COPENHAGEN BUSINESS SCHOOL
HANDELSHØJSKOLEN

Assessing Clause Retrieval Pipelines for Legal Contract Drafting

Course professors:	[redacted]
Course:	Natural Language Processing and Text Analytics (CDSCO1002U.LA_F25)
Program:	M.Sc. in BA & Data Science, Copenhagen Business School
Date of submission:	2025-05-30
Character Count:	34,095
Page Count:	15
Students & IDs:	Bastian Meyer-Karlsen [redacted], Elias Salvador Smidt Torjani [redacted], Francesco Esposito [redacted], and Mattia Malipiero [redacted]

Abstract

The manual search for precedent clauses remains a significant inefficiency in commercial contracting, partly because prevailing retrieval benchmarks fail to capture the length, terminology, as well as graded relevance standards of legal drafting. This study adapts the ACORD corpus, combining its official files into a single set of 126,659 query-clause pairs across nine query categories, and evaluate three retrieval tiers: BM25, semantic bi-encoders (MiniLM and Sentence-BERT), and a compact cross-encoder (GPT 4.1 nano) used as a reranker. Results show that BM25 attains an NDCG@5 of 0.475 and retrieves four- or five-star clauses in roughly 40% of cases. Sentence-BERT alone reaches NDCG@5 0.686; when reranked with GPT-4.1 nano it retains a comparable NDCG score while boosting three- and four-star precision (0.833 and 0.639, respectively) though five-star precision remains low. These findings indicate that modern retrieval stacks substantially ease clause discovery yet do not remove the need for expert review, given the limited number of extremely relevant clauses.

Keywords: Retrieval-augmented drafting, Normalized Discounted Cumulative Gain (NDCG), information retrieval, bi-directional encoder transformer (BERT), cross-encoder

Table of Contents

Introduction

Negotiating and drafting commercial contracts remains one of the most resource-intensive and financially consequential activities in corporate practice (Menear, 2024). Industry analyses estimate that ineffective contract development and oversight "leaks" approximately 9 percent of expected contract value, eroding enterprise revenue through ambiguous language and protracted negotiation cycles (World Commerce & Contracting, 2023). The human effort behind those losses is equally stark. A 2024 survey of 150 in-house and law-firm lawyers found that nearly half spend at least three hours reviewing a single contract, with a third requiring three-to-five hours for routine red-lining (Bachman, 2024). Such time sinks translate directly into legal spend, opportunity cost, reduced organizational agility, and heightened risk.

Against this backdrop, retrieval-augmented drafting has gained substantial traction in recent times. Instead of composing clauses de novo, lawyers are now able to first surface high-quality precedent language using modern natural-language processing (NLP) pipelines. Transformer encoders map clauses to dense semantic vectors, enabling sub-second retrieval across corpora of hundreds of thousands of provisions, while retrieval-augmented generation (RAG) lets large language models (LLMs) draft tailored red-lines from the retrieved text (Devlin et al., 2018; Lewis et al., 2020)

However, despite these advances, the dominant public benchmarks, most notably MS MARCO and BEIR, provide limited guidance for the contract-review setting (Bajaj et al., 2016; Thakur et al., 2021). Both corpora are drawn largely from web passages or general domain question answering, thus lacking the specialized vocabulary that commercial agreements entail. Moreover, conventional benchmarks employ binary relevance labels, whereas practicing lawyers distinguish finer shades of applicability. Progress is therefore heavily contingent on the availability of comprehensive datasets that expose models to the full stylistic and semantic variance of legal drafting.

To expand on the scope of this, the study leverages the Atticus Clause Retrieval Dataset (ACORD), an expert-annotated corpus that casts clause retrieval as a ranked-relevance task, effectively evaluating whether modern NLP can bridge the efficiency and accuracy gap that persists in this field (Wang et al., 2025). As such, this paper will systematically compare: (1) classical term-frequency methods, (2) dense-embedding retrieval, and (3) LLM-based reranking, reporting results on lawyer-centric metrics. In grounding the evaluation in ACORD, the study provides an empirically strong assessment of how far current techniques advance contract analytics and identifies the residual challenges that must be solved before retrieval-augmented drafting can deliver its full economic promise.

Related Work

Early information-retrieval (IR) research has relied on web-based collections such as MS MARCO and BEIR, yet, as previously mentioned, these datasets contain short, general-domain passages and binary relevance labels that do not reflect contractual prose (Bajaj et al., 2016; Thakur et al., 2021). Recognizing this mismatch, researchers have begun curating corpora that mirror the length, terminology, and nuance of real-world contracts. As such, the Contract Understanding Atticus Dataset (CUAD) represents a prominent example, providing 13,000 expert annotations across 510 commercial agreements and covering 41 clause categories that lawyers routinely inspect in mergers and financing deals (Hendrycks et al., 2021). CUAD's span-level labels, supplied by practicing attorneys, enable models to learn highly granular signals that are absent from earlier benchmarks.

Subsequent resources extend CUAD's spirit. For instance, the LexGLUE dataset aggregates seven legal tasks, including contract clause classification, to encourage multi-task evaluation, and the Competition on Legal Information Extraction/Entailment (COLIEE) events add annual retrieval challenges in case and statute law (Chalkidis et al., 2021; Goebel et al., 2024). Nevertheless, clause-centric retrieval datasets remain sparse. CUAD therefore remains the principal foundation on which recent clause-ranking benchmarks, including ACORD, have been constructed.

Traditional IR methods continue to underpin legal text systems, despite the ascendancy of neural models. Term frequency approaches such as TF-IDF and the BM25 ranking function remain the first stage of many pipelines as they scale linearly, handle rare legal terminology without supervision, and deliver strong recall in both statutory and contractual corpora (Robertson & Zaragoza, 2009). Beyond overall retrieval, classical methods have been adapted to contract structure. Paragraph-level scoring mitigates the dilution effect of lengthy agreements, while header aware weighting prioritizes sections that lawyers consult most. Consequently, BM25 serves not only as a benchmark but as a practical backbone for hybrid systems; an insight that has motivated our own use of BM25 as the entry point for clause retrieval.

A second strand of research adapts general purpose transformers to the stylistic and structural idiosyncrasies of legal text. LEGAL-BERT was the first family trained from scratch on statutes and case law, as well as contracts, achieving consistent gains across classification and clause extraction tasks (Chalkidis et al., 2020). Later work extends this approach to long form inputs. Mamakas et al. (2022) warm-start a Longformer from Legal-BERT and raise the context window to 8,000 sub-words, a scale needed for unabridged agreements and judicial opinions. Parallel advances in dense retrieval map queries and passages into a shared embedding space, enabling rapid nearest-neighbour search

without lexical overlap. As such, Dense Passage Retrieval (DPR) remains a canonical method and has been replicated on statutory and contract corpora (Karpukhin et al., 2020). Together, pretrained encoders and dense indexing signify a shift from keyword matching toward semantic retrieval that respects the jargon and hierarchical cross-references characterized by legal drafting.

Building on these advances in legal IR, a growing body of work integrates search with LLMs and RAG pipelines to automate downstream drafting tasks (Lewis et al., 2020). Case studies show that prototypes mediated with RAG pipelines can cut lawyer review time by 30-40% on standard agreements (Bommarito & Katz, 2022). As such, recent surveys catalogue a wave of domain-specific deployments; from suggestion widgets in contract lifecycle platforms to chat-based assistants that surface governing law or indemnity language on demand (Johnston, 2025).

Problem Definition & Research Objective

This study addresses the practical problem a lawyer faces when drafting or negotiating an agreement: *given a specific drafting question*, *quickly locate the most appropriate precedent clauses in a large contract library*. Each search query must be matched against thousands of clauses. To reflect professional practice, every query-clause pairing in ACORD is graded on a five-point scale, where one star denotes irrelevance and five stars indicate a clause the annotating attorneys would reuse verbatim. The retrieval task thus requires producing a ranked list in which highly rated clauses appear near the top, minimizing the manual search effort during drafting. In the analysis, we compare alternative retrieval pipelines (lexical, dense, and LLM-based) on their ability to deliver these legal clauses efficaciously.

Ultimately, model effectiveness is captured with graded, lawyer-centric metrics, encompassing normalized discounted cumulative gain (NDCG) as well as precision, as these scores reflect how effectively practitioners identify clauses they could plausibly reuse. Moreover, efficiency is gauged by the computational cost of each pipeline, including indexing footprint and average query latency, as actual adoption depends on both accuracy and speed. All comparisons are performed on the original ACORD relevance labels. As such, the following research question will serve as the academic cornerstone for this study:

To what extent do current information retrieval stacks close the gap between the clauses lawyers need and the clauses they can find, without incurring prohibitive computational overhead?

Methodology

Data Description

The ACORD dataset was compiled to furnish a specialized benchmark for contract-clause search, complementing resources such as CUAD with similar focus on extraction (Wang et al., 2025). Its source corpus comprises roughly 450 public agreements, approximately 400 US SEC-filed contracts inherited from CUAD, as well as 50 Fortune 500 terms of service documents, from which annotators extracted all passages responsive to nine clause categories: (1) Limitation of Liability , (2) Indemnification, (3) Affirmative Covenants, (4) Restrictive Covenants, (5) Term & Termination, (6) Governing Law, (7) Liquidated Damages, (8) Third-party Beneficiary, and (9) IP Ownership/Licence.

For each category, practicing lawyers authored 114 natural language queries that mirror real drafting tasks, yielding a "Clause Corpus" of candidate provisions. Exhaustive pairing of these clauses with the queries produced approximately 126,000 query instances, each graded on a five-point relevance scale (1 = irrelevant … 5 = exemplary). The annotation effort, carried out by 12 attorneys and 10 trained students and reconciled by a senior lawyer panel, would be highly resource intensive, emphasizing the dataset's depth and reliability. Ultimately, we reconstruct a unified working table by concatenating the three official "qrels" files (train.tsv, valid.tsv, test.tsv) and merging them with the clause corpus to form a single Pandas DataFrame, containing all 126.659 query-clause pairs.

Data Preprocessing

Before starting with the main text preprocessing we did a basic one to prepare the legal contract clause for topic modelling using Latent Dirichlet Allocation (LDA). This initial phase did not include the lawyer-annotated relevance scores.

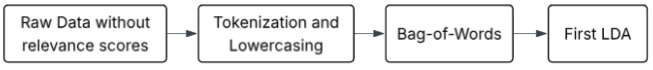


Figure 1: Overview of data preparation.

We start with a simple tokenization and lowercasing, using the genism package to tokenize each clause, remove accent marks, and convert to lowercase (Řehůřek & Sojka, 2010). This process follows standard normalization practices in NLP to reduce lexical variability and ensure consistent input for downstream models (Manning et al., 2008). The resulting tokenized text was then used to construct a term-frequency dictionary and transformed into a Bag-of-Words (BoW) representation. LDA was then applied to the BoW data to extract latent topics. However, the results were semantically vague and lacked interpretability. Due to these limitations, we transitioned to enhanced modeling strategies and a more robust preprocessing workflow.

To support subsequent modeling stages, we constructed a unified DataFrame by merging corpus identifiers, clause texts, query identifiers, and their associated relevance scores. We opted for a streamlined representation where the query identifiers were not separated from the query text itself, given that each query in our dataset is unique the query text itself functions as a unique identifier, simplifying indexing and model input formatting. After forming the base DataFrame, we further enriched it by incorporating query categories, as defined in the original ACORD paper (Wang et al., 2025). Including these labels allows us to conduct category-specific evaluation of retrieval performance-providing insights into how well each model performs across distinct legal clause types.

The final DataFrame includes: (1) the query-identifier (which as explained is also the text); (2) the clause identifier; (3) the expert-assigned relevance score; (4) the clause text; and (5) the category of the query.

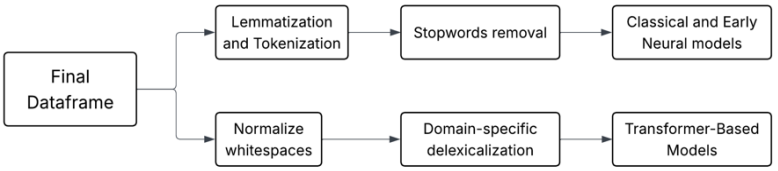


Figure 2: Pre-processing workflow

We thus start with the lemmatization using the WordNetLemmatizer from the NLTK library (Bird et al., 2009). We decided to use lemmatization instead of stemming as it preserves the base or dictionary form of words while maintaining syntactic validity–making it more suitable for downstream NLP tasks that require semantic precision (Manning et al., 2008). Afterwards we

reapplied tokenization and proceeded to remove stopwords. Each token was checked against the genism STOPWORDS list, and any match was excluded.

The preprocessing steps outlined above were applied consistently across all models up to and including the classical approaches and early neural models such as LDA and BM25. However, for transformer-based architectures such as MiniLM, SBERT, and Legal-BERT, we adopted a different preprocessing strategy. These models rely on contextualized embeddings and are typically pre-trained on raw or lightly processed text. To preserve alignment with their pre-training conditions, we bypassed aggressive normalization and instead applied only minimal preprocessing. We thus only normalize whitespaces by stripping them, and convert typographic (smart) quotes and other non-standard Unicode characters to their ASCII equivalents to ensure encoding consistency.

The last cleaning applied is the substitution of all references to other legal documents. These references–often introduced by patterns such as "§" followed by digits–we programmatically identified using regular expressions. To avoid misleading the models into learning superficial correlations between reference numbers and clause content, we replaced all such patterns with the placeholder token SECREF. This approach is similar to delexicalization strategies used in task-oriented dialogue systems and sensitive-domain NLP, where placeholders replace structured tokens to improve model generalization (Sharma et al., 2017).

Models

In this paper, we evaluate the performance of several retrieval techniques. Given the great variety of approaches, the methods introduced can be divided into simple models, advanced models and rerankers. These approaches range from basic lexical matching to advanced neural architectures. Across most approaches, cosine similarity is used as the standard measure to quantify the relevance between queries and documents, following their transformation into appropriate vector representations. In the following sections, we briefly describe each retrieval technique evaluated in this study.

Keyword Matching is one of the simplest information retrieval techniques. It involves quantifying the exact lexical overlap between a user query and a document by counting the number of shared terms. For each query-document pair, the number of overlapping keywords is counted, with a higher count indicating a greater presumed relevance

TF-IDF (Term frequency-inverse document frequency) is a statistical measure used to evaluate the importance of a word in a document relative to a corpus (Salton & Buckley, 1987;

Robertson, 2004). The technique boosts terms that are frequent in a document but infrequent across the corpus, helping to emphasize distinctive content over common words.

BM25 (Best matching 25) is a ranking function used to estimate the relevance of a document to a query. BM25 score increases with the frequency of query terms in the document while mitigating the impact of prevalent words (Robertson & Zaragoza, 2009).

LDA is a generative probabilistic model of a corpus. It assumes that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words (Blei et al., 2003). During inference, LDA uncovers these hidden topic structures by estimating two sets of probability distributions: the distribution of topics within each document and the distribution of words within each topic. 9 topics were given as input for LDA topic modelling, to reflect the number of queries' categories and the results were visualized using pyLDAvis. Additionally, the LDA-based similarity between documents and queries was computed using the cosine similarity of their respective topic distribution. Both results are stored in the columns *lda_topic* and *lda_score*.

NMF (Non-negative matrix factorization) is the second technique used for topic modelling. Unlike LDA, which uses a probabilistic model, NMF relies on algebraic decomposition, thus offering a different and non-probabilistic approach (Choo et al., 2013). NMF factorizes a term-document matrix, in this case derived from TF-IDF, into two lower-dimensional non-negative matrices (Lee & Seung, 2000). The first one representing the document-topic distributions and the other the topic-word distributions.

Word2Vec. To classify documents based on their semantic content, we employed a technique that combines pre-trained word embeddings with a logistic regression classifier (Egger & Yu, 2022). By using the pre-trained Word2Vec model trained on part of the Google News dataset, an average vector was constructed for each document. The resulting vectors were then used as input features to train a logistic regression model for a multi-class classification task. The training used an 80-20 stratified train-test split and a balanced class weight to mitigate the negative effects of an unbalanced dataset.

MiniLM. With the term MiniLM we refer to the "All-MiniLM-L6-v2" model developed by Microsoft. Compared to BERT, MiniLM has a reduced number of layers and hidden units, making the model lighter and faster, while maintaining a comparable performance (Wang et al., 2020). The model was used to generate dense vector embeddings for both queries and documents, allowing for more fine-grained comparison than traditional approaches.

LEGAL-BERT is a family of BERT models for the legal domain. Being pre-trained on general domain text (e.g. Wikipedia), BERT underperforms in specialized domains (Lee et al., 2019). In fact, legal text has peculiar characteristics compared to generic corpora, such as specialised vocabulary and particularly formal syntax. Given the nature of our dataset, using a legal-specialized model was a justified choice to better capture domain-specific semantics and improve performance.

SBERT is a modification of the BERT network designed to produce semantically meaningful sentence embeddings. Traditional BERT uses a cross-encoder, and thus requires two sentences to be input simultaneously, making it inefficient for large-scale similarity comparisons. By allowing independent sentence encoding, SBERT significantly reduces computational overhead while maintaining or improving the performance (Reimers & Gurevych, 2019).

MS MARCO Reranker. Rerankers are models designed to rescore and reorder a small set of candidate documents to obtain better results. By working on a smaller set of documents, rerankers can evaluate query-document pairs more deeply and with better accuracy. It is important to mention that rerankers do not rely on the original scores, instead they independently assign a new score to each pair. Two rerankers are used in the project: MS MARCO and GPT4.1.

To enhance retrieval precision, we employed a cross-encoder re-ranker based on the pre-trained MS MARCO MiniLM-L-6-v2 model (Nguyen et al., 2016). Unlike traditional retrievers that embed queries and documents independently, the cross-encoder jointly encodes each query-document pair, making it perfectly suited for our dataset structure (Bajaj et al., 2016). Reranker input is selected by taking the top 50 documents for each query based on different scores: the BM25, MiniLM, LEGAL-BERT, and SBERT. The model outputs an independent relevance score for each pair, and the results are stored in columns named similarly to the input column.

GPT 4.1 Reranker. The second reranker works using OpenAI's GPT-4.1-nano, a lightweight variant of the GPT-4 architecture. The procedure is similar to the one used for the MS MARCO reranker, that is we selected as input each query and the top 50 documents for each query and they were included in a single prompt. The prompt instructed the model to rate each document's relevance to the query on a scale from 1 (not relevant) to 5 (highly relevant), mimicking lawyers' scoring task.

Since we accessed the model through an application programming interface (api), we were subject to OpenAI's usage cost and limits on use. As a result, a thoughtful optimization of resources was necessary. By including multiple documents in a single prompt, we were able to reduce the total token usage and the number of prompts per minute. In addition, query and document text were truncated to control total token usage; while error handling and retry logic were implemented to address rate limits and occasional failures. This GPT-based reranking strategy allowed us to

incorporate high-level reasoning and contextual understanding into the final ranking phase, offering a great culmination to our techniques pipeline.

Computational Performance and Cost Analysis

All experiments were conducted on a u1-standard-64 virtual machine, equipped with 64 vCPUs (Intel Xeon Gold 6130), 384 GB of RAM, and no GPU. The code takes approximately 45 minutes to run, including the download time for datasets and pretrained models. A noticeable amount of time (9 minutes) is used to compute the MiniLM embeddings. In fact, it is an inefficient process as for each pair both the query and corpus vector are computed every time, despite them being heavily duplicated. However, both Legal-BERT and SBERT are computed efficiently, and despite being more advanced models, they take much less time to run (2 minutes). Nevertheless, we decided to keep the original MiniLM embeddings computation to emphasize the importance of optimization, efficiency and the evolution of our models workflow.

Including GPT's model limitations and cost analysis seems dutiful here. As of May 2025, a newly created OpenAI API PlatformTier 1 account requires a 5$ (+tax) minimum billing deposit. The model selected is GPT-4.1-nano, despite being the "weakest" of the 4.1 family, it is one of the most recent OpenAI's models and has a great cost/performance, completing a reranking task using 1.14M token and just 0,12$. The GPT Reranker takes around 4 minutes to complete its task.

Metrics

We rely on two metrics to evaluate model performance: Star Precision and Normalized Discounted Cumulative Gain (NDCG). Star Precision measures the proportion of documents within the top k retrieved clauses that meet a [human-annotated] relevance threshold, indicated by the "star" rating (e.g., '5-star-precision@5' measures the fraction of the model's top 5 results rated as 5-star by the human annotators). NDCG assesses ranking quality by considering both the position of retrieved clauses in the ranked list (up to k) and their human-annotated relevance scores, assigning higher value to more relevant documents ranked higher. Both metrics provide insight into how effectively the models retrieve and prioritize the most relevant legal clauses for a given query. While NDCG is sensitive to the relative order of relevance within the ranked list, Star Precision attributes credit based solely on the count of retrieved clauses meeting the specified score (star-rating) within the top k, irrespective of their precise order within those top results

These metrics are applied within our two-stage retrieval pipeline. First, an initial retrieval model (e.g., BM25) scores all query-clause pairs from our dataset, totaling over 126,000 initial scores.

For each query, this model's top 50 scoring clauses are then passed as candidates to a second-stage reranking model (e.g., MS-MARCO or GPT-4.1-nano), which re-evaluates and re-scores only this subset. The final Star Precision and NDCG metrics are then computed on the top 5 or 10 results of the list generated by the reranker (or the single-stage model), using the human lawyer scores as ground truth. Critically, when averaging these metrics across categories, a category is excluded from the calculation if the set of documents processed by the reranker contains fewer than two items that were also marked as relevant (human score > 0) by human lawyers. This approach means that the reported overall performance of a reranker, and consequently the potentially skewed average metrics, is highly dependent on the initial retriever's success in finding a sufficient number of relevant clauses within its top 50 candidates.

Results

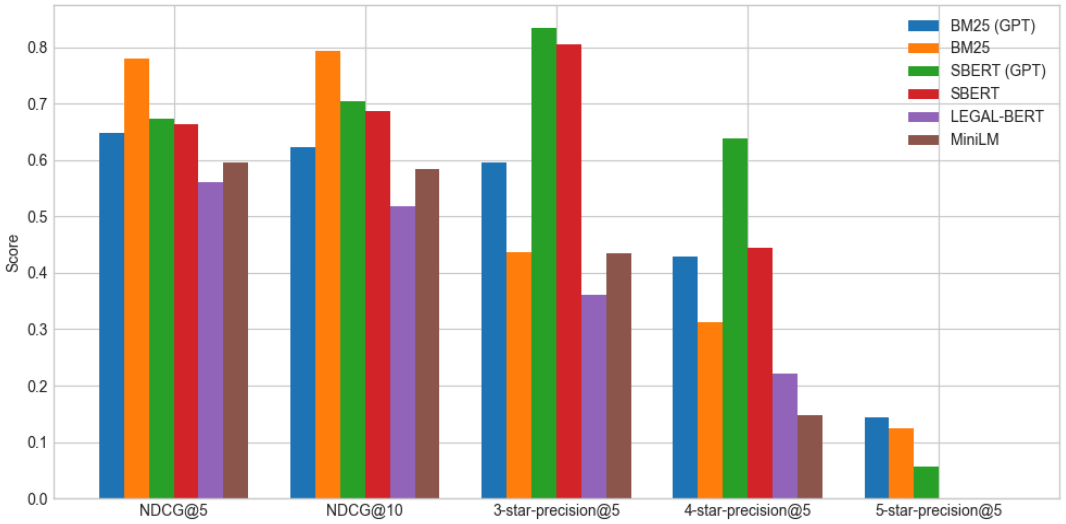


Figure 3: Scores ranging across the 5 selected metrics

When reranked with ms-marco BM25 stands out with best NDCG-scores (@5 of 0.781 and @10 of 0.794) and only non-zero 5-star-precision@5 score (0.125). GPT reranked SBERT is the second best performing model with NDCG scores of 0.674 and 0.705 respectively. Compared to BM25, SBERT has superior 4- and 3-star-precision@5 (0.639 and 0.833 respectively), showing that it is better at catching the general query-clause relevance (see appendix A for full results table).

Despite obtaining the highest NDCG results, the MS MARCO reranked BM25 needs to be carefully scrutinized. In fact, the model is completely unbalanced as evaluation of it only considers its ability in 4 out of the 9 categories due to limited data points. This is partially the result of the original BM25 scores, which assigns score zero on all metrics to certain queries, thus making the reranker results biased. For example, the query "indemnity of broad-based claims" achieved perfect scores, whereas "as-is clause" got zero NDCG@5 score. A fairer comparison would be comparing BM25 and SBERT on the same GPT reranker. When reranking the two top-performing models with GPT-4.1-nano, the BM25 was inferior to SBERT at nearly all metrics. In view of these considerations, and considering SBERT's superior ability at retrieving relevant clauses (according to lawyers' definition), GPT reranked SBERT can be considered the best overall model.

Discussion

Traditional lexical matching remains a reliable launch-pad, quickly recalling many on-topic clauses, but begins to falter when finer conceptual or stylistic nuance is required. Introducing dense encoders narrows that gap, signaling that semantic representations trained on large, mixed-domain text already generalize surprisingly well to contract language. The most substantial uplift, however, arrives when a modern LLM reranker is interposed as a second stage. Point-wise cross-encoders consistently reorder the candidate list so that clauses lawyers deem "ready to paste" rise to the top. Yet even with this two-stage pipeline, a residual quality gap persists. The very best clauses appear less often than practicing attorneys would consider reliable for LLM-mediated drafting. Taken together, the findings validate retrieval-augmented drafting as a credible efficiency lever while also underscoring the need for targeted domain adaptation and cost-aware model design before such systems can be relied upon in high-stakes contract negotiation.

Several design choices highlight why LLM reranking delivers the sharpest gains yet still falls short of "plug-and-play" reliability. First, ACORD's five-point rubric is inherently ordinal. Point-wise rerankers, optimized to predict that very grade, therefore align naturally with the annotation scheme, whereas pair-wise objectives can misallocate learning capacity to swaps among mid-tier clauses. Second, ACORD clauses are long (median 146 words) and densely cross-referenced. Cross-encoder architectures such as GPT 4.1 nano compute full token-level attention between query and clause. Third, the lawyer-centric precision metrics expose a residual "last-mile" gap. Even when normalized ranking scores look strong, genuinely paste-ready clauses remain scarce. This echoes recent findings

in legal question answering, where high accuracy masks downstream editing burden (Niklaus et al., 2023).

The performance hierarchy we observe has direct consequences for how IR technology can be embedded in day-to-day contracting. A purely lexical stack offers the speed required for interactive search, but its limited precision obliges lawyers to read far more candidate language than is commercially adequate. By appending a point-wise cross-encoder reranker, implemented here with the lightweight GPT-4.1 nano model, we raise the likelihood of surfacing a four- or five-star clause in the top five results.

Our internal timing study therefore places the combined pipeline well within the sub-second thresholds typical of contract-lifecycle management (CLM) interfaces, yet at a fraction of the energy cost associated with larger LLMs. On the user side, evidence suggests that when high-quality precedent appears this early in the ranked list, lawyers can shorten review and red-lining cycles by 30-40% (Bommarito & Katz, 2022). Nevertheless, the modest incidence of exemplary five-star clauses means that human vetting remains indispensable. The technology currently serves best as a precision-boosted triage layer rather than an autonomous drafting engine. Designing interfaces that expose the model's relevance score and clause category will therefore be as critical as further algorithmic refinements.

Limitations

Notwithstanding these encouraging throughput figures, several constraints temper the generalizability of our results. Dataset scope is the first limiting factor since ACORD covers only nine clause genres drawn largely from US-law, English-language agreements, omitting financially pivotal sections such as representations-and-warranties or payment terms. Clause quality assessments therefore reflect a narrow slice of global contracting practice and may not transfer to other jurisdictions. In addition, although ACORD's five-point scale is richer than binary labels, relevance remains subjective; inter-annotator disagreement reaches 21%, and the rubric does not explicitly encode stylistic compatibility (e.g., UK versus US drafting conventions). Finally, our project methodology isolates retrieval logic but forgoes the potential gains of domain fine-tuning.

Conclusion & Future Inquiry

Commercial contracting still expends disproportionate time and money on locating precedent language, a problem aggravated by the absence of clause-level retrieval benchmarks that mirror legal drafting practice. To address this gap, we adopted the ACORD corpus and reframed it into a coherent dataset appropriate for evaluating retrieval pipelines. Within this unified approach we compared three tiers of technology: (a) classical lexical ranking; (b) semantic bi-encoders; and (c) lightweight cross-encoders, used purely for second-stage reranking. This design allowed us to ask, in practical rather than algorithmic terms, to what extent modern IR stacks narrow the gap between the clauses lawyers need and the clauses they can actually find during a drafting session.

Our evaluation traces a consistent arc. BM25 delivers fast recall but raises four- or five-star clauses for barely 40% of queries, showing lexical overlap alone is too coarse. Furthermore, semantic bi-encoders (MiniLM, SBERT) close roughly half that gap, almost doubling graded scores without hurting latency. The decisive lift, however, comes from the GPT-4.1 nano reranker, which greatly improves relevant clause retrieval. Yet exemplary five-star provisions surface only 5% of the time; a ceiling partly set by the data itself, as only 130 five-star label pairs are present in the entire ACORD dataset. Consequently, today's best pipelines markedly reduce search friction but cannot eliminate the lawyer's screening burden.

Looking ahead, bridging the residual gap between "good enough" retrieval and lawyer-ready output turns on three fronts of which future research could inquire into. On the model front, compact, domain-tuned cross-encoders and ordinal listwise objectives could raise graded precision without the cost of a full scale large language model. On the data front, enlarging ACORD beyond its nine US-law clause types to include representations and warranties, payment terms, and agreements in other languages would lift the ceiling on "paste ready" retrieval and give a broader test of generality. Equally important are human factor studies that embed these pipelines in contract lifecycle software, measure real drafting time savings, and examine interface cues such as relevance grades and clause category tags that help users calibrate trust.

In sum, retrieval augmented drafting already offers a credible efficiency boost: modern stacks surface high quality precedent in milliseconds at modest cost. Nevertheless, full autonomy in drafting remains elusive, underscoring the complementary rather than substitutive role NLP currently plays in legal workflows. Bridging this "last mile" gap through targeted model refinement, richer legal signals, and broader clause corpora sets a clear agenda for future research in this field.

References

Bachman, J. (2024, August 8). AI's first serious foray into legal may be contract review. Legal Dive. https://www.legaldive.com/news/ais-first-serious-foray-legal-contract-review/723775/

Bajaj, P., Campos, D., Craswell, N., Deng, L., Gao, J., Liu, X., Majumder, R., McNamara, A., Mitra, B., Nguyen, T., Rosenberg, M., Song, X., Stoica, A., Tiwary, S., & Wang, T. (2016). MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1611.09268

Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: Analyzing text with the natural language toolkit. O'Reilly Media. https://www.nltk.org/book/

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022. https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

Bommarito, M. J., & Katz, D. M. (2022). GPT takes the bar exam. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4314839

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., & Androutsopoulos, I. (2020). LEGAL-BERT: The Muppets straight out of Law School. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2010.02559

Chalkidis, I., Jana, A., Hartung, D., Bommarito, M., Androutsopoulos, I., Katz, D. M., & Aletras, N. (2021). LexGLUE: a benchmark dataset for legal language understanding in English. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2110.00976

Choo, J., Lee, C., Reddy, C. K., & Park, H. (2013). Utopian: User-driven topic modeling based on interactive nonnegative matrix factorization. IEEE transactions on visualization and computer graphics, 19(12), 1992-2001. https://www.researchgate.net/profile/Chandan-Reddy-2/publication/256837226_UTOPIAN_User-Driven_Topic_Modeling_Based_on_Interactive_Nonnegative_Matrix_Factorization/links/55b8b0d108aed621de067d2d/UTOPIAN-User-Driven-Topic-Modeling-Based-on-Interactive-Nonnegative-Matrix-Factorization.pdf

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1810.04805

Egger, R., Yu, J. (2022, May). A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. Front. Sociol., 06 May 2022. Sec. Sociological Theory, Volume 7 – 2022. https://doi.org/10.3389/fsoc.2022.886498

Goebel, R., Kano, Y., Kim, M., Rabelo, J., Satoh, K., & Yoshioka, M. (2024). Overview and discussion of the Competition on Legal Information, Extraction/Entailment (COLIEE) 2023. The Review of Socionetwork Strategies, 18(1), 27–47. https://doi.org/10.1007/s12626-023-00152-0

Hendrycks, D., Burns, C., Chen, A., & Ball, S. (2021). CUAD: An Expert-Annotated NLP Dataset for Legal Contract Review. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2103.06268

Johnston, P. (2025, April 2). Retrieval-augmented generation (RAG): towards a promising LLM architecture for legal work? (K. Raditya & P. Sayed, Eds.). Harvard Journal of Law & Technology. https://jolt.law.harvard.edu/digest/retrieval-augmented-generation-rag-towards-a-promising-llm-architecture-for-legal-work

Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. (2020). Dense passage retrieval for Open-Domain question answering. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2004.04906

Lee, D., & Seung, H. S. (2000). Algorithms for non-negative matrix factorization. Advances in neural information processing systems, 13. https://proceedings.neurips.cc/paper_files/paper/2000/file/f9d1152547c0bde01830b7e8bd60024c-Paper.pdf

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), 1234-1240. https://arxiv.org/pdf/1901.08746

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP tasks. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2005.11401

Mamakas, D., Tsotsi, P., Androutsopoulos, I., & Chalkidis, I. (2022). Processing Long Legal Documents with Pre-trained Transformers: Modding LegalBERT and Longformer. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2211.00974

Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. Cambridge University Press. https://nlp.stanford.edu/IR-book/

Menear, H. (2024, November 12). Organisations “lose millions” in revenue thanks to “poor contract negotiation.” CPOstrategy. https://cpostrategy.media/blog/2024/11/12/organisations-lose-millions-in-revenue-thanks-to-poor-contract-negotiation/

Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., & Deng, L. (2016). Ms marco: A human-generated machine reading comprehension dataset. https://openreview.net/pdf?id=Hk1iOLcle

Niklaus, J., Matoshi, V., Rani, P., Galassi, A., Stürmer, M., & Chalkidis, I. (2023). LEXTREME: A Multi-Lingual and Multi-Task Benchmark for the Legal Domain. ACL Anthology. https://doi.org/10.18653/v1/2023.findings-emnlp.200

Řehůřek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks (pp. 45–50). European Language Resources Association. https://www.fi.muni.cz/usr/sojka/papers/lrec2010-rehurek-sojka.pdf

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 3982–3992). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410ACL Anthology+3ACL Anthology+3SCIRP+3

Robertson, S. (2004). Understanding inverse document frequency: on theoretical arguments for IDF. Journal of Documentation, Vol. 60 No. 5, pp. 503-520. https://doi.org/10.1108/00220410410560582

Robertson, S., & Zaragoza, H. (2009). The Probabilistic Relevance Framework: BM25 and beyond. Foundations and Trends® in Information Retrieval, 3(4), 333–389. https://doi.org/10.1561/1500000019

Robertson, S., Zaragoza, H., & Taylor, M. (2004, November). Simple BM25 extension to multiple weighted fields. In Proceedings of the thirteenth ACM international conference on Information and knowledge management (pp. 42-49). https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=10b2752c9940e9c761a161a0fef7ae08cae66301

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information processing & management, 24(5), 513-523. https://ecommons.cornell.edu/server/api/core/bitstreams/fc18789c-6a03-48e6-8226-7dba0ce94e32/content

Sharma, S., He, J., Suleman, K., Schulz, H., & Bachman, P. (2017). Natural Language Generation in Dialogue using Lexicalized and Delexicalized Data [Workshop paper]. In Workshop Track – ICLR 2017. https://arxiv.org/abs/1606.03632

Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., & Gurevych, I. (2021). BEIR: a heterogenous benchmark for zero-shot evaluation of information retrieval models. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2104.08663

Wang, S. H., Zubkov, M., Fan, K., Harrell, S., Sun, Y., Chen, W., Plesner, A., & Wattenhofer, R. (2025). ACORD: An Expert-Annotated Retrieval Dataset for Legal Contract Drafting. Cornell University. https://doi.org/10.48550/arXiv.2501.06582

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., & Zhou, M. (2020). Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. Advances in neural information processing systems, 33, 5776-5788. https://arxiv.org/pdf/2002.10957

World Commerce & Contracting. (2023). CCM: the journey to operational excellence. In World Commerce & Contracting. Icertis. https://www.worldcc.com/Portals/IACCM/Reports/Benchmark-report-2023.pdf

Appendices

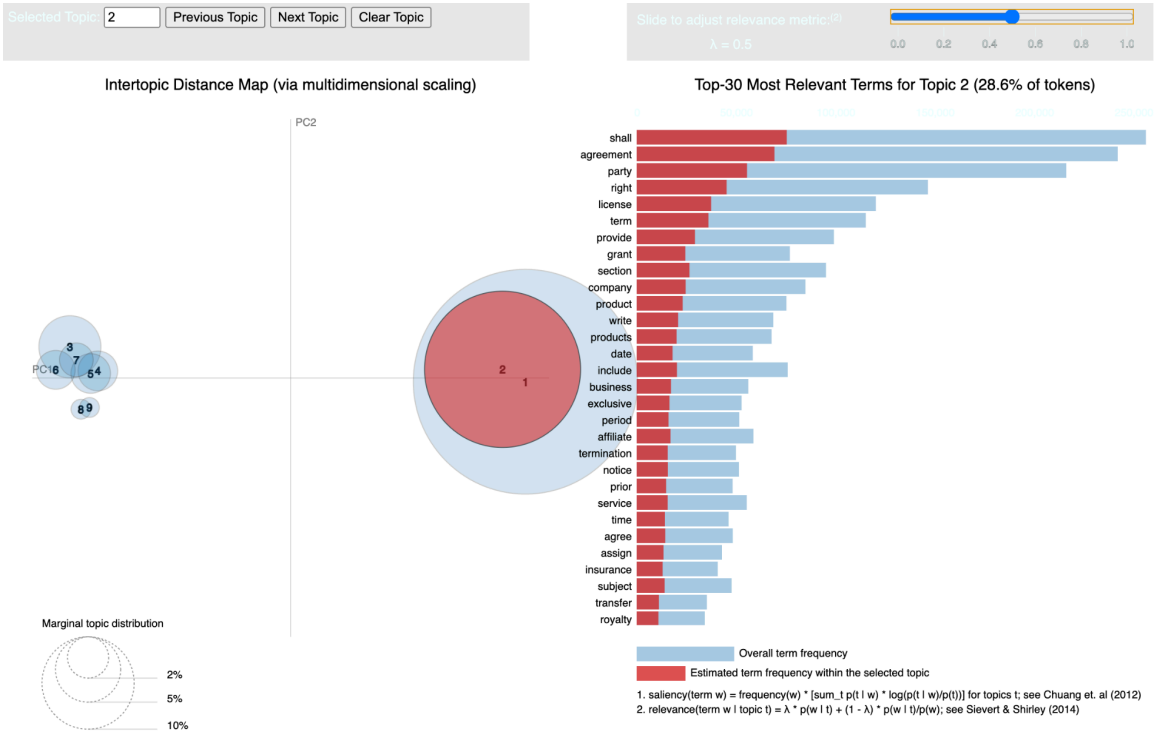
A: Results

Model	Reranker	NDCG@5	NDCG@10	3-Star Precision@5	4-Star Precision@5	5-Star Precision@5
TF-IDF	None	0,468	0,488			
LDA	None	0,539	0,553			
BM25	None	0,475	0,494			
	MS MARCO	0,781	0,794	0,437	0,312	0,125
	GPT4.1-nano	0,647	0,623	0,595	0,429	0,125
MiniLM	None	0,453	0,467			
	MS MARCO	0,597	0,583	0,435	0,148	0
Legal-BERT	None	0,462	0,463			
	MS MARCO	0,561	0,518	0,361	0,222	0
SBERT	None	0,686	0,673			
	MS MARCO	0,663	0,687	0,806	0,444	0
	GPT4.1-nano	0,674	0,705	0,833	0,639	0,056

B: *merged_df* Dataframe

	query-id	corpus-id	score	text	category
0	"as is" clause with carveouts	e279fd792f	3	Section 9.3 Disclaimer. EXCEPT AS EXPRESSLY SE...	Limitation of Liability
1	"as is" clause with carveouts	4f06e1c658	3	12.1 Disclaimer of Warranties. THE WARRANTIES ...	Limitation of Liability
2	"as is" clause with carveouts	adec05476d	1	WARRANTY DISCLAIMER\nTHE LICENSED TECHNOLOGY I...	Limitation of Liability
3	"as is" clause with carveouts	494c10db01	3	Disclaimer of Warranties. UNLESS SPECIFIED OTH...	Limitation of Liability
4	"as is" clause with carveouts	e282637c4e	3	NO WARRANTIES.\nExcept as expressly set forth ...	Limitation of Liability

C: Intertopic distance of predefined categories lexical attributes



D: Intertopic distance of LDA defined categories lexical attributes