CLERC: A Dataset for U. S. Legal Case Retrieval and Retrieval-Augmented Analysis Generation

Anonymous ACL submission

Abstract

Legal professionals need to write analyses that rely on citations to relevant precedents, i.e., previous case decisions. Intelligence systems assisting legal professionals in writing such documents provide great benefits but are challenging to design. Such systems need to help locate, summarize, and reason over salient precedents in order to be useful. To enable systems for such tasks, we work with legal professionals to create a colossal dataset[1] supporting two important backbone tasks: information retrieval (IR) and retrieval-augmented generation (RAG). This dataset CLERC (**C**ase **L**aw **E**valuation and **R**etrieval **C**orpus), is constructed for training and evaluating models on their ability to (1) find corresponding citations for a given piece of legal analysis and to (2) compile the text of these citations (as well as previous context) into a cogent analysis that supports a reasoning goal. We benchmark state-of-the-art models on CLERC, showing that current approaches still struggle: GPT-4o generates analyses with the highest ROUGE F-scores but hallucinates the most, while zero-shot IR models only achieve 48.3% recall@1000.

1 Introduction

Writing legal analyses is a central task of lawyers. For example, in the U.S. legal system, litigation attorneys draft motions and briefs, citing relevant cases (case law) to support their analysis. From a corpus of millions of cases, finding the relevant files and weaving them into a persuasive whole requires substantial time and effort. Enhancing lawyers' capacity to retrieve relevant cases and generate legal analyses presents difficult challenges across various fields, including information retrieval, natural language processing, and the broader AI community.

To support the development of information retrieval and generative language models in writing legal documents, we build a dataset of case law, i.e., the *precedents*. Case law is written by judges to explain the court's reasoning for its resolution of legal disputes based on concrete facts and laws. It is foundational for legal analyses for Common Law and some Civil Law jurisdictions (Ma et al., 2021; Mahari et al., 2023). Curating a quality case law dataset is valuable for two main reasons. First, given that case law is carefully written and serves as a strong prediction of future case law, the data is valuable for developing legal intelligence systems. Second, given the desire for retrieval accuracy and text generation quality, high-quality training data is critical for producing effective models.

Thus, we build a dataset, CLERC (**C**ase **L**aw **E**valuation and **R**etrieval **C**orpus), as a step towards enabling collaborative AI systems that assist legal professionals, while also providing AI researchers with a challenging real-world evaluation set. Compared with existing precedent datasets (Henderson et al., 2022; Santosh et al., 2024; Mahari et al., 2023; Ma et al., 2021; Guha et al., 2023), we offer the following novel contributions:

1. We build the largest (25.5 millions of instances in total) Legal IR and RAG dataset using U.S. case law data, and a pipeline for converting raw data into our task formats, built with advice and knowledge from legal experts.

2. Through working with legal professionals, we provide a formulation of legal case retrieval and analysis generation tasks that balance the needs and perspectives of legal professionals with computational feasibility.

3. We conduct a comprehensive evaluation of case retrieval and retrieval-augmented legal analysis generation on CLERC with state-of-the-art methods, revealing that IR models struggle to retrieve relevant documents and LLMs frequently hallucinate.

[1] Code/data is publicly available at `anonymous link`.

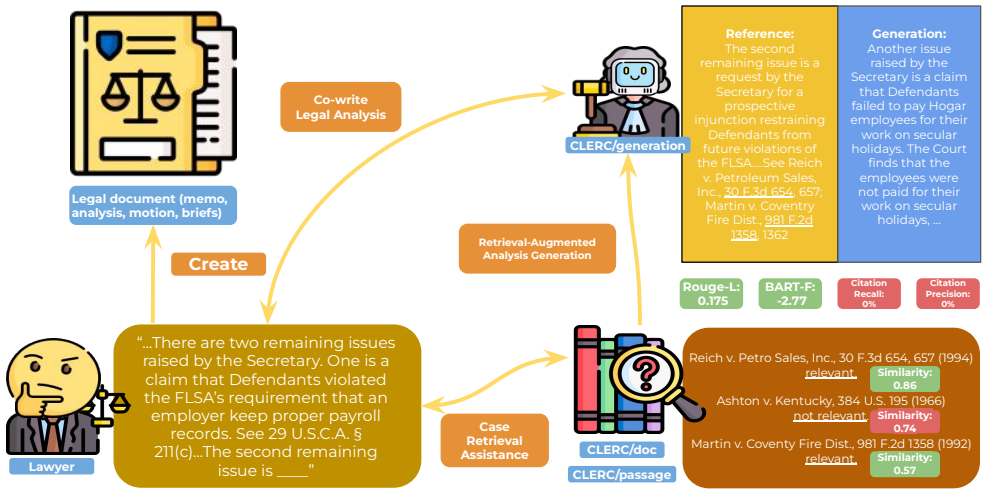

Figure 1: An overview of how CLERC enables systems that empower legal professionals: retrieval models for `CLERC/doc` and `CLERC/passage` that assist with finding relevant cases to support the analysis, and retrieval-augmented generation systems with `CLERC/generation` that aid in legal analysis generation.

Query: ...Summary judgment should be granted where "the pleadings, depositions, answers to interrogatories and admissions on file, together with the affidavits, if any, show there is no genuine issue as to any material fact and that the moving party is entitled to judgment as a matter of law." Fed.R.Civ.P. 56(c). The moving party has the responsibility of informing the Court of portions of the record or affidavits that demonstrate the absence of a triable issue. Celotex Corp. v. Catrett, 477 U.S. 317, 322, 106 S.Ct. 2548, 91 L.Ed.2d 265 (1986). The moving party may meet its burden of showing an absence of disputed material facts by demonstrating *"that there is an absence of evidence to support the non-moving party's case."* `Id. at 325, 106 S.Ct. 2548`. Any doubt as to the existence of a genuine issue for trial is resolved against the moving party. `Anderson v. Liberty Lobby, Inc., 477 U.S. 242, 255, 106 S.Ct. 2505, 91 L.Ed.2d 202 (1986)`;...

Figure 2: A depiction of a query from CLERC. "Fed.R.Civ.P. 56(c)" cites a federal statute. Underlined violet is the query's **central citation sentence**. *The brown italicized sentence* is a **direct quote** from Celotex Corp. v. Catrett, making this query a *direct query*. `The red citations in teletype fonts` are *non-central citations*. We define `single-removed` as a data view which masks only the central citation and its salient sentence, while `all-removed` masks all citations from the passage. All references to statutes are retained.

2 Related Work

In this section, we provide an overview of long-context retrieval and RAG, including existing applications in the legal domain.

2.1 Legal Case Retrieval and Generation

Case law retrieval is a longstanding problem aimed at helping legal experts find cases to cite (Wilson, 1962). Techniques range from lexical search (Blair and Maron, 1985) to more recent semantic similarity-based approaches (Ebietomere and Ekuobase, 2019; Askari et al., 2021). Overall, most legal retrieval is done through WestLaw[2] or LexisNexis,[3] or services offered by Thomson Reuters,[4] which charge large sums for access to a manually annotated and custom-based ontology. Although many new datasets exist for countries outside of the United States (Ma et al., 2021; Li et al., 2024; Santosh et al., 2024; Goebel et al., 2024), **few existed in the United States** before the large U.S. case law resource, Caselaw Access Project (CAP), was released to the public in 2023 (Mahari, 2021). A summary of more related work of legal IR can also be found in Locke and Zuccon (2022). Due to the recent advances of Large Language Models (LLMs) and RAG, legal generation is a new field with preliminary works, including analysis generation using small datasets or small-scale models (Tuvey and Sen, 2023; Li and Zhang, 2021; Lin and Cheng, 2024), generating stories to educate about legal concepts (Jiang et al., 2024), generating legal summaries (Gesnouin et al., 2024; Shen et al., 2022), and profiling hallucinations in legal generation (Dahl et al., 2024; Magesh et al., 2024). In comparison, we provide an extensive benchmark on **retrieving citations and generating long-form legal analysis**, which is a core task for lawyers that requires substantial time and effort.

[2] https://legal.thomsonreuters.com/en/westlaw
[3] https://www.lexisnexis.com/
[4] https://legal.thomsonreuters.com/en/products/court-express

2.2 Long-Context Generation

LLMs are limited in context length due to concerns of efficiency and performance (Li et al., 2023). Benchmarks have been introduced to measure the effectiveness of LLMs, including Shaham et al. (2022); Tay et al. (2020); Dong et al. (2024). Various methods were proposed to solve this problem, such as efficient attention (Beltagy et al., 2020; Choromanski et al., 2021; Qin et al., 2024a, inter alia), recurrence (Dai et al., 2019; Bulatov et al., 2022), context length generalization (Su et al., 2024; Chen et al., 2023), and RAG (Lewis et al., 2020), while some recent work challenges the effectiveness of those methods (Liu et al., 2024; Qin et al., 2023). Recent proprietary LLMs are reported to extend the context length up to 1 million tokens, such as Gemini 1.5 (Reid et al., 2024), Claude 3 (Anthropic, 2024), and GPT-4 (Josh Achiam et al., 2024). More specific to legal texts, generative models on long contexts are commonly used for legal case outcome prediction (Qin et al., 2024b; Cao et al., 2024), but long-context generation datasets on legal cases are largely absent from the literature.

2.3 Long-Document Retrieval

Many state-of-the-art neural retrieval models, because of their reliance on BERT models (Devlin et al., 2018), have a context limit of 512 tokens (Khattab and Zaharia, 2020; Santhanam et al., 2022; Ren et al., 2021; Karpukhin et al., 2020a). Accompanied by the advances in long-context LLMs (Brown et al., 2020; Touvron et al., 2023; Josh Achiam et al., 2024), there are also rising interests in long-context retrieval models (Fu et al., 2023; Günther et al., 2023; Wang et al., 2023). However, existing benchmarks lack insights on how both BERT-based and long-context models perform on well-defined legal retrieval tasks (Thakur et al., 2021; Saad-Falcon et al., 2024; Shen et al., 2022).

3 Dataset Construction

3.1 Dataset Preprocessing

CLERC contains a collection of long case documents (`CLERC/doc`) and chunked documents for retrieval (`CLERC/passage`) as well as for generating paragraphs of legal analysis (`CLERC/generation`). We obtain the collection of raw case documents from CAP (CAP, 2024), including all federal case law in U.S. history up to September 21, 2021. To prepare `CLERC/doc`, we concatenate all types of opinions (e.g., majority, dissenting, concurring opinions) in every case document and remove the newline characters. To prepare `CLERC/passage`, we split documents into chunks of 350 words, with a sliding window of 175 words. We discuss the `CLERC/generation` preparation process in Section 3.3 in more detail.

Our final dataset consists of 1.84M documents with over 20.7M total citations, 23.7M passages for retrieval, and 6k passages for generation. Details and summary statistics are presented in Table 1.

3.2 Case Retrieval

Query Definition As lawyers write legal analyses, they need to find case documents to support their texts. To simulate this use case, we define a query as a block of text, taken from a case document, with its citation in the middle (a.k.a the central citation) removed. The *relevant passage* of a query is defined as the passage its central citation cites to. We provide both left and right contexts, as legal analysis involves context from both sides.

Direct and Indirect Queries A case document may cite another case in the form of a direct quote. (Review, 2020). We define two kinds of legal queries based on whether the central citation contains a direct quote (`direct`) or not (`indirect`). Ideally, using *only* this direct quote to retrieve should yield an almost perfect accuracy through an exact string match. We build a heuristic program to recognize such direct quotes (detailed in Appendix C.1), and classify queries into direct and indirect queries. We retrieve with *only* direct quotes in Table 5 and Section C.2.

Data Views We define two data views of legal queries depending on whether all (`all-removed`) or only the central (`single-removed`) citation information is removed. Since each kind of query has two data views, this leads to a total of four categories of queries, illustrated in Figure 2.

Query Construction First, we identify case citations in documents. For every case citation, we expand a window of 300 words centered on this citation to construct a query, as legal analysis uses context from both sides. We experimented with various query lengths and found that including 300 words yields the highest recall (summarized in Figure 4), which indicates that this length provides enough critical information for retrieving relevant cases without distracting the models to retrieve other cases, as we discuss in Section 4.1.2. We

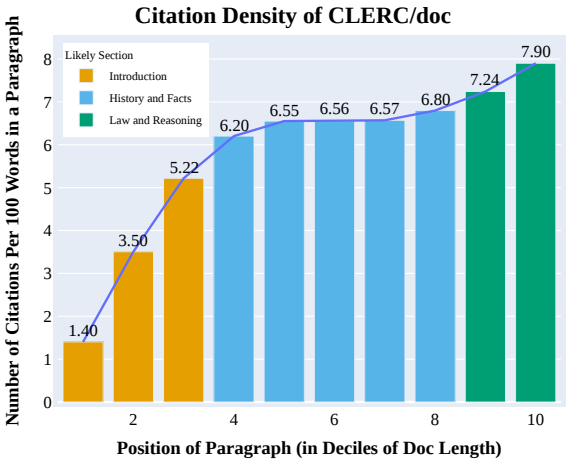

Figure 3: Number of citations per 100 words in a paragraph, arranged by the positions (in deciles) of paragraphs in the document. **Documents tend to cite more often in later paragraphs.** Case law typically has different sections (e.g. Intro, Facts, Reasoning, etc). Deciles with the same color code likely belong to the same general sections of case law.

remove the central citation in this paragraph for `single-removed` queries and all case citations in this paragraph for `all-removed` queries.

Query Extraction To identify the case citation in a query, we use `eyecite` (Cushman et al., 2021), a Regex-based parser that recognizes U.S. case citations. We develop a heuristic and rule-based Python program to detect the sentence boundary where the case citation is located. We create the heuristic rules based on the *Bluebook* (Review, 2020) (detailed in C.1), which is the standard citation guide for U.S. legal professionals. We manually evaluate the extraction accuracy on 1000 sentences and discovered that 87.8% of citation sentences can be accurately identified and removed.

3.3 Legal Analysis Generation

We prepare `CLERC/generation` to test the ability of LLMs to generate realistic legal analyses in a similar manner to how human legal experts write case law. In practice, lawyers write a paragraph of legal analysis based on the previously written context and relevant case law to the current analysis. To simulate this, we first select a target analytical paragraph from a case document. Then, we feed in the preceding context of the case and the external citations from the target paragraph as input. Lastly, we prompt the LLM to generate the target paragraph. We evaluate generation quality by comparing how close the generated paragraph is to the reference. This section will elaborate each step.

Dataset Statistic	# instances (M)	Avg len (words)
`CLERC/doc`	1.84	2279
`CLERC/passage`	23.7	350
`CLERC/queries`	0.105	300
CLERC	25.5	N/A
LePaRD	0.375	200 char
ECtHR-PCR	0.015	5530
LeCaRDv2	0.055	4766
`CLERC/generation`	0.006	3371

Table 1: Dataset statistics. CLERC has the largest document and passage combined collection compared to other legal retrieval datasets (Li et al., 2024; Mahari et al., 2023; T.Y.S.S et al., 2024). Additionally, CLERC contributes a novel subset for legal analysis generation.

Analytical Paragraphs Identification We observe that case documents usually start with introducing the basic facts about the cases, and then make multiple legal analyses while citing other relevant cases (see Figure 5) (Lippmann, 2018). To prepare the reference for evaluating analysis generation, we only use paragraphs (blocks of text separated by newline characters) containing legal analyses and discard the rest.

Our rule of thumb to determine if a paragraph is a suitable reference is whether it has a large number of case citations, as the analytical claims are supported by case. In Figure 3, we plot the trend of citation density in case law documents (only the majority opinion), discovering that cases cite more often, which is consistent with our prior observation.

Reference Selection We sample analytical paragraphs from case texts as reference for evaluating analysis generation. As we observe that analysis tends to occur in later paragraphs of a case document, we randomly sample from the last 33.3% of all paragraphs but exclude the last two paragraphs (which are usually short conclusions) and paragraphs with fewer than two case citations.

Input Preparation After we sample the reference paragraph, we take all texts preceding the reference as a part of the input. We concatenate the most salient passages of relevant cases cited by the reference paragraph, whose chunks are selected via RankLlaMa (Ma et al., 2023).

4 Experiments

In this section, we present experiment results of legal citation retrieval (Section 4.1) and legal analysis generation (Section 4.2) with the proposed CLERC.

Type	R@5	R@10	R@100	R@1K	nDCG@10
ISR	**7.9**	**11.8**	**27.6**	48.3	**5.4**
IAR	7.5	11.1	25.1	43.9	5.2
DSR	7.0	11.0	26.7	**48.5**	5.0
DAR	6.8	9.88	24.9	46.3	4.6

Table 2: Retrieval results of BM25 with different types of CLERC queries. I stands for `indirect`, D for `direct`, SR for `single-removed`, and AR for `all-removed`. E.g. ISR refers to `indirect single-removed`. All numbers are in percentage. R@X represents Recall@X. **ISR queries are the easiest queries to retrieve and are used for our main retrieval experiments.**

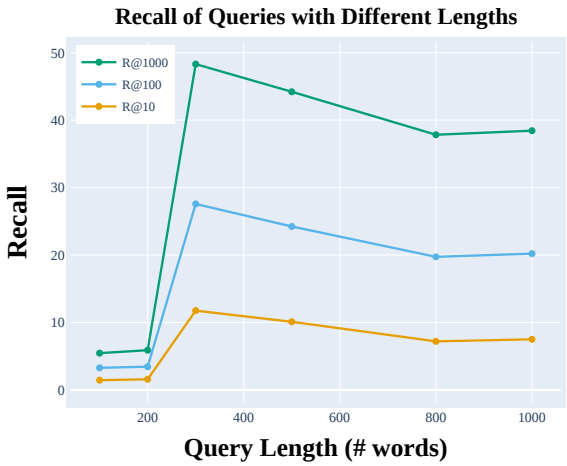

Figure 4: Retrieval results of BM25 with CLERC queries of length 100 to 1000 words. R@X represents Recall@X. **Queries with 300 words maximize recall and are used for our main experiments.**

4.1 Finding Legal Citations

Task Formulation Given a query paragraph q with a central citation sentence s with left and right contexts l and r ($q = l \circ s \circ r$, where $\circ$ indicates string concatenation), retrieve the relevant cited passages or documents $R = \{d_a, ..., d_b\}$ from a collection $D = \{d_1, ...d_N\}$.

4.1.1 Experimental Setup

Metrics and Data We benchmark the performance of popular retrieval methods and models in Table 3 and evaluate with Recall@10, 100, 1000, and nDCG@10. We evaluate with 2851 `indirect single-removed` queries and provide ablation studies on the effect of query types and lengths in Table 2 and Figure 4.

Late Interaction Models We run ColBERTv2 (Santhanam et al., 2022) and long-context Jina-ColBERT (Günther et al., 2023) experiments with the PLAID-X implementation (Yang et al., 2024). Due to the large size of the collection and GPU memory required by ColBERT indexing, we split `CLERC/doc` and `CLERC/passage` into four shards for processing. For ColBERTv2, which has a context limit of 512 tokens,[5] we conduct passage-level retrieval and aggregate the scores by taking the highest-ranking passage of a document to represent the overall document score (MaxP) (Zhang et al., 2021). For long-context ColBERT, we perform and evaluate document-level retrieval and truncate documents over the context lengths (2048, 4096).

Bi-Encoders We also benchmark the performance of state-of-the-art Bi-Encoders, including BGE-large-en-v1.5 (Xiao et al., 2023), E5-large-v2 (Wang et al., 2022), Contriever-MSMarco (Izacard et al., 2021). For DPR models, we adopt the Tevatron pipeline (Gao et al., 2022) and evaluate the base DPR model (Karpukhin et al., 2020b) trained on QA datasets. We also fine-tune BERT-base-uncased[6] (Devlin et al., 2018) and LegalBERT[7] (Holzenberger et al., 2020) on 327K triples from `CLERC/passage` for 10 epochs with a temperature of 0.01 and a learning rate of 10^{-4}, with a contrastive learning-to-rank objective.

Compute Please see Appendix A for the compute setup and hours of the experiments.

4.1.2 Ablation Studies

Query Types We run retrieval experiments with BM25 on all four types of CLERC queries. Table 2 shows that `single-removed` are the easiest queries to retrieve, whereas `all-removed` queries remove non-central case citations and lose additional contextual information contained by them.

Query Length We also evaluate the effect of query length on BM25 across the range of 100 to 1000 words. Figure 4 shows that longer query length improves recall and maximizes it at 300 words. Queries with more than 300 words are harder to retrieve, since as length increases, there are also more distracting contextual information about non-central citations that cause the retriever to find irrelevant passages (see Section 5.1 for examples and qualitative analysis). Another limitation of increased query length is that it will raise the computational cost and storage, which is especially challenging to token-level late-interaction models like ColBERT (Khattab and Zaharia, 2020).

[5] ColBERT generally assumes a max query token limit of 32, which we remove to incorporate our longer queries.

[6] anonymous link, will release after camera-ready

[7] anonymous link, will release after camera-ready

Type	Params	Length	Model	Metrics			
				R@10 ↑	R@100 ↑	R@1K ↑	nDCG@10 ↑
Sparse	N/A	N/A	BM25	11.7	27.8	48.3	5.40
Late Interaction	110M	512	ColBERTv2	2.2	7.1	17.6	1.05
	110M	2048	Jina-ColBERT-v1-en	2.1	5.8	14.2	1.15
	110M	4096	Jina-ColBERT-v1-en	2.2	6.9	16.1	1.19
Bi-Encoders	110M	512	DPR	4.9	13.1	26.2	2.24
	110M	512	BGE En v1.5	7.7	23.2	42.4	3.43
	110M	512	E5 v2	8.4	24.3	42.3	3.67
	110M	512	Contriever-MSMarco	9.3	24.8	41.4	4.22
Fine-Tuned	110M	512	ft-DPR	18.6	40.9	63.1	11.59
	110M	512	ft-LegalBERT DPR	**23.2**	**46.9**	**68.5**	**14.67**

Table 3: Retrieval results on `CLERC/doc` and `CLERC/passage`, evaluated with Recall@10, @100, @1000 and nDCG@10. Overall, zero-shot models are particularly bad due to distribution shifts between standard retrieval training data and the legal domain (Section 5.1). Models fine-tuned on legal data show effectively improved results.

Model	R1 ×100 ↑	R2×100 ↑	RL×100 ↑	BF ↑	CR ×100 ↑	CP×100 ↑	CFP×100 ↓
Mistral-7B-v0.3	23.78 / 18.48	8.00 / 4.63	21.89 / 16.78	-3.35 / -3.57	42.67 / 0.32	32.65 / 0.70	5.27 / 25.28
gemma-1.1-7b-it	18.33 / 16.82	5.28 / 3.99	16.70 / 14.81	-3.71 / -3.62	37.22 / 0.00	36.59 / 0.00	**4.34** / 84.62
GPT-4o	**26.82** / 19.87	**10.04** / 5.56	**24.89** / 17.93	-3.33 / -3.64	**89.87** / 3.28	**52.84** / 7.66	6.41 / 71.49
Llama-3	25.16 / 18.68	8.91 / 4.68	23.42 / 16.86	**-3.16** / -3.48	62.64 / 0.07	33.41 / 0.84	4.60 / 7.17
Avg Score	18.55	4.73	16.71	-3.52	0.84	2.13	40.64
+ ref	23.50	7.96	21.70	-3.38	54.88	36.39	5.13
Avg Gain (%)	+ 26.63	+ 68.08	+ 29.87	+ 3.98	+ 6471.43	+ 1603.76	- 87.37

Table 4: Analysis generation results on `CLERC/generation`, evaluated with F-scores of ROUGE and BARTScore (BF), CR, CP, and CFP. All models are instruct versions, and Llama-3 is shortened from Meta-Llama-3-8B-Instruct. Left / right column results correspond to prompting with / without the texts of cited cases. **GPT-4o generates analyses with the highest ROUGE F-scores but hallucinates the most**. **Prompting with cited cases shows effective improvement across all metrics**, especially boosting CR and CP and lowering CFP, making more factual citations in generations.

4.1.3 Main Results and Analysis

We show main results of the retrieval task in Table 3. Overall, we see that **no zero-shot model scores higher than 50% R@1K, indicating poor performance at retrieving case citations**. Due to the large size of the corpus, we find that late interaction models like ColBERTv2 (Santhanam et al., 2022) perform the worst, followed by Bi-Encoder models which score around 42% R@1K. BM25 performs the best at 48.3% R@1K among all zero-shot models. We show later in Section 5.1 that this poor performance is due to domain shift to legal text and the struggles of models with long queries.

Domain shift can be alleviated through training: fine-tuning BERT-base with the DPR architecture on CLERC's training set, we find a 36+ percent gain in R@1K, indicating that training on longer queries and legal text effectively improves performance in legal case retrieval, and it achieves the second highest performance after fine-tuned LegalBERT DPR, which is based on a model pretrained with legal texts (Holzenberger et al., 2020) and experiences less domain shift.

We also experiment with multiple Cross-Encoders and attach the results and analysis in C.3.

4.2 Generating Legal Analyses from Citations

Task Formulation Suppose a case document consisting of paragraphs $\{p_1, p_2, \ldots, p_N\}$, given $R_t = \{r_1, r_2, ...r_T\}$, a set of relevant passages or documents, and a prefix of case document $(p_1, p_2, \ldots, p_{t-1})$, a model is asked to continue to generate the next paragraph $\hat{p}_t$. The generation quality is measured by the similarity between p_t and $\hat{p}_t$. In practice, the gold paragraph p_t is randomly sampled from the last third of the paragraphs, i.e. $\lfloor \frac{2N}{3} \rfloor \leq t \leq N-2$, and R_t is selected as the citations used in p_t.

4.2.1 Experimental Setup

Citation Metrics To measure the faithfulness and degree of hallucination of legal analysis generation, we adopt Citation Recall (CR) and Citation Precision (CP) from Liu et al. (2023) and also

FINDINGS OF FACT AND CONCLUSIONS OF LAW
PAEZ, District Judge.
I. INTRODUCTION
The Board of Trustees of the Aircondi-tioning and Refrigeration Industry Health and Welfare Trust Fund ("Health Fund"), the Board of Trustees of the Aircondition-ing and Refrigeration Industry Retirement Trust Fund ("Retirement Fund"), and the Board of Trustees of the Airconditioning and Refrigeration Industry Defined Contribution Retirement Plan ("401(k) Plan") (collectively, the "Trust Funds") brought this action against J.R.D. Mechanical Services, Inc. ...(Omitted for brevity)...

II. FINDINGS OF FACT
1. Defendant J.R.D. is a California corporation and at all material times was an employer within the meaning of section 3(5) of ERISA, 29 U.S.C. § 1002(5) ...(Omitted for brevity)...

III. CONCLUSIONS OF LAW
Based on the forgoing findings of fact, the Court makes the following conclusions of law: ...(Omitted for brevity)...

"'Any' control over disposition of plan money makes the person who has the control a fiduciary." *IT Corp. v. General Am. Life Ins. Co., 107 F.3d 1415, 1421 (9th Cir.1997), cert. denied, 522 U.S. 1068, 118 5.Ct. 738, 139 L.Ed.2d 675 (1998).* Thus, "a person with authority to direct payment of a plan's money [is] deemed a fiduciary." Id. Divers' status as a corporate officer does not exempt him from liability. See *Kayes v. Pacific Lumber Co., 51 F.3d 1449, 1459 (9th Cir.1995)* ("This court has held corporate officers to be liable as fiduciaries on the basis of their conduct and authority with respect to ERISA plans."). Rather, he is liable because, as established by the factual findings, he exercised authority and control over the management and disposition of the Trust Funds' assets.
...(Omitted for brevity)...

Figure 5: Example case document (*99 F. Supp. 2d 1115*). Blue texts are analytical texts and *italicized texts* are case citations. This is a typical case document which starts with introductions and facts (**I** and **II**) about the case and then dives into legal analyses and reasoning at later paragraphs (**III**). See casetext link for full text: `https://casetext.com/case/board-of-trustees-of-aircondition-v-jrd-mech`.

propose Citation False Positive rate (CFP). Suppose a piece of generated text contains citations $\{c_1, \ldots, c_M\}$ and a set of citations that should be cited, C_r, we define the following metrics:

$$CR = \frac{\sum_i^M \mathbb{1}\left[c_i \in C_r\right]}{|C_r|}$$

$$CP = \frac{\sum_i^M \mathbb{1}\left[c_i \in C_r\right]}{M}$$

$$CFP = 1 - \frac{\sum_i^M \mathbb{1}\left[c_i \in C_r \vee \left(\bigvee_{j=1}^{t-1} c_i \sqsubseteq p_j\right)\right]}{M}$$

where $\mathbb{1}$ is the indicator function, and $a \sqsubseteq b$ indicates that a is a substring of b. CR measures the fraction of relevant citations being generated, and CP measures the fraction of generated citations being relevant. CFP measures the fraction of the generated citations being hallucinated, which means that they come from unlisted sources. See an example calculation in Figure 7.

Metrics, Data, and Models We evaluate legal analysis generation with ROUGE (Lin, 2004), BARTScore (Yuan et al., 2021), CP, CR, and CFP.

We run greedy generations on the test set (1k instances) of `CLERC/generation` with Meta-Llama-3-8B-Instruct,[8] Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), GPT-4o (Josh Achiam et al., 2024), and Gemma-1.1-7b-it (Reid et al., 2024). We use FastChat (Zheng et al., 2023) to host the open source models and the OpenAI API to access GPT-4o.[9] All the experiments are done using greedy decoding with a temperature of 1. Readers may refer to Appendix B for the prompts used.

Ablation study To study the effect of retrieved cases on generation, we perform an ablation study by not providing any citation documents to the model. The model is asked to generate the text solely based on the previous text, i.e. $(p_1, p_2, \ldots, p_{t-1})$.

4.2.2 Results and Analysis

Table 4 shows results of analysis generation on `CLERC/generation`. GPT-4o overall demonstrates the highest performance in terms of F-scores of ROUGE, CP, and CR, while Meta-Llama-3-8B-Instruct scores highest in BARTScore and second-highest across multiple metrics.

Under the full prompt, GPT-4o hallucinates the most, as reflected by its CFP score (see Figure 7 for an extended example). Gemma-1.1-7b-it achieves both the minimum CFP with the full prompt and the worst CFP with the shortened prompt, which shows that it is most sensitive to prompting under our setting.

[8] `https://github.com/meta-llama/llama3`

[9] `https://platform.openai.com/playground`

Prompting with the texts of cited cases effectively boosts performance, helping models cite more comprehensively, precisely, and with less hallucination, which is crucial for legal writing.

However, even with the full prompt, **generations have low CP and CR and considerably high CFP, which is concerning for legal applications**. Moreover, **CP, CR, and CFP overestimate analysis quality**. As we will discuss in Section 5.2, high scores on the citation metrics do not imply that outputs are factual and analytically sound.

5 Discussion

5.1 Limitations in Case Retrieval

We analyze why popular retrieval models struggle on the CLERC retrieval task, as shown in Table 3. As an example, consider this shortened query:

> ··· (Omitted for brevity) ··· St. Paul Reins. Co., Ltd. v. Commercial Fin. Corp., 198 F.R.D. 508, 512 (N.D.Iowa 2000) (Boilerplate objections, including that a particul ar document request was oppressive, burdensome and harassing, were "[i]n eve ry respect ... textbook examples of what federal courts have routinely deeme d to be improper objections.") ··· (Omitted for brevity) ··· Elk Run's Second Set of Interrogatories The court notes that Plaintiffs answers to Elk Run's second set of interrogatories are utterly deficient because they are not answered under oath, signed by the person making the answers as required by Rule 33(b)(1) of the Federal Rules of Civil Procedure and **REDACTED** In addition to the deficiency outlined above, the court finds that the answers to interrogatory numbers 1 and 2 are inadequate, unresponsive, and inconsistent with the requirements of Rule 33. ··· (Omitted for brevity) ···

where REDACTED is a placeholder for the removed central citation sentence, and we highlight underlined red words as "distractor" words.

After examining the full texts of top-ranked results when searching with BM25, we discover that **models can be easily be "distracted" and retrieve non-central citations, usually due to the many overlaps of distractor words**. In the highest ranked case, *228 F.R.D. 536*, the word "Elk Run" appears 45 times, which is the name of a sued company in the referenced case. In the second-highest case, *309 F.R.D. 476*, the word "Boilerplate" appears 67 times, whereas "Boilerplate objection" is a common type of objection raised in court (Baicker-McKee and Janssen, 2022). These distractor words are often common words and legal jargon, which can have many occurrences in irrelevant documents and make it harder for popular models to retrieve. Hence, this domain shift (and our results in Table 3) show how important fine-tuning on legal data is for legal retrieval models.

5.2 Limitations in Analysis Generation and Metrics

As discussed in Section 4.2.2, the current models struggle on the citation metrics. However, we show an example to illustrate that **scoring high on current citation metrics does not imply good generation quality.** We sample and show a fragment of gold paragraph from 99 F. Supp. 2d 1115 (C.D. Cal. 1999) in Figure 5:

> *Divers' status as a corporate officer does not exempt him from liability.* See Kayes v. Pacific Lumber Co., 51 F.3d 1449, 1459 (9th Cir.1995)

where the *red italicized* texts describe the analytical claim, supported by the case citation in blue. On the other hand, Llama-3-8B outputs:

> *As previously stated, 29 U.S.C. § 1002(21)(A) defines a fiduciary as a person who exercises any authority or control respecting management or disposition of [a plan's] assets.* P51 F.3d 1449, 1459 (9th Cir.1995).

While the generation makes the correct citation ("P51 F.3d 1449, 1459 (9th Cir.1995)") and scores high on the citation metrics, **it fails to generate an analytical claim that can be supported by the citation**. From the gold paragraph, the citation should support the claim "corporate officer has liability." However, the generated claim only defines what a fiduciary is, which was already defined in the previous context, rather than conveying useful new information for reasoning. See an additional example of GPT-4o hallucination in Appendix D.

Therefore, we welcome the development of automated metrics that focus on evaluating the *factual recall* aspect of text generation. One such approach could be extending metrics such as FActScore (Min et al., 2023) to the legal domain. However, as extracting the legal analyses itself is non-trivial in the legal domain, we leave this to future work.

6 Conclusion

Case law data is a high-quality resource for building intelligence systems to assist legal professionals. Through an interdisciplinary effort with legal professionals, we build CLERC, a dataset designed for training and evaluating models on long-context case retrieval and analysis generation tasks. We hope our work opens the gate for more exciting discoveries in AI built with legal text data.

7 Limitations

Our work provides an initial step into developing systems that can aid legal professionals. However, as we've shown, current systems are not well-suited to helping legal professionals on these tasks at the current moment, even those that are fine-tuned.

We use the case data from CAP, which gathered the data through OCR. As such, our data may contain small OCR errors, although our analysis has found it to be generally of high quality.

Ethical Considerations

In this work, we curate our dataset based on the corpus of historical U.S. federal case law. However, past case law might reflect racisim, sexism, and all other ideologies that people have considered controversial or undesirable for society and people's well-being. Building legal intelligence systems on case law data has the risk of propagating such historical biases and ideologies, which we caution the NLP community to be aware of and encourage future endeavors to address.

References

Anthropic. 2024. The Claude 3 Model Family: Opus, Sonnet, Haiku.

Arian Askari, Suzan Verberne, O Alonso, S Marchesin, M Najork, and G Silvello. 2021. Combining lexical and neural retrieval with longformer-based summarization for effective case law retrieval. In *DESIRES*, pages 162–170.

Steven Baicker-McKee and William M. Janssen. 2022. *Federal Civil Rules Handbook*. Thomson Reuters.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer.

David C Blair and Melvin E Maron. 1985. An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Communications of the ACM*, 28(3):289–299.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Aydar Bulatov, Yuri Kuratov, and Mikhail S. Burtsev. 2022. Recurrent Memory Transformer. In *Proceedings of Conference on Neural Information Processing Systems (NeurIPS)*.

Lang Cao, Zifeng Wang, Cao Xiao, and Jimeng Sun. 2024. PILOT: Legal Case Outcome Prediction with Case Law.

CAP. 2024. Caselaw access project.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. 2023. Extending Context Window of Large Language Models via Positional Interpolation.

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, David Belanger, Lucy Colwell, and Adrian Weller. 2021. Rethinking Attention with Performers. In *Proceedings of International Conference on Learning Representations (ICLR)*.

Jack Cushman, Matthew Dahl, and Michael Lissner. 2021. eyecite: A tool for parsing legal citations. *Journal of Open Source Software*, 6(66):3617.

Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E. Ho. 2024. Large legal fictions: Profiling legal hallucinations in large language models. *ArXiv*, abs/2401.01301.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Zican Dong, Tianyi Tang, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. 2024. BAMBOO: A Comprehensive Benchmark for Evaluating Long Text Modeling Capacities of Large Language Models. In *Proceedings of Language Resources and Evaluation Conference (LREC)*.

Esingbemi Princewill Ebietomere and Godspower Osaretin Ekuobase. 2019. A semantic retrieval system for case law. *Applied Computer Systems*, 24(1):38–48.

Daniel Y Fu, Simran Arora, Jessica Grogan, Isys Johnson, Sabri Eyuboglu, Armin W Thomas, Benjamin Spector, Michael Poli, Atri Rudra, and Christopher Ré. 2023. Monarch mixer: A simple sub-quadratic gemm-based architecture. In *Advances in Neural Information Processing Systems*.

Luyu Gao, Xueguang Ma, Jimmy J. Lin, and Jamie Callan. 2022. Tevatron: An efficient and flexible toolkit for dense retrieval. *ArXiv*, abs/2203.05765.

Joseph Gesnouin, Yannis Tannier, Christophe Gomes Da Silva, Hatim Tapory, Camille Brier, Hugo Simon, Raphael Rozenberg, Hermann Woehrel, Mehdi El Yakaabi, Thomas Binder, Guillaume Marie, Emilie Caron, Mathile Nogueira, Thomas Fontas, Laure Puydebois, Marie Theophile, Stephane Morandi, Mael Petit, David Creissac, Pauline Ennouchy, Elise Valetoux, Celine Visade, Severine Balloux, Emmanuel Cortes, Pierre-Etienne Devineau, Ulrich Tan, Esther Mac Namara, and Su Yang. 2024. Llamandement: Large language models for summarization of french legislative proposals. *ArXiv*, abs/2401.16182.

Randy Goebel, Yoshinobu Kano, Mi-Young Kim, Juliano Rabelo, Ken Satoh, and Masaharu Yoshioka. 2024. Overview and discussion of the competition on legal information, extraction/entailment (coliee) 2023. *The Review of Socionetwork Strategies*.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, Jessica Wu, Joe Nudell, Joel Niklaus, John Nay, Jonathan H. Choi, Kevin Tobia, Margaret Hagan, Megan Ma, Michael Livermore, Nikon Rasumov-Rahe, Nils Holzenberger, Noam Kolt, Peter Henderson, Sean Rehaag, Sharad Goel, Shang Gao, Spencer Williams, Sunny Gandhi, Tom Zur, Varun Iyer, and Zehua Li. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Preprint*, arXiv:2308.11462.

Michael Günther, Georgios Mastrapas, Bo Wang, Han Xiao, and Jonathan Geuter. 2023. Jina embeddings: A novel set of high-performance sentence embedding models. In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, pages 8–18, Singapore. Association for Computational Linguistics.

Peter Henderson, Mark Krass, Lucia Zheng, Neel Guha, Christopher Manning, Dan Jurafsky, and Daniel E Ho. 2022. Pile of law: Learning responsible data filtering from the law and a 256gb open-source legal dataset.

Nils Holzenberger, Andrew Blair-Stanek, and Benjamin Van Durme. 2020. A dataset for statutory reasoning in tax law entailment and question answering. In *NLLP@KDD*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning.

Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L'elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *ArXiv*, abs/2310.06825.

Hang Jiang, Xiajie Zhang, Robert Mahari, Daniel Kessler, Eric Ma, Tal August, Irene Li, Alex 'Sandy' Pentland, Yoon Kim, Jad Kabbara, and Deb Roy. 2024. Leveraging large language models for learning complex legal concepts through storytelling. *ArXiv*, abs/2402.17019.

Josh Achiam, Steven Adler, Sandhini Agarwal, and et al. 2024. GPT-4 Technical Report.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020a. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Yu Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. 2020b. Dense passage retrieval for open-domain question answering. *ArXiv*, abs/2004.04906.

Omar Khattab and Matei Zaharia. 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, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 39–48. ACM.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen-tau Yih, Tim Rocktaschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Proceedings of Conference on Neural Information Processing Systems (NeurIPS)*.

Dacheng Li, Rulin Shao, Anze Xie, Ying Sheng, Lianmin Zheng, Joseph E Gonzalez, Ion Stoica, Xuezhe Ma, and Hao Zhang. 2023. How Long Can Context Length of Open-Source LLMs truly Promise? In *Proceedings of Workshop on Instruction Tuning and Instruction Following*.

Haitao Li, Yunqiu Shao, Yueyue Wu, Qingyao Ai, Yixiao Ma, and Yiqun Liu. 2024. Lecardv2: A large-scale chinese legal case retrieval dataset. *ArXiv preprint*.

Quanzhi Li and Qiong Zhang. 2021. Court opinion generation from case fact description with legal basis. In *AAAI Conference on Artificial Intelligence*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Annual Meeting of the Association for Computational Linguistics*.

Chun-Hsien Lin and Pu-Jen Cheng. 2024. Legal documents drafting with fine-tuned pre-trained large language model. *Software Engineering & Trends*.

Matthew Lippmann. 2018. *Criminal Procedure*, 3rd edition. SAGE Publications, Inc, Thousand Oaks, CA.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics (TACL)*, pages 157–173.

Nelson F. Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating verifiability in generative search engines. *ArXiv*, abs/2304.09848.

Daniel Locke and Guido Zuccon. 2022. Case law retrieval: problems, methods, challenges and evaluations in the last 20 years. *arXiv preprint arXiv:2202.07209*.

Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. 2023. Fine-tuning llama for multi-stage text retrieval. *ArXiv*, abs/2310.08319.

Yixiao Ma, Yunqiu Shao, Yueyue Wu, Yiqun Liu, Ruizhe Zhang, Min Zhang, and Shaoping Ma. 2021. Lecard: A legal case retrieval dataset for chinese law system. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2342–2348.

Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning, and Daniel E. Ho. 2024. Hallucination-free? assessing the reliability of leading ai legal research tools.

Robert Mahari. 2021. Autolaw: Augmented legal reasoning through legal precedent prediction. *ArXiv*, abs/2106.16034.

Robert Mahari, Dominik Stammbach, Elliott Ash, and Alex'Sandy' Pentland. 2023. Lepard: A large-scale dataset of judges citing precedents. *arXiv preprint*.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy J. Lin. 2020. Document ranking with a pre-trained sequence-to-sequence model. In *Findings*.

Guanghui Qin, Yukun Feng, and Benjamin Van Durme. 2023. The NLP task effectiveness of long-range transformers. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3774–3790, Dubrovnik, Croatia. Association for Computational Linguistics.

Guanghui Qin, Corby Rosset, Ethan C Chau, Nikhil Rao, and Benjamin Van Durme. 2024a. Dodo: Dynamic contextual compression for decoder-only lms. In *Proceedings of the 2024 Conference on Association for Computational Linguistics (ACL)*.

Weicong Qin, Zelin Cao, Weijie Yu, Zihua Si, Sirui Chen, and Jun Xu. 2024b. Explicitly Integrating Judgment Prediction with Legal Document Retrieval: A Law-Guided Generative Approach.

Machel Reid, Nikolay Savinov, Denis Teplyashin, and et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context.

Ruiyang Ren, Yingqi Qu, Jing Liu, Wayne Xin Zhao, QiaoQiao She, Hua Wu, Haifeng Wang, and Ji-Rong Wen. 2021. RocketQAv2: A joint training method for dense passage retrieval and passage re-ranking. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2825–2835, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Columbia Law Review. 2020. *The Bluebook: A Uniform System of Citation*. The Harvard LawReview Association Gannett House, 1511 Massachusetts Avenue Cambridge, Massachusetts 02138 U.S.A.

Jon Saad-Falcon, Daniel Y. Fu, Simran Arora, Neel Guha, and Christopher Ré. 2024. Benchmarking and building long-context retrieval models with loco and m2-bert. *arXiv preprint arXiv:2402.07440*.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 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*, pages 3715–3734, Seattle, United States. Association for Computational Linguistics.

T. Y. S. S. Santosh, Rashid Haddad, and Matthias Grabmair. 2024. Ecthr-pcr: A dataset for precedent understanding and prior case retrieval in the european court of human rights. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 5473–5483. ELRA and ICCL.

Uri Shaham, Elad Segal, Maor Ivgi, Avia Efrat, Ori Yoran, Adi Haviv, Ankit Gupta, Wenhan Xiong, Mor Geva, Jonathan Berant, and Omer Levy. 2022. SCROLLS: Standardized CompaRison over long language sequences. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 12007–12021, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zejiang Shen, Kyle Lo, Lauren Jane Yu, Nathan Dahlberg, Margo Schlanger, and Doug Downey. 2022. Multi-lexsum: Real-world summaries of

civil rights lawsuits at multiple granularities. *ArXiv*, abs/2206.10883.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. RoFormer: Enhanced transformer with Rotary Position Embedding. *Neurocomputing*, page 127063.

Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. 2020. Long Range Arena: A Benchmark for Efficient Transformers.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *ArXiv preprint*, abs/2104.08663.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Oscar Tuvey and Procheta Sen. 2023. Automated argument generation from legal facts. *arXiv:2310.05680*.

Santosh T.Y.S.S, Rashid Gustav Haddad, and Matthias Grabmair. 2024. Ecthr-pcr: A dataset for precedent understanding and prior case retrieval in the european court of human rights. *ArXiv preprint*.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *ArXiv*, abs/2212.03533.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2023. Improving text embeddings with large language models. *arXiv preprint arXiv:2401.00368*.

Orion Weller, Benjamin Chang, Sean MacAvaney, Kyle Lo, Arman Cohan, Benjamin Van Durme, Dawn Lawrie, and Luca Soldaini. 2024. Followir: Evaluating and teaching information retrieval models to follow instructions. *ArXiv*, abs/2403.15246.

Robert A Wilson. 1962. Computer retrieval of case law. *Sw. LJ*, 16:409.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding. *Preprint*, arXiv:2309.07597.

Eugene Yang, Dawn Lawrie, James Mayfield, Douglas W. Oard, and Scott Miller. 2024. Translate-distill: Learning cross-language dense retrieval by translation and distillation. In *Proceedings of the 46th European Conference on Information Retrieval (ECIR)*.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *ArXiv*, abs/2106.11520.

Xinyu Crystina Zhang, Andrew Yates, and Jimmy J. Lin. 2021. Comparing score aggregation approaches for document retrieval with pretrained transformers. In *European Conference on Information Retrieval*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

A Compute Infrastructure

We run our experiments on 24GB Nvidia Quadro RTX 6000 GPUs for retrieval and 80GB Nvidia A100s for RAG experiments. Retrieval indexing takes around 60 hours for a BERT-base sized model on one GPU. For RAG experiments, they take approximately 2 hours per model on one GPU.

B Prompt for Argument Generation

User prompt:

```
Here are some reference articles for legal cases:
# Reference case {case_key_1}
{text of cited case 1}
# Reference case {case_key_2}
{text of cited case 2}
...
# Reference case {case_key_N}
{text of cited case N}

Here is the text I've written so far:
# Paragrah
{previous_text}

Continue to write it following the style of my writeup. Your answer contains 100 to 400 words. You must explicitly use the reference cases and mention their reference ids, i.e. {case_key_1}, {case_key_2} . . . {case_key_N}. Wrap your answer with <answer></answer>. Make your answer concise and avoid redundant languages.
```

{previous_text} refers to the previous context before the reference paragraph, i.e.

$(p_1, p_2, \ldots, p_{t-1})$ defined in Section 4.2. `{text cited case i}` refers to the reference articles, i.e. R_t defined in Section 4.2, and `{case_key_i}` are the associated citation keys.

For the ablation study, we experiment prompting LLMs without explicitly providing the text of cited cases with the following prompt:

```
User prompt:
Here is the text I've written so far:
# Paragrah
{previous_text}

Continue to write it following the style of my writeup. Your answer contains 100 to 400 words. Wrap your answer with <answer></answer>. Make your answer concise and avoid redundant languages.
```

C Additional Experimental Results

C.1 Parse Direct Quotes and Filter Queries

We design a simple heuristics program to parse direct quotes. CAP uses special double quotation marks for all direct quotes, which have unicode values of 8220 and 8221. We extract all texts quoted by these special quotation marks, and we pair them with the nearest case citation.

To locate and filter the central citation sentences in the queries, we design an involved RegEx program. The high-level overview is that it will find the start of the citation sentence, which is usually some Entity A v. Entity B, and we recognize the start by referring to the metadata of CAP. There are other variant formats to start a citation sentence, such as *Id.*, *See.*, and *In re*, and we also accomodate these formats in the filter program (Review, 2020). We find the end of the citation by locating some notable punctuations, such as closing parenthesis, semicolon, etc. In the event that we find outlier citation sentences, we would report them as failure cases. The average failure rate is around 11%.

For more specific details, please refer to our code and repo.

C.2 Retrieve with Direct Quote

Around 33% of the time, we find that case documents cite other cases in the form of direct quotations. We parse such direct quotes from case documents and used them as queries for retrieval, expecting a near-perfect accuracy. However, conducting exact matches with `grep`, we only retrieve 16% of the direct quotes. We discover that at least 25.4% of the direct quotes experience subtle alterations with the use of square brackets (see Figure 6), which partly causes the low recall.

Therefore, we further conduct retrieval experiments of direct quotes with BM25 and n-gram match, in the hope that these "soft" term matching methods could overcome subtle alterations and yield a higher recall. Results in Table 5 show that the highest Recall@1000 is only 66.7%, which is far from perfect.

By examining the case texts, we attribute this result to mainly three reasons. First, direct quote, when being cited, commonly experiences slight alterations in punctuation, wording, and pronouns for clarification purposes or due to human errors (see Figure 6).

Second, when retrieving a direct quote that consists of common words or widely used legal terms, soft term matching methods could retrieve multiple false positive documents that contain multiple occurrences of these common terms and push down the ranking of the actual positive document.

Third, direct quotes often appear in contexts of various neighboring case citations that point to a set of candidate documents. Unless investigating and searching all candidate documents, it is not apparent which direct quote corresponds to which case. While we develop a program (see Section C.2) to determine the correspondence based on the proximity of direct quotes to citations, this correspondence determination is imperfect and remains an open question to be addressed in future works.

C.3 Retrieve with Cross-Encoders

We conduct additional retrieval experiments with the same data and metrics setup in Section 4.1 using multiple Cross-Encoders in Table 6. We rerank the top-100 outputs of E5 (Wang et al., 2023) with RankLlama (Ma et al., 2023), MonoT5-3B (Nogueira et al., 2020), and FollowIR (Weller et al., 2024). We discover that reranking with cross-encoders degrades the retrieval performance. A possible explanation is that due to domain mismatch, rerankers are unable to rerank long-context legal data that have distinct styles and length from the data rerankers were trained on.

Types of alterations in direct quote:

Contextual clarification:
Query: "A reading of the entire [Wage] Act clearly shows that the purpose of the Act is to assist..."
Original: "A reading of the entire Act clearly shows that the purpose of the Act is to assist..."

Rephrases:
Query: "punishable for a term of imprisonment exceeding one year."
Original: "punishable by imprisonment for more than one year."

Punctuation changes:
Query: "an account of the time, place, and specific content of the false representations as well as the identities of the parties to the misrepresentations."
Original: an account "of the time, place, and specific content of the false representations as well as the identities of the parties to the misrepresentations."

Figure 6: Direct quote may experience different types of alterations and degrade retrieval performance.

	Recall@5 ↑	Recall@10 ↑	Recall@100 ↑	Recall@1000 ↑	nDCG@10 ↑
BM25	18.8	25.2	**45.7**	**66.7**	15.8
5-gram	30.3	**36.4**	44.7	44.8	**23.0**
12-gram	**31.1**	35.7	40.1	40.1	22.8

Table 5: Retrieval experiment results using direct quote as query. All numbers are in percentage.

D Qualitative Example

Please see Figure 7 for an example of problematic legal analysis generated by GPT-4o, which exhibits hallucination and achieves low citation precision and recall. We discover that citations from the cited cases are a possible source of hallucination, meaning that the model confuses the cases they should cite with the citations in those cases. It is possible to mitigate this with more refined prompt engineering, which we will leave to future works.

Type	Params	Length	Model	Metrics	
				R@10	nDCG@10
Bi-Encoders	110M	512	E5	**8.4**	**3.67**
Cross-Encoders	7B	512	E5 + RankLlama	4.7	2.14
	3B	512	E5 + MonoT5	4.2	2.00
	7B	512	E5 + FollowIR	5.1	2.14

Table 6: Retrieval results with Cross-Encoders. All numbers are in percentage

Gold paragraph:
A pro se complaint, "however unlawfully pleaded," is held "to less stringent standards than formal pleadings drafted by lawyers." Hughes v. Rowe, *449 U.S. 5, 9, 101 S.Ct. 173, 66 L.Ed.2d 163 (1980)*. The court's role is to ensure that claims of pro se litigants are given "fair and meaningful consideration." Matzker v. Herr, *748 F.2d 1142, 1146 (7th Cir.1984)*. Accordingly, pro se complaints must be liberally construed. Estelle v. Gamble, *429 U.S. 97, 106, 97 S.Ct. 285, 50 L.Ed.2d 251 (1976)*. The complaint need not specify the correct legal theory, nor point to the correct statute in order to survive a motion to dismiss. Bartholet v. Reishauer A.G. (Zurich), `953 F.2d 1073, 1078 (7th Cir.1992)`. Finally, a district court may dismiss a complaint only if "it appears beyond doubt that the plaintiff can prove no set of facts in support of his claim which would entitle him to relief." Hughes, supra, 449 U.S. at 10, 101 S.Ct. 173.

GPT-4o generation:
Liberal Construction of Pro Se Complaints

The City contends that Mr. Cleaves' complaint should be dismissed because it fails to state a claim upon which relief can be granted. However, it is well-established that complaints filed by pro se litigants must be held to less stringent standards than those drafted by lawyers. As the Supreme Court noted in Haines v. Kerner, "however inartfully pleaded," a pro se complaint must be liberally construed and should not be dismissed unless it appears beyond doubt that the plaintiff can prove no set of facts in support of his claim which would entitle him to relief (404 U.S. 519, 520-21 (1972)).

Similarly, in Hughes v. Rowe, the Court reiterated this principle, emphasizing that the district court's role is to ensure that pro se litigants' claims are given fair and meaningful consideration (*449 U.S. 5, 9-10*, 101 S.Ct. 173, 175, 66 L.Ed.2d 163 (1980)). Courts must determine whether the plaintiff should be allowed to present evidence on his claims before deciding on a motion to dismiss (*748 F.2d 1142*).

Additionally, considering the liberal standards applied to such cases as set forth in Estelle v. Gamble (*429 U.S. 97, 106 (1976)*), and in light of the notice-pleading standard of the Federal Rules of Civil Procedure (Rule 8(a)), Mr. Cleaves' allegations, though detailed, sufficiently provide the City with notice of the claims and underlying facts.

Therefore, the City's motion to dismiss Mr. Cleaves' pro se complaint is denied as it fails to demonstrate that there are no conceivable facts under which Mr. Cleaves could prevail.

Reference texts:
449 U.S. 5
purports to justify or explain the segregation of petitioner for two days in advance of the disciplinary hearing. II Petitioner's complaint, like most prisoner complaints filed in the Northern District of Illinois, was not prepared by counsel. It is settled law that the allegations of such a complaint, "however inartfully pleaded" are held "to less stringent standards than formal pleadings drafted by lawyers ...." Haines v. Kerner, 404 U. S. 519, 520 (1972). See also Maclin v. Paulson, 627 F. 2d 83, 86 (CA7 1980); French v. Heyne, 547 F. 2d 994, 996 (CA7 1976). · · · (omitted for brevity) · · ·
748 F.2d 1142
magistrate erred in applying the deliberate indifference standard in dismissing this complaint. III. Our inquiry turns, therefore, to the question of whether Matzker stated a cause of action when judged under due process standards. A complaint drafted by a pro se litigant "however inartfully pleaded," is held "to less stringent standards than formal pleadings drafted by lawyers." Hughes v. Rowe, 449 U.S. 5, 9, 101 S.Ct. 173, 175, 66 L.Ed.2d 163 (1980). The district court's role is to ensure that the claims of pro se litigants are given "fair and meaningful consideration." · · · (omitted for brevity) · · ·
429 U.S. 97
to evidence deliberate indifference to serious medical needs. It is only such indifference that can offend "evolving standards of decency" in violation of the Eighth Amendment. Ill Against this backdrop, we now consider whether respondent's complaint states a cognizable § 1983 claim. The handwritten pro se document is to be liberally construed. · · · (omitted for brevity) · · ·
953 F.2d 1073
it done so, ERISA still would govern unless the plan were "unfunded." The exemption recognizes that a suit concerning an unfunded plan is one directly against the employer's assets, and as ERISA leaves excess benefit plans substantively unregulated there is no reason to oust state law. · · · (omitted for brevity) · · ·

Figure 7: Example GPT-4o generation that exhibits hallucination and low citation precision and recall. The input of generation is all referenced texts concatenated with all paragraphs before the gold paragraph in the original case document. The generation is supposed to generate all citations in the gold paragraph. *Green italicized texts* are citations successfully generated, `yellow texts in teletype fonts` are citation missed in generation, and red underlined texts are hallucinated citations. GPT-4o should cite all referenced cases (449 U.S. 5, 748 F.2d 1142, 429 U.S. 97, 953 F.2d 1073). Instead, since we provide the actual texts of referenced cases, **GPT-4o hallucinates citations from citations in referenced cases**, which we underline with red color. In summary, the generation achieves the following results on citation metrics: $CP = 3/5$, $CR = 3/4$, $CFP = 2/5$.