and excluded from synthetic query generation, candidate retrieval, teacher scoring and student training. Evaluation on the 100-query benchmark therefore measures generalisation to unseen statutory instruments rather than memorisation of training documents.

#### 4.1.6 Training Dynamics

Figure 2 shows the evolution of both training and validation mean squared error (MSE) during knowledge distillation to ModernBERT over three epochs. Both curves decrease rapidly from an initial value around 0.25 and converge to a small value without divergence between training and validation loss, indicating stable optimisation and no apparent overfitting on the regression dataset. A more detailed quantitative characterisation of the student–teacher agreement, including correlations between the teacher scores  $\{s_i\}$  and student predictions  $\{\hat{s}_i\}$ , is reported in Section 5.



**Fig. 2:** Training and validation mean squared error (MSE) during knowledge distillation from the proprietary teacher reranker to ModernBERT over three epochs. Both curves decrease rapidly and converge to a small value without divergence, indicating stable optimisation.

### 4.1.7 Final Reranking Pipeline

At inference time, we adopt a two-stage retrieval architecture over UK-STATUTECORPUS. Given a user query:

1. A first-stage retriever (BM25 or MPNet bi-encoder) retrieves a candidate set of passages from the statutory corpus. In our experiments, we use BM25 as the default first-stage retriever, and additionally report performance when the same reranker is applied to MPNet candidate pools (Section 5).
2. The Distilled-Voyage-ModernBERT cross-encoder then scores each query-candidate pair  $(q, p)$ , producing a continuous relevance prediction  $\hat{s}_\theta(q, p)$ .

Candidates are ranked by  $\hat{s}_\theta(q, p)$ , and the highest-scoring passages are returned as the final retrieval output. The resulting reranker is fully open source, runs entirely locally without external API dependencies and is designed for legal settings that require transparency and strict data governance. In Section 5 we show empirically how closely its rankings align with those of the proprietary teacher on UK statutory text.

## 5 Results and Analysis

We evaluate ten reranking models over two retrieval settings: a dense pool based on MPNet embeddings and a sparse pool based on BM25. For each of the 100 queries in our benchmark, the underlying corpus provides exactly three relevant passages with graded relevance labels (grades 3, 2, and 1), yielding 300 expert-verified relevance annotations. For each query and each pool, we retrieve the top 100 candidate passages and then apply rerankers to this fixed candidate set.

### 5.1 Evaluation Metrics

Let  $Q$  denote the set of queries with  $|Q| = 100$ . For each query  $q \in Q$ , let  $R_q = \{p_1, \dots, p_k\}$  be the ranked list of passages returned by a reranker (restricted to the top- $k$  positions), and let  $\text{rel}(p_i) \in \{0, 1, 2, 3\}$  denote the relevance grade of passage  $p_i$ .

***Normalised Discounted Cumulative Gain (nDCG@k):***

nDCG measures ranking quality under graded relevance and position discount. For a query  $q$ , DCG at cutoff  $k$  is

$$\text{DCG}@k(q) = \sum_{i=1}^k \frac{2^{\text{rel}(p_i)} - 1}{\log_2(i + 1)}. \quad (1)$$

The corresponding ideal DCG,  $\text{IDCG}@k(q)$ , is obtained by sorting passages in descending order of their relevance grades. The normalised score is

$$\text{nDCG}@k = \frac{1}{|Q|} \sum_{q \in Q} \frac{\text{DCG}@k(q)}{\text{IDCG}@k(q)}. \quad (2)$$

Metric	Dense Pool: MPNet Retrieval										
	MPNet -only*	Mini LM-6 <sup>1</sup>	Mini LM-12 <sup>2</sup>	mono T5 <sup>3</sup>	Dist. V-MB <sup>4</sup>	Cohere r-v3.5 <sup>5</sup>	Voyage r-2.5 <sup>6</sup>	Col BERT <sup>7</sup>	Rank Zeph <sup>8</sup>	BGE base <sup>9</sup>	BGE large <sup>10</sup>
MRR@5	.3835	.4418	.4818	.4615	<b>.5577</b>	.5275	<b>.5743</b>	.3840	.3835	.5308	.5303
MRR@10	.3984	.4571	.4982	.4712	<b>.5647</b>	.5362	<b>.5859</b>	.4058	.3984	.5389	.5442
MRR@25	.4032	.4625	.5023	.4765	<b>.5679</b>	.5434	<b>.5872</b>	.4111	.4032	.5416	.5473
MRR@50	.4053	.4637	.5034	.4768	<b>.5683</b>	.5437	<b>.5872</b>	.4130	.4053	.5425	.5479
MRR@100	.4062	.4639	.5034	.4770	<b>.5683</b>	.5437	<b>.5872</b>	.4132	.4062	.5425	.5479
nDCG@5	.2695	.3013	.3391	.3059	<b>.3991</b>	.3574	<b>.4136</b>	.2723	.2695	.3423	.3431
nDCG@10	.3117	.3445	.3742	.3456	<b>.4211</b>	.3863	<b>.4500</b>	.3157	.3117	.3766	.3917
nDCG@25	.3329	.3741	.3995	.3850	<b>.4489</b>	.4276	<b>.4751</b>	.3465	.3329	.4111	.4213
nDCG@50	.3570	.3909	.4163	.3967	<b>.4627</b>	.4369	<b>.4821</b>	.3600	.3570	.4221	.4303
nDCG@100	.3684	.3992	.4220	.4007	<b>.4665</b>	.4388	<b>.4835</b>	.3748	.3684	.4253	.4326
Recall@5	.2500	.2567	.2900	.3000	<b>.3300</b>	.3267	<b>.3600</b>	.2467	.2500	.3267	.3167
Recall@10	.3467	.3700	.3967	.3867	.4000	.4067	<b>.4567</b>	.3633	.3467	.3967	<b>.4367</b>
Recall@25	.4433	.4800	.4933	.5167	.5133	.5367	<b>.5600</b>	.4567	.4433	.5200	<b>.5400</b>
Recall@50	.5333	.5567	.5667	.5733	.5800	<b>.5867</b>	<b>.5967</b>	.5267	.5333	.5767	<b>.5867</b>
Recall@100	.6100	.6100	.6100	.6100	.6100	.6100	.6100	.6100	.6100	.6100	.6100

**Table 7: Dense-pool reranking effectiveness.** Results on 100 queries with 100 candidates per query. Best scores are highlighted in green; second-best in cyan. <sup>1</sup>MiniLM-6, <sup>2</sup>MiniLM-12, <sup>3</sup>monoT5, <sup>4</sup>Distilled-Voyage-ModernBERT, <sup>5</sup>Cohere rerank-v3.5, <sup>6</sup>Voyage rerank-2.5, <sup>7</sup>ColBERTv2, <sup>8</sup>RankZephyr-7B, <sup>9</sup>BGE-base, <sup>10</sup>BGE-large.

### Mean Reciprocal Rank (MRR@k):

MRR focuses on the position of the first relevant passage. For each query  $q$ , let

$$\text{rank}_q = \min\{i \leq k : \text{rel}(p_i) > 0\}$$

be the position of the first passage with non-zero relevance within the top- $k$ . We define the reciprocal rank for  $q$  as  $1/\text{rank}_q$  if such a passage exists and 0 otherwise. The mean reciprocal rank is then

$$\text{MRR@}k = \frac{1}{|Q|} \sum_{q \in Q} \frac{1}{\text{rank}_q}. \quad (3)$$

MRR@ $k$  therefore emphasises whether the system surfaces any relevant passage very early in the ranking, which is important when practitioners need to quickly locate at least one authoritative statutory provision.

### Recall@k:

Recall@ $k$  measures coverage of relevant passages among the top- $k$  candidates. Let  $G_q$  denote the set of all relevant passages for query  $q$  (here  $|G_q| = 3$  for all  $q$ ), and let  $R_q^{(k)}$  be the top- $k$  elements of  $R_q$ . We compute

$$\text{Recall@}k = \frac{1}{|Q|} \sum_{q \in Q} \frac{|\{p \in R_q^{(k)} : \text{rel}(p) > 0\}|}{|G_q|}. \quad (4)$$

## 5.2 Dense Pool Results

Table 7 reports effectiveness over the MPNet dense candidate pool. All rerankers deliver substantial gains over the MPNet-only baseline across all metrics. For instance, nDCG@10 improves from 0.3117 (MPNet-only) to 0.4500 with Voyage rerank-2.5 and 0.4211 with Distilled-Voyage-ModernBERT. The corresponding MRR@10 scores increase from 0.3984 to 0.5859 and 0.5647. These improvements mean that the most relevant statutory provisions, especially grade-3 passages, are usually surfaced near the very top of the ranking.

Voyage rerank-2.5 is the strongest overall model in this setting, and achieves the highest scores on all nDCG cut-offs and on MRR@10. Distilled-Voyage-ModernBERT is consistently the best-performing open model and closely tracks the teacher: its nDCG@10 is only 0.029 below Voyage and it outperforms all other open baselines, including BGE-large, BGE-base, MiniLM-12, monoT5, ColBERTv2 and RankZephyr-7B. That a distilled model trained on only 5,221 teacher-scored pairs matches or surpasses larger open rerankers underscores the value of targeted distillation for statutory retrieval.

The performance spread is most pronounced at shallow depths. At nDCG@10, Voyage improves over the weakest reranker (RankZephyr-7B) by 0.138 absolute (0.4500 vs. 0.3117), i.e., about 44% relative; by nDCG@100, the corresponding gap falls to around 0.115 (0.4835 vs. 0.3684; roughly 31% relative). As the cutoff increases, most rerankers eventually place the three relevant passages somewhere within the top 100 whenever MPNet has already recalled them so differences become less dramatic.

Recall metrics provide a complementary view. Because all rerankers act on the same candidate pool, Recall@100 is fixed at 0.61 (183/300 relevant passages retrieved by MPNet) and does not distinguish models. At smaller cut-offs, better rerankers achieve higher Recall@k: at Recall@10, Voyage reaches 0.4567 and Distilled-Voyage-ModernBERT 0.4000, compared with 0.3467 for MPNet-only. These gains indicate that rerankers are not merely reordering a handful of good candidates but substantially increase the likelihood that multiple relevant provisions appear near the top of the ranking.

From a legal perspective, improvements at these shallow cut-offs are crucial. In our benchmark, grade-3 passages are typically narrow operative provisions (for example sections that explicitly “must” or “may” do something), while grade-2 and grade-1 passages provide supporting or contextual material. High nDCG@10 and MRR@10 scores therefore reflect an ability to correctly privilege the provisions that instantiate statutory effect, rather than merely retrieving passages that talk about similar topics.

## 5.3 BM25 Pool Results

Table 8 shows the corresponding results when the candidate pool is obtained via BM25. The relative ordering of rerankers closely mirrors the dense setting: Voyage rerank-2.5 again achieves the highest nDCG@10 (0.4309), and Distilled-Voyage-ModernBERT remains the strongest open model with nDCG@10 = 0.3975. The relative gap between

Metric	BM25 Pool: Sparse Retrieval										
	BM25 only	Mini LM-6	Mini LM-12	mono T5	Dist. V-MB	Cohere r-v3.5	Voyage r-2.5	Col BERT	Rank Zeph	BGE base	BGE large
MRR@5	.3638	.4312	.4867	.4490	<b>.5343</b>	.5325	<b>.5853</b>	.4148	.3638	.5138	.5312
MRR@10	.3708	.4471	.5009	.4611	.5402	.5427	<b>.5940</b>	.4280	.3708	.5226	<b>.5453</b>
MRR@25	.3800	.4501	.5039	.4653	<b>.5432</b>	.5461	<b>.5946</b>	.4343	.3800	.5255	.5461
MRR@50	.3829	.4509	.5044	.4658	<b>.5437</b>	.5463	<b>.5946</b>	.4345	.3829	.5258	.5464
MRR@100	.3831	.4511	.5044	.4660	<b>.5438</b>	.5463	<b>.5946</b>	.4349	.3831	.5258	.5464
nDCG@5	.2301	.2976	.3300	.2904	<b>.3796</b>	.3575	<b>.4042</b>	.2726	.2301	.3354	.3460
nDCG@10	.2509	.3270	.3549	.3261	<b>.3975</b>	.3843	<b>.4309</b>	.3037	.2509	.3621	.3778
nDCG@25	.2784	.3480	.3743	.3490	<b>.4148</b>	.4071	<b>.4443</b>	.3282	.2784	.3823	.3931
nDCG@50	.2978	.3601	.3835	.3591	<b>.4206</b>	.4102	<b>.4504</b>	.3372	.2978	.3874	.3978
nDCG@100	.3111	.3647	.3871	.3637	<b>.4255</b>	.4126	<b>.4511</b>	.3454	.3111	.3915	.4017
Recall@5	.2300	.2800	.3000	.2733	.3167	<b>.3267</b>	<b>.3400</b>	.2633	.2300	.3167	.3200
Recall@10	.2833	.3700	.3867	.3667	.3767	.3967	<b>.4200</b>	.3533	.2833	.3833	<b>.4133</b>
Recall@25	.3667	.4400	.4567	.4500	<b>.4600</b>	.4833	<b>.4867</b>	.4367	.3667	.4667	.4767
Recall@50	.4567	.4933	.5067	.5000	.4967	<b>.5133</b>	<b>.5200</b>	.4800	.4567	.5033	.5067
Recall@100	.5267	.5267	.5267	.5267	.5267	.5267	.5267	.5267	.5267	.5267	.5267

**Table 8:** BM25-pool reranking effectiveness (100 queries; 100 candidates per query). Best scores are highlighted in green; second-best in cyan.

Distilled-Voyage-ModernBERT and Voyage is modest (about 7.8% in nDCG@10) and the two models have very similar MRR@10.

Cohere rerank-v3.5 forms the next tier, followed by BGE-large and BGE-base, with MiniLM-12, MiniLM-6, monoT5, ColBERTv2 and RankZephyr-7B trailing. The spread in nDCG@10 across rerankers is similar in magnitude to the dense pool and, again, largest at shallow depths.

Absolute scores in the BM25 pool are uniformly lower than in the dense pool. This is consistent with weaker candidate coverage: BM25 retrieves only 158/300 relevant passages (Recall@100 = 0.5267), compared to 183/300 (Recall@100 = 0.61) for MPNet. Since rerankers can only reorder the candidates they receive, this upper bound on recall constrains downstream nDCG and MRR. The stability of the model ranking across the two pools indicates that differences between rerankers largely reflect their ability to assign appropriate scores to statutory passages, rather than idiosyncrasies of a particular retrieval backbone.

From a legal informatics perspective, the lower BM25 coverage suggests that sparse lexical retrieval is more likely to miss highly operative provisions that do not share obvious surface forms with the query. Dense MPNet retrieval appears better able to retrieve such provisions somewhere in the candidate set, after which rerankers can identify and promote them. This aligns with recent findings in legal retrieval-augmented generation settings where lexical term matching alone is insufficient to surface the key statutory clause (Kalra et al. 2024; Zhang et al. 2025; Louis et al. 2024).

## 5.4 Cross-Pool Comparison

Comparing results across candidate pools reveals two key patterns. First, for all rerankers, nDCG and MRR are consistently higher in the dense pool than in the

BM25 pool. For example, Distilled-Voyage-ModernBERT attains  $\text{nDCG@10} = 0.4211$  on MPNet versus 0.3975 on BM25 (approximately 6% relative difference), while Voyage improves from 0.4309 to 0.4500 (around 4% relative). This reflects the higher  $\text{Recall@100}$  of the dense MPNet pool and confirms that rerankers can capitalise on richer candidate sets to produce better top-ranked lists.

Second, the *relative* standing of models is remarkably stable across pools. In both settings, Distilled-Voyage-ModernBERT is the top open model and the closest competitor to Voyage, consistently outperforming other open rerankers such as BGE-large, BGE-base, Cohere rerank-v3.5, MiniLM variants, monoT5, ColBERTv2 and RankZephyr-7B. The relative gap between Distilled-Voyage-ModernBERT and Voyage remains in the 6–8% range across  $\text{nDCG@10}$  and  $\text{MRR@10}$  in both pools. This suggests that the distillation process has captured a substantial portion of the teacher’s domain-specific ranking behaviour rather than overfitting to a particular retrieval regime.

The most consequential differences between models arise at shallow depths, where practitioners typically inspect results. In the dense pool, for example, Voyage’s  $\text{nDCG@10}$  exceeds RankZephyr-7B’s by 0.138 absolute (0.4500 vs. 0.3117), whereas by  $\text{nDCG@100}$  the gap shrinks to 0.115 (0.4835 vs. 0.3684); analogous patterns hold in the BM25 pool. This pattern indicates that high-quality rerankers add most value in the top ranks, ensuring that the first few passages presented to the user are not merely on-topic but legally operative. For statutory retrieval scenarios such as compliance checking or advisory work, these top-ranked differences are substantially more important than marginal gains at deeper ranks.

Taken together, these results suggest that dense retrieval plus strong reranking is preferable to sparse retrieval alone for statute-focused tasks and a carefully distilled cross-encoder can deliver robust gains across retrieval backends, which allows for open and locally deployable systems.

## 5.5 Statistical Significance Testing

To determine whether observed differences in  $\text{nDCG@10}$  reflect genuine improvements rather than sampling variability, we conduct query-level statistical tests comparing Distilled-Voyage-ModernBERT to each baseline reranker. All analyses in this subsection use the dense MPNet pool. The BM25 results exhibit qualitatively identical patterns.

### *Testing procedure.*

For each baseline model  $b$ , we compute the per-query difference in  $\text{nDCG@10}$ :

$$\Delta_q^{(b)} = \text{nDCG@10}_{\text{Distilled}}(q) - \text{nDCG@10}_b(q), \quad q \in Q.$$

We then apply a paired Wilcoxon signed-rank test to the distribution  $\{\Delta_q^{(b)}\}_{q \in Q}$  to test the null hypothesis that the median difference is zero. To control the family-wise error rate across the nine baselines, we use Holm-Bonferroni correction with global

Baseline Model	Mean Diff.	Mean 95% CI	Cohen’s $d$	$p$ -value (Holm)	Outcome
<i>Significant improvements (<math>p &lt; 0.01</math>):</i>					
RankZephyr	+0.143	[0.085, 0.204]	0.47	<b>¡0.001</b>	<b>Win</b>
ColBERT	+0.131	[0.075, 0.187]	0.45	<b>¡0.001</b>	<b>Win</b>
monoT5	+0.088	[0.041, 0.135]	0.36	<b>0.002</b>	<b>Win</b>
MiniLM-6	+0.093	[0.038, 0.148]	0.33	<b>0.003</b>	<b>Win</b>
<i>No significant difference (<math>p &gt; 0.05</math>):</i>					
MiniLM-12	+0.061	[0.007, 0.114]	0.22	0.066	No sig. diff.
Cohere	+0.038	[-0.007, 0.084]	0.16	0.396	No sig. diff.
BGE-base	+0.034	[-0.020, 0.088]	0.12	0.360	No sig. diff.
BGE-large	+0.023	[-0.027, 0.071]	0.09	0.516	No sig. diff.
Voyage	-0.031	[-0.079, 0.016]	-0.13	0.333	No sig. diff.

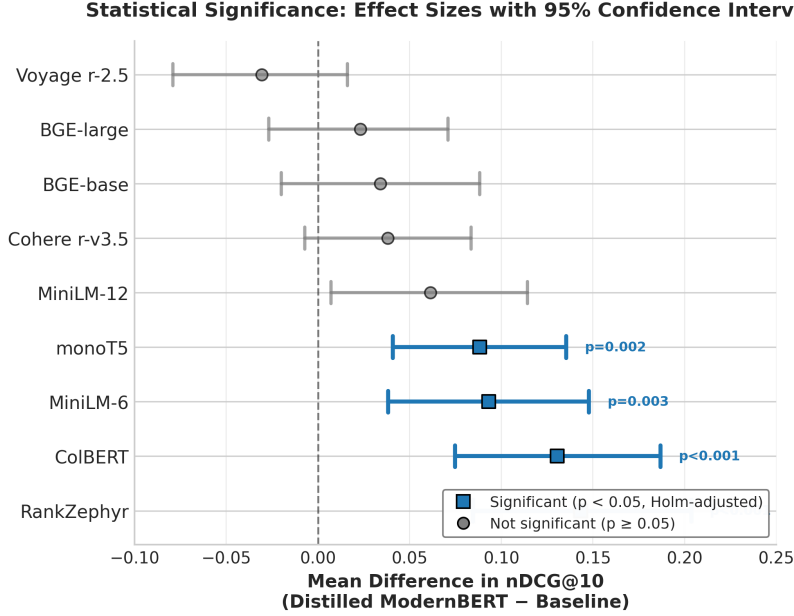
**Table 9:** Statistical comparison of Distilled-Voyage-ModernBERT with baseline rerankers on nDCG@10 (dense MPNet pool, 100 queries). Positive mean differences favour Distilled-Voyage-ModernBERT.

$\alpha = 0.05$ . For each comparison we also report Cohen’s  $d$  effect size (computed on the per-query differences) and 95% bootstrap confidence intervals for the mean difference (10,000 resamples). Table 9 reports, for each baseline reranker, the mean nDCG@10 difference between Distilled-Voyage-ModernBERT and the baseline, its 95% bootstrap confidence interval, Cohen’s  $d$ , the Holm–Bonferroni adjusted  $p$ -value and an outcome label indicating whether the student model achieves a statistically significant win or shows no statistically significant difference.

### **Findings:**

Distilled-Voyage-ModernBERT achieves statistically significant improvements over four widely used open rerankers: RankZephyr, ColBERT, monoT5 and MiniLM-6. For these models, the Holm-adjusted  $p$ -values are below 0.01, and the 95% confidence intervals for the mean difference are strictly positive. Effect sizes are in the small-to-moderate range (Cohen’s  $d$  between 0.33 and 0.47), indicating that the improvements are not only statistically reliable but also practically meaningful in a 100-query evaluation. **Figure 3 visualises these results, showing the effect sizes and 95% confidence intervals for the mean nDCG@10 difference for each baseline comparison.**

For the remaining baselines of MiniLM-12, Cohere rerank-v3.5, BGE-base, BGE-large and Voyage, the adjusted  $p$ -values exceed 0.05 and the confidence intervals for the mean nDCG@10 difference all include zero. We therefore conclude that these comparisons show no statistically significant difference after Holm–Bonferroni correction. On our 100-query statutory benchmark, this suggests that Distilled-Voyage-ModernBERT performs comparably to the strongest open and commercial rerankers, rather than clearly underperforming any of them. Voyage rerank-2.5 remains slightly ahead in mean nDCG@10 (by 0.031 points; Cohen’s  $d = -0.13$ ), but this gap is small and not statistically significant after correction, which is consistent with the distilled



**Fig. 3:** Effect sizes with 95% confidence intervals for the mean nDCG@10 difference between Distilled-Voyage-ModernBERT and each baseline. Blue markers indicate statistically significant differences after Holm–Bonferroni correction; grey markers indicate non-significant differences. The vertical dashed line at zero corresponds to no difference. Models are ordered by effect size.

student achieving teacher-level effectiveness within the resolution of our evaluation. MiniLM-12 is a borderline case: its uncorrected  $p$ -value would favour Distilled-Voyage-ModernBERT, but the effect is no longer significant after Holm adjustment ( $p_{\text{adj}} = 0.066$ ), so we conservatively report this as no statistically significant difference after correction. Taken together with the significant improvements over RankZephyr, ColBERT, monoT5 and MiniLM-6 reported above, this pattern supports our claim that the distilled model delivers top-tier statutory retrieval quality while remaining fully open and deployable without reliance on a closed API.

**Query-level behaviour:**

Per-query win/tie/loss counts reinforce this picture. Against ColBERT, Distilled-Voyage-ModernBERT achieves higher nDCG@10 on 56 of 78 non-tied queries (72%; sign test  $p < 0.001$ ). Against RankZephyr, it wins 51 of 71 non-tied queries (also 72%; sign test  $p < 0.001$ ). In contrast, against Voyage, Distilled-Voyage-ModernBERT wins 30 of 66 non-tied queries (45%; sign test  $p > 0.05$ ), consistent with a statistical tie between the two models. Bootstrap distributions over mean differences (10,000 resamples) exhibit the same pattern: for weaker baselines, the entire distribution lies above zero and for top-tier models, it straddles zero.

### ***Summary:***

Overall, the statistical analysis confirms that distillation from a proprietary teacher using only 5,221 scored pairs yields an open reranker that significantly outperforms several widely deployed baselines, while achieving performance indistinguishable from stronger proprietary and large open rerankers on our statutory benchmark. The effect sizes are sufficiently large to be relevant in practice and the results are robust under conservative multiple-comparison control. This strengthens the methodological conclusion that knowledge distillation is a viable pathway for constructing high-quality open rerankers for legal retrieval without continued reliance on closed APIs.

## **5.6 Qualitative Analysis**

Aggregate metrics summarise overall performance, so we also qualitatively inspect the queries using expert relevance labels to see where Distilled-Voyage-ModernBERT differs from the Voyage teacher.

### ***Success on directly operative provisions:***

For queries whose grade-3 passage is a clearly operative provision, Distilled-Voyage-ModernBERT typically matches Voyage’s behaviour. Consider Q0055, which asks:

*“Under the Acts of the Scottish Parliament, what are the requirements for preparing a circular economy strategy?”*

The grade-3 passage is the section entitled **Circular economy strategy**, which explicitly states that the Scottish Ministers *must* prepare a strategy and *must set out* specified objectives, plans and monitoring arrangements. This is a paradigmatic operative clause as it contains clear deontic language (“must”) and enumerates statutory requirements. Both Voyage and Distilled-Voyage-ModernBERT rank this passage first, whereas ColBERT ranks it fourth. Here, the distilled model correctly aligns the query intent with the statutory language that actually creates the obligation, rather than with surrounding explanatory or contextual material.

A similar pattern appears in Q0047, which asks:

*“Under the Acts of the Scottish Parliament, how are disqualifications for charity trustees specified?”*

The grade-3 passage inserts Section 69A titled **Disqualification: specified offences**, which enumerates the offences triggering disqualification. This is again a narrowly drafted high-salience operative provision. Voyage ranks this passage first, Distilled-Voyage-ModernBERT second and ColBERT thirteenth. In such cases, the distilled model is able to prioritise the specific section that implements the disqualification regime, capturing the legal structure rather than merely matching surface terms like “trustee” or “charity”.

### *Difficulties with contextual statutory provisions.*

The same query Q0047 also exposes a recurring weakness. The grade-1 passage titled **Record of persons removed from office** requires the Scottish Charity Regulator (OSCR) to keep a record of persons removed by court order from management roles. This provision is contextually related to disqualification. It concerns administrative consequences for persons who have already been removed but it does not itself specify the disqualification criteria. It is therefore annotated as grade 1: relevant context but not a direct answer.

Voyage ranks this grade-1 passage at position 23, recognising its peripheral but non-negligible relationship to the query. Distilled-Voyage-ModernBERT, by contrast, relegates it to position 92, while ColBERT ranks it 57th. This suggests that the distilled model is relatively conservative in assigning non-zero relevance to passages whose connection to the query is mediated by institutional roles (for example OSCR’s record-keeping duty) rather than by explicit repetition of the query’s core concepts (“disqualification”, “specified offences”).

### *Patterns across relevance grades:*

To systematically characterise this behaviour, we compute Recall@5 separately for each relevance grade, treating passages of other grades as non-relevant.<sup>3</sup> On grade-3 passages, the highest typically operative relevance level—Distilled-Voyage-ModernBERT achieves 78.2% Recall@5, slightly surpassing Voyage’s 75.6%. This confirms the impression from Q0055 and Q0047: the distilled model is particularly strong at identifying the single most operative provision that answers the query.

On grade-1 passages, however, performance drops: Distilled-Voyage-ModernBERT attains 38.5% Recall@5 compared to Voyage’s 48.1%. Grade-1 passages often encode background conditions, ancillary procedures, or cross-referenced powers that situate the operative provision within a broader statutory scheme. From a legal standpoint, they are useful to a reader seeking a complete understanding of the regulatory context but are not strictly necessary to answer the query. The distilled model appears more inclined than the teacher to concentrate probability mass on provisions that explicitly restate the query’s key terms or deontic operators, at the expense of these context-setting clauses.

One interpretation of this pattern is that knowledge distillation, as implemented here, encourages the student to mimic the teacher’s scores most faithfully in the high-confidence regime (that is, near the most relevant passages), while compressing score differences among lower-confidence passages. In statutory retrieval, this manifests as excellent performance on directly operative provisions (grade 3), competitive performance on strongly supporting provisions (grade 2) and a tendency to under-rank contextual but legally adjacent provisions (grade 1). In practice, this behaviour may be acceptable or even desirable in applications where the primary goal is to retrieve

---

<sup>3</sup>Grade-specific recall is computed as:  $\text{grade-}g \text{ Recall@5} = (\text{number of grade-}g \text{ passages retrieved within the top 5}) / (\text{total number of grade-}g \text{ passages})$ , aggregating over queries. This differs from the Recall@k values in Tables 7 and 8, which aggregate over all relevance grades.

the operative clause. However, for tasks that require reconstructing the full legislative narrative around an issue, it suggests that additional mechanisms (for example diversification or structure-aware reranking) may be beneficial.

***Legal implications:***

The qualitative analysis thus complements the aggregate statistics in two ways. First, it confirms that much of the nDCG and MRR advantage of strong rerankers is driven by their ability to place operative statutory provisions very high in the ranking. Second, it reveals a nuanced trade-off: the Distilled-Voyage-ModernBERT model occasionally sacrifices “soft” contextual coverage in favour of sharper focus on provisions that directly create, modify or extinguish legal effects. For legal practitioners concerned primarily with the existence and content of obligations, offences or powers, this behaviour is often appropriate. For research tasks that require tracing the full network of statutory conditions and consequences, future work could explore combining such distilled cross-encoders with explicit models of legislative structure, or augmenting training data with examples where contextual provisions are labelled as desirable secondary targets.

Additional examples illustrating these patterns, including queries where the student surpasses the teacher and queries that highlight failure modes on multi-step statutory reasoning, are provided in Appendix A.

## 6 Discussion

Our experiments show that, on the 100-query evaluation dataset over UK-STATUTECORPUS, neural rerankers improve provision-level retrieval for UK legislation compared with using BM25 or MPNet alone. Across both sparse (BM25) and dense (MPNet) candidate pools, the rerankers we evaluate increase nDCG and MRR at cut-offs such as @5 and @10. Because the evaluation labels distinguish legally operative, strongly supporting and contextual provisions, higher scores correspond to ranking provisions that directly address the statutory information need ahead of passages that are only supporting or contextual.

The comparison between dense and sparse retrieval shows how the two stages interact in this setting. MPNet attains higher Recall@100 than BM25 and, with the same reranker, yields higher nDCG and MRR values on the evaluation dataset. In the two-stage design we use, the first-stage retriever fixes the set of candidate provisions and the reranker orders this set. The observed differences are therefore due to changes in the ranking of these fixed candidates rather than to changes in recall.

The distillation experiments show that these gains can be realised without continued reliance on a proprietary reranking API. Distilled-Voyage-ModernBERT, trained on 5,221 query–passage pairs scored by Voyage rerank-2.5, achieves nDCG@10 and MRR@10 on the MPNet candidate pool that are close to the teacher’s values on this evaluation dataset and higher than those of the other open rerankers we consider. The same pattern holds, with slightly lower absolute scores, on the BM25 candidate pool. Taken together, the corpus, evaluation dataset, distillation data and open

reranker indicate that, for UK legislation in the period we study, provision-level statutory retrieval can be improved by adding a neural reranking stage and that a compact cross-encoder can match or exceed the other open baselines while approaching the ranking quality of a proprietary reranker.

## 7 Limitations and Future Directions

This study introduces a provision-level evaluation benchmark for UK primary and secondary legislation and compares neural rerankers, including a distilled ModernBERT cross-encoder, in a controlled intra-instrument setting. We fix BM25 and MPNet as first-stage retrievers and evaluate on a 100-query, expert-validated test set drawn from legislation enacted between 2020 and 2024 so that performance differences can be attributed to reranking rather than retrieval. Our results therefore characterise provision-level, single-instrument retrieval over recent UK legislation, extending the framework to cross-instrument retrieval and to larger or more diverse query sets is left for future work.

## 8 Conclusion

This paper examined provision-level retrieval and neural reranking for UK primary and secondary legislation via UK-STATUTECORPUS, a corpus of 12,604 instruments segmented into 124,796 passages, and a 100-query benchmark whose graded labels distinguish legally operative, strongly supporting and contextual provisions. With BM25 and MPNet as fixed first-stage retrievers, modern neural rerankers substantially improve statutory retrieval as, across both sparse and dense candidate pools, they raise nDCG and MRR and, crucially, bring the clauses that create or modify rights and duties to the top of the ranking. Because evaluation is based on expert-verified provision-level relevance, these gains directly measure an increased ability to surface the clauses that create, modify or clarify legal rights and duties, not just passages on the same topic.

Within this setting we introduced Distilled-Voyage-ModernBERT, a ModernBERT-based cross-encoder distilled from the proprietary Voyage rerank-2.5 on 5,221 query-passage pairs. On the dense MPNet pool it matches the teacher in nDCG@10 up to non-significant differences and is numerically close in MRR, while achieving statistically significant nDCG@10 gains over RankZephyr-7B, ColBERTv2, monoT5 and MiniLM-6 with small-to-moderate effect sizes. Across both dense and sparse pools it is the strongest open reranker and the closest open competitor to the Voyage teacher. The corpus, benchmark, distillation dataset and open reranker released with this work therefore provide an experimental basis for statutory retrieval that is both legally meaningful and technically reproducible, and demonstrate that high-quality provision-level reranking for UK legislation need not rely on closed APIs. In essence, with carefully designed resources and targeted distillation, an open model can approximate proprietary legal rerankers closely enough to support AI-assisted legal research and retrieval-augmented systems under demanding requirements of transparency, audit and data governance.

## References

- Ashley KD (1991) Reasoning with cases and hypotheticals in HYPO. *International journal of man-machine studies* 34(6):753–796
- Ashley KD (2002) An AI model of case-based legal argument from a jurisprudential viewpoint. *Artificial Intelligence and Law* 10(1):163–218
- Askari A, Yang Z, Ren Z, et al (2024) Answer retrieval in legal community question answering. In: *Advances in Information Retrieval: 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24–28, 2024, Proceedings, Part III*. Springer, pp 273–289
- Bajaj P, Campos D, Craswell N, et al (2016) MS MARCO: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:161109268*
- Baldelli D, Jiang J, Aizawa A, et al (2024) Twolar: A two-step llm-augmented distillation method for passage reranking. In: Goharian N, Tonellotto N, He Y, et al (eds) *Advances in Information Retrieval*. Springer Nature Switzerland, Cham, pp 470–485
- Chalkidis I, Fergadiotis M, Malakasiotis P, et al (2020) LEGAL-BERT: The muppets straight out of law school. *arXiv preprint arXiv:201002559*
- Chalkidis I, Fergadiotis M, Androutsopoulos I (2021) MultiEURLEX—a multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp 6974–6996
- Chalkidis I, Garneau N, Goanta C, et al (2023) LeXFiles and LegalLAMA: Facilitating English multinational legal language model development. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp 15513–15535
- Congressional Research Service (2020) *Understanding Federal Legislation: A section-by-section guide to key legal considerations* Available at <https://sgp.fas.org/crs/misc/R46484.pdf>
- Hendrycks D, Burns C, Chen A, et al (2021) Cuad: An expert-annotated nlp dataset for legal contract review. In: *Proceedings of the Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*, pp 1–12, URL [https://datasets-benchmarks-proceedings.neurips.cc/paper\\_files/paper/2021/file/6ea9ab1baa0efb9e19094440c317e21b-Paper-round1.pdf](https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/6ea9ab1baa0efb9e19094440c317e21b-Paper-round1.pdf)
- Hofstätter S, Althammer S, Schröder M, et al (2021a) Improving efficient neural ranking models with cross-architecture knowledge distillation. In: *arXiv preprint arXiv:2010.02666*

- Hofstätter S, Lin SC, Yang JH, et al (2021b) Efficiently teaching an effective dense retriever with balanced topic aware sampling. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 113–122
- Hou AB, Weller O, Qin G, et al (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. Association for Computational Linguistics, Albuquerque, New Mexico, pp 7898–7913
- Johnson J, Douze M, Jégou H (2019) Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data* 7(3):535–547
- Kalra R, Wu Z, Gulley A, et al (2024) HyPA-RAG: A hybrid parameter adaptive retrieval-augmented generation system for AI legal and policy applications. arXiv preprint arXiv:240909046
- Khattab O, Zaharia M (2020) ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In: Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, pp 39–48
- Kim Y, Choi Y, Choi E, et al (2024) Developing a pragmatic benchmark for assessing Korean legal language understanding in large language models. In: Findings of the Association for Computational Linguistics: EMNLP 2024, pp 5573–5595
- Lewis P, Perez E, Piktus A, et al (2020) Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems* 33:9459–9474
- Louis A, van Dijck G, Spanakis G (2023) Interpretable long-form legal question answering with retrieval-augmented large language models. In: arXiv preprint arXiv:2309.17050
- Louis A, van Dijck G, Spanakis G (2024) Interpretable long-form legal question answering with retrieval-augmented large language models. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 22266–22275
- Muennighoff N, Tazi N, Magne L, et al (2023) MTEB: Massive text embedding benchmark. In: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, Dubrovnik, Croatia, pp 2014–2037
- Nguyen C, Le NK, Nguyen DH, et al (2022) A legal information retrieval system for statute law. In: Asian conference on intelligent information and database systems, Springer, pp 370–382

- Nogueira R, Jiang Z, Pradeep R, et al (2020) Document ranking with a pretrained sequence-to-sequence model. In: Findings of the Association for Computational Linguistics: EMNLP 2020, pp 708–718
- Nussbaum Z, Morris JX, Duderstadt B, et al (2024) Nomic embed: Training a reproducible long context text embedder. [arXiv:2402.01613](https://arxiv.org/abs/2402.01613)
- Office of the Parliamentary Counsel (2013) When laws become too complex. URL [https://assets.publishing.service.gov.uk/media/5a7a2ce9e5274a34770e4c80/GoodLaw\\_report\\_8April\\_AP.pdf](https://assets.publishing.service.gov.uk/media/5a7a2ce9e5274a34770e4c80/GoodLaw_report_8April_AP.pdf), cabinet Office Report
- Ollama Inc. (2025) Ollama: Local large language model inference framework. <https://github.com/ollama/ollama>, accessed 9 December 2025
- Pipitone N, Alami GH (2024) LegalBench-RAG: A benchmark for retrieval-augmented generation in the legal domain. arXiv preprint arXiv:240810343
- Pradeep R, Sharifymoghaddam S, Lin J (2023a) RankVicuna: Zero-shot listwise document reranking with open-source large language models. In: arXiv preprint arXiv:2309.15088
- Pradeep R, Sharifymoghaddam S, Lin J (2023b) RankZephyr: Effective and robust zero-shot listwise reranking is a breeze! In: arXiv preprint arXiv:2312.02724
- Robertson S, Zaragoza H, et al (2009) The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval* 3(4):333–389
- Santhanam K, Khattab O, Saad-Falcon J, et al (2022) ColBERTv2: Effective and efficient retrieval via lightweight late interaction. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp 3715–3734
- Song K, Tan X, Qin T, et al (2020) MPNet: Masked and permuted pre-training for language understanding 33:16857–16867
- Sun W, Yan L, Ma X, et al (2023) Is ChatGPT good at search? investigating large language models as re-ranking agent. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp 14918–14937
- Thakur N, Reimers N, Rücklé A, et al (2021) BEIR: A heterogenous benchmark for zero-shot evaluation of information retrieval models. arXiv preprint arXiv:210408663
- Wang K, Thakur N, Reimers N, et al (2022) GPL: Generative pseudo labeling for unsupervised domain adaptation of dense retrieval. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp 2345–2360

Wang S, Scardigli A, Tang L, et al (2023) MAUD: An expert-annotated legal NLP dataset for merger agreement understanding. In: Bouamor H, Pino J, Bali K (eds) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Singapore, pp 16369–16382, <https://doi.org/10.18653/v1/2023.emnlp-main.1019>, URL <https://aclanthology.org/2023.emnlp-main.1019/>

Xiao S, Liu Z, Zhang P, et al (2024) C-pack: Packed resources for general chinese embeddings. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, <https://doi.org/10.1145/3626772.3657878>

Yang R (2024) CaseGPT: a case reasoning framework based on language models and retrieval-augmented generation. arXiv preprint arXiv:240707913

Zhang Z, Li X, Zhao Y, et al (2025) LexRAG: Benchmarking retrieval-augmented generation in multi-turn legal consultation conversation. In: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval

## Appendix A Additional Qualitative Examples

This appendix provides additional examples of model ranking behavior across different relevance grades, complementing the analysis in Section 5.6.

### A.1 Cases Where Distilled-Voyage-ModernBERT Excels

*Q0013: Distilled outperforms Voyage on grade-3 passage.*

The Distilled-Voyage-ModernBERT ranks the relevant passage at position 1, while Voyage ranks it at position 7 and ColBERT at position 4. This demonstrates that on some high-relevance passages, the Distilled-Voyage-ModernBERT can exceed the teacher’s performance, likely due to learning patterns in the synthetic training data that generalize well to this particular query-document pair.

- **Rankings:** Voyage=7, Distilled=1, ColBERT=4
- **Relevance:** Grade 3 (directly relevant)

*Q0054: Near-perfect agreement across models.*

All models perform well on this grade-3 passage, with the Distilled-Voyage-ModernBERT ranking it at position 1, demonstrating successful knowledge transfer for clearly relevant statutory content.

- **Rankings:** Voyage=2, Distilled=1, ColBERT=2
- **Relevance:** Grade 3 (directly relevant)

***Q0088: Distilled significantly outperforms Voyage.***

The Distilled-Voyage-ModernBERT ranks the grade-3 passage at position 2, while Voyage ranks it at position 10, with ColBERT also at position 2. This suggests the Distilled-Voyage-ModernBERT has learned effective ranking strategies for certain statutory patterns.

- **Rankings:** Voyage=10, Distilled=2, ColBERT=2
- **Relevance:** Grade 3 (directly relevant)

## A.2 Cases Showing Distillation Gaps

***Q0005: Distilled-Voyage-ModernBERT lags on grade-3 passage.***

While Voyage ranks this relevant passage at position 5, the Distilled-Voyage-ModernBERT ranks it at position 31, with ColBERT at position 99. This example shows that even for high-relevance passages, Distilled-Voyage-ModernBERT occasionally underperforms when query-document alignment is less obvious.

- **Rankings:** Voyage=5, Distilled=31, ColBERT=99
- **Relevance:** Grade 3 (directly relevant)

***Q0041: ColBERT outperforms both dense models.***

In this unusual case, ColBERT ranks the grade-3 passage at position 3, while Voyage ranks it at 6 and the Distilled-Voyage-ModernBERT at 34. This suggests the passage contains strong lexical overlap with the query that dense models fail to exploit.

- **Rankings:** Voyage=6, Distilled=34, ColBERT=3
- **Relevance:** Grade 3 (directly relevant)

## A.3 Contextual Relevance Challenges

***Q0026: Grade-3 and grade-1 passages.***

This query illustrates the performance gap across relevance grades. For the grade-3 passage, all models rank it in mid-range (Voyage=42, Distilled=57, ColBERT=57), suggesting challenging query-document alignment. For the grade-1 passage, Voyage ranks it at position 10 while the Distilled-Voyage-ModernBERT ranks it at position 49, demonstrating Distilled-Voyage-ModernBERT’s difficulty with contextual relevance.

- **Grade-3 rankings:** Voyage=42, Distilled=57, ColBERT=57
- **Grade-1 rankings:** Voyage=10, Distilled=49, ColBERT=58

## A.4 Interpretation

The Distilled-Voyage-ModernBERT achieves strong performance on directly relevant passages (grade-3) where explicit query-passage alignment is present. On 78.2% of grade-3 passages, Distilled-Voyage-ModernBERT ranks them in the top 5, compared

to 75.6% for Voyage. However, on contextually related content (grade-1), performance drops to 38.5% recall@5 versus 48.1% for Voyage.

This reflects a fundamental challenge in distillation from large proprietary models to compact student models: while the student successfully learns to mimic the teacher’s ranking of explicitly relevant content, it captures less of the teacher’s nuanced understanding of implicit statutory relationships, cross-references, and broader legal context. The Distilled-Voyage-ModernBERT’s mean rank on grade-3 passages (6.2) is competitive with the best open-weight baseline (BGE-large, also 6.2), but the gap to Voyage (4.8) reveals room for improvement in capturing subtle relevance signals.

Notably, in some cases (Q0013, Q0088), the Distilled-Voyage-ModernBERT outperforms Voyage, suggesting that the synthetic training data and distillation process can sometimes produce rankings that are equally or more effective than the teacher model. This provides evidence that distillation with carefully constructed synthetic queries can transfer ranking knowledge successfully within specific domains.