A Reasoning-Focused Legal Retrieval Benchmark

Lucia Zheng*
Stanford University
Stanford, California, USA
zlucia@stanford.edu

Neel Guha*
Stanford University
Stanford, California, USA

Javokhir Arifov
Stanford University
Stanford, California, USA

Sarah Zhang
Stanford University
Stanford, California, USA

Michal Skreta
Stanford University
Stanford, California, USA

Christopher D. Manning
Stanford University
Stanford, California, USA

Peter Henderson[†]
Princeton University
Princeton, New Jersey, USA

Daniel E. Ho[†]
Stanford University
Stanford, California, USA

Abstract

As the legal community increasingly examines the use of large language models (LLMs) for various legal applications, legal AI developers have turned to retrieval-augmented LLMs ("RAG" systems) to improve system performance and robustness. An obstacle to the development of specialized RAG systems is the lack of realistic legal RAG benchmarks which capture the complexity of both legal retrieval and downstream legal question-answering. To address this, we introduce two novel legal RAG benchmarks: Bar Exam QA and Housing Statute QA. Our tasks correspond to real-world legal research tasks, and were produced through annotation processes which resemble legal research. We describe the construction of these benchmarks and the performance of existing retriever pipelines. Our results suggest that legal RAG remains a challenging application, thus motivating future research.

CCS Concepts

• **Computing methodologies → Natural language processing**;
• **Applied computing → Law**.

Keywords

retrieval, reasoning, benchmark, dataset

---

*Equal contribution.
[†]Equal advising.

---

1 Introduction

There is significant excitement towards using large language model (LLM) tools to improve the quality and cost of legal services [24]. Already, lawyers across the world have begun incorporating LLMs into legal practice, and applying them towards a range of tasks: answering questions about the law in various jurisdictions, identifying potential legal issues in client cases, drafting agreements, and more [1].

Applying LLMs towards legal tasks requires resolving distinctive challenges posed by the legal domain. For instance, legal tasks are often fact-intensive [8] and LLMs have a tendency to produce factually-ungrounded statements ("hallucinations") [4]. The law is also constantly changing—through new statues or judicial opinions—and the knowledge contained in LLM parameters is static [18]. To address these challenges, legal AI developers have begun deploying "RAG" systems, where LLMs are augmented with retrievers over corpora of case law, statutes, and other legal documents [25]. Given a lawyer query, the retriever fetches documents from the corpora relevant to the query, and requires the LLM to answer the query with respect to the retrieved documents.

However, a significant bottleneck in the development of legal RAG systems is the lack of realistic English legal open-domain question answer benchmarks. In particular, existing benchmarks suffer from one or more of the following weaknesses.

(1) First, they fail to represent tasks where where the query and relevant document have little lexical overlap, and identifying the relevant document requires multi-hop or analogical reasoning. In practice this setting is ubiquitous. Producing the legal cases relevant to a client's factual circumstances, for instance, requires extracting higher-order legal issues and identifying other cases which present those issues—even if the specific factual descriptions are quite different.

(2) Second, existing benchmarks are often exclusively retrieval benchmarks, and do not contain paired question-answers to evaluate downstream reasoning based on the retrieved information [2, 6, 11, 26, 36]. As a result, they do not capture the downstream impacts of improvements in retrievers.

(3) Finally, benchmarks rely on query-document distributions extracted from datasets built for other purposes, where queries

do not correspond to the types of questions lawyers might actually ask [6, 11, 32].

To address this gap, we introduce two new benchmark datasets for evaluating retrieval-augmented LLMs: Bar Exam QA and Housing Statute QA. These datasets address the deficiencies discussed above. In Bar Exam QA, queries correspond to reasoning-intensive Bar Exam hypothetical fact-patterns, and documents correspond to judicial opinion passages necessary for answering the hypothetical. In Housing Statute QA, queries correspond to practically useful questions about housing law, and documents correspond to statutes from different jurisdictions. Our datasets provide ~10K labeled, paired query, gold passage, answer examples for training and evaluating language models on legal retrieval and retrieval-augmented downstream QA tasks. The gold passages for these datasets are hand-annotated and validated by law students and researchers, through annotation processes modeled off of a lawyer's legal research process.

Concretely, our work makes three contributions. First, we describe the construction of these datasets (Section 3) and compare them to existing benchmarks (Section 4). We show that relative to existing benchmarks, ours captures query-document distributions where the lexical similarity between the query and document is low. Second, we benchmark existing state-of-the-art retrieval pipelines on these datasets and find that because of the low lexical similarity, common retrieval methods like BM25 struggle (Section 5). Third, we present results of a simple heuristic to verify that our benchmarks accurately measure improved legal reasoning in retrieval (Section 5.3). Specifically, we describe a law-inspired query expansion strategy with generative reasoning roll-outs. We find that this approach improves performance on our datasets.

Our work suggests that developers of retrieval-augmented legal LLM products may need to go further than simple retrievers to improve the performance of their approaches. In particular, they may need to ensure that retrievers can also be legal reasoners too, either through query expansion or increased embedding model capacities.

2 Related Works

2.1 Open-Domain QA Datasets

BEIR [35] is a widely used information retrieval (IR) benchmark, which consists of 18 datasets across 9 task types. Our datasets are most similar to the question answering tasks in BEIR, Natural Questions (NQ) [17], HotpotQA [41], and FiQA-2018 [27]. For this style of task, the retrieved passage is used as context to help the model on downstream question answering. For NQ and HotpotQA, the answer is an extractive span of the context passage. For Bar Exam QA and Housing Statute QA, the answer is in a multiple-choice format.

Past works have also discussed several limitations of existing general IR benchmarks: a skew towards web/search-engine style retrieval tasks, tasks with low lexical and syntactic variance between queries and gold passages, tasks with short query lengths, and propose new datasets to address these challenges [15, 35]. Most similar to our tasks, BIRCO [40] and BRIGHT [34] introduce new benchmark IR tasks with more complex task objectives and that require greater reasoning capabilities to solve. Compared to BEIR

[35], these tasks have significantly lower lexical similarities between queries and gold passages and longer query lengths. BIRCO and BRIGHT include reasoning tasks in the natural sciences, computer science, and theorem-based mathematics, but neither contains legal reasoning tasks that share similar types of deductive reasoning processes to mathematical reasoning tasks.

2.2 Legal Information Retrieval Datasets

Early work on retrieval of statutory law focus on building systems using lexical matching and extensive annotation of semantic features [7, 33].

More recently, several works release legal retrieval datasets constructed by leveraging case document structure/metadata to link a citing context in a new case (query) to precedential (prior) cases cited to support arguments made in the citing context (gold passage) [2, 6, 11, 26]. Though this automatic extraction approach enables the collection of large-scale datasets, the citing contexts often summarize the high-level rule from the cited case relevant to the argument to justify its citation.[1] As a result, we find the lexical similarity of the query and the gold passage is often quite high and comparable to ODQA datasets for these legal IR datasets (Section 4)

Additionally, since the queries are extracted directly from case opinions written by judges, they often do not reflect the natural distribution of user question-style queries that might be asked by a person seeking legal information or a lawyer conducting legal research. To our knowledge, there are few English-language legal IR datasets with natural question-style queries and expert gold passage annotations. Existing datasets [21, 22, 43] are in other languages.

Lastly, few legal IR datasets are paired with downstream tasks akin to open-domain QA. Thus, few of the available datasets are suitable for end-to-end evaluation of retrieved-augmented LLMs. CLERC [11] includes both a retrieval and retrieved-augmented generation task: given the beginning of a case and retrieved reference paragraphs from cited cases, a model is evaluated on its ability to generate continuing analysis paragraphs of the case. However, as discussed by the authors, automatically measuring factual recall of open-ended text generations is a challenging, unsolved problem [11]. Our datasets are linked to multiple-choice QA tasks. The classification setting makes it easier to automatically evaluate retrieval-augmented LLMs for factual correctness. In law, commons principles or rules are restated many times across the corpus (of case law). In settings where many references passages may be helpful or gold annotations are limited, downstream retrieval-augmented task performance can also be valuable for further contextualizing retrieval performance.

3 Datasets

Our datasets advance beyond existing open-domain QA datasets and legal IR datasets by offering concrete, substantive legal questions paired with both supporting gold passages and answers.

We highlight the key ways in which our datasets differ from existing datasets. First, our benchmark allows for evaluation of both retrieval and downstream question-answering. In contrast,

---

[1]The ECtHR-PCR dataset [36] is an exception. The authors leverage additional case document structure to separate facts from the argument in the citing case context when constructing queries.

many prior legal datasets, are either exclusively intended for retrieval evaluation (no associated question-answer pairs) [6, 11], or intended exclusively for question-answering (no associated document corpora) [8, 10, 42]. As our paper highlights, datasets that enable evaluation on both tasks allow for a more fine-grained understanding of end-to-end system performance.

Second, our questions and passage labels were hand-annotated by legal experts, bar exam writers and law students for Bar Exam QA and legal researchers for Housing Statute QA. We believe this makes the retrieval task more realistic compared to popular extractive constructions. In these extractive constructions, the query and passage pairs are derived from case citation relationships, where both the query and passage are sections of text extracted from the opinions of the citing case and the cited case. Because of their extractive nature, these datasets simulate the task of retrieving a rule from a cited case using the citing context, but the query is not typically a well-formed question and the passage is not always closely related to the query (due to challenges with localizing the relevant rule within the cited case) [11]. In contrast, our query and passage pairs represent substantive legal questions from multiple areas of law and explanatory rules or passages justifying the answer. To our knowledge, few (if any) English legal retrieval datasets were constructed with hand-annotated passage pairs; existing datasets cover French or Chinese law [21, 22, 43].

Third, our retrieval corpora, particularly for Housing Statute QA, are substantially larger (~1-2M documents) than the retrieval corpora used in several other legal and general retrieval benchmarks for reasoning intensive retrieval tasks (~10,000-100,000 documents) [6, 34, 40]. Retrieval corpora size matters because retrieval becomes harder to perform as the corpora increases in size and the relative fraction of irrelevant documents increases.

Finally, our datasets focus on specific types of documents and questions, legal rule-application questions over cases (spanning the traditional areas of law tested on the Bar Exam) and housing questions over state statutes, which existing benchmarks do not capture. This is important because the performance of retrievers and LLMs can vary, sometimes significantly, across question and document types.

We recognize our datasets cannot capture the full spectrum of complexities involved in real-world legal tasks. Our datasets are restricted in subject-matter domains and restricted to multiple-choice answer forms to enable automatic evaluation of the downstream task. They may not represent a "realistic" approximation of the full natural distribution of legal questions. In particular, the BarExamQA questions are drawn from practice bar exams, where the questions involve stylized, fictional, short fact patterns, which may not be similar to the distribution of real-world fact patterns attorneys encounter. However, these queries advance beyond existing legal retrieval datasets [6, 11] by offering concrete, substantive legal questions and enable comparison to bar exam QA tasks in other popular general reasoning benchmarks [9, 16]. But we acknowledge that as new efforts move towards creating more realistic bar exam questions with multistage factual scenarios, reapplying the techniques described to construct these datasets to those questions would provide even more realistic tasks [28]. The Housing Statute QA question are drawn from the LSC Eviction Laws Database, a

real-world resource designed to help address tenants' questions related to the legal process of eviction.

We describe the datasets in the benchmark and the process on construction in the following sections. Table 1 provides a summary of the datasets. Dataset release and license information is provided in Appendix C. We show representative examples from each dataset in Appendix D.

3.1 Bar Exam QA

The Bar Exam QA dataset is a dataset of multistate bar exam (MBE) questions. The multistate bar exam is a professional exam that certifies law students to practice law in the U.S. The Bar Exam QA datasets consists of MBE questions from historical bar exams released by the National Conference of Bar Examiners (NCBE) and practice bar exams from Barbri MBE test preparation workbook (2013 Ed.). Each MBE question contains a novel legal scenario, a question about a specific legal issue implicated in the scenario, and four answer choices. The task is to select the correct answer choice.

We transform the dataset into a retrieval task by collecting gold explanation passages for each example. For the Barbri practice bar exams, we extract the explanation passages from the answer key for each question as the gold passage. For the historical bar exams, for which no explanation passages are available, a law student hand-annotated each example with a gold passage that helps or supports the correct answer to the question. The law student's annotation process simulates the legal research process. We provide a detailed description of the annotation process in Appendix A. The authors and research assistants manually validated subsets of the examples and gold passages. Annotations took approximately 6 months for the team to complete.

The retrieval passage pool contains ~900K passages. The passage pool consists of the gold passages, U.S. caselaw from 2019-2021 (case decision text split at the paragraph-level), and Cornell Law School Legal Information Institute (LII) Wex legal encyclopedia entries and select primary sources.[2]

We release the subset of the dataset containing historical publicly released MBE questions (Historical MBE). We treat the Barbri questions as a private held-out subset and report separate results on this subset (Barbri). Because the historical publicly released MBE questions are contained in the MMLU auxiliary train set for professional law [9], these examples may have been used for model training. To our knowledge, the Barbri set has not been previously released in any dataset, and thus, are more likely to be true, unseen examples, for model evaluation. The Barbri set examples also reflect more modern styles of bar exam questions.

3.2 Housing Statute QA

Housing Statute QA is a question answering dataset covering statutory housing law across 50+ U.S. jurisdictions. Each sample in the dataset contains a Yes/No (Y/N) question about housing law in a particular state, the answer to the question, and a small number (≤ 10) of "relevant" statutes (which contain text support the correct answer). These statutes are mapped to an individual statute in a

---

[2]The source documents are segmented at the paragraph-level using this tool: https://github.com/neelguha/legal-segmenter.

Dataset	Total Number		Avg. Length		Examples
	Q	P	Q	P	
Ours					
Bar Exam QA	1,195/1,815	856,835	172/157	131	Table 7
Housing Statute QA	6,853	1,837,403	15	349	Table 8
Comparison					
Natural Questions [17]	3,452	2,681,468	11	102	Table 9
HotpotQA [41]	7,405	5,233,329	20	63	Table 10
COLIEE [6]	1,278	5,616	6,730	6,768	Table 11
CLERC [11]	2,851	1,842,422	415	3,303	Table 12

Table 1: Summary of datasets. We report number of queries (Q), number of passages (P), the average length of queries and passages (calculated with the GPT-2 tokenizer [29]), and examples. For Bar Exam QA, the query statistics are reported for the Historical MBE/Barbri subsets.

larger database of state law. In the retrieval setting, the objective is to identify the relevant statutes from the larger database.

Housing Statute QA was created by adapting the Legal Services Corporation (LSC) Eviction Laws Database [23]. The original database was constructed by legally trained researchers and students, who manually answered questions about housing law for different jurisdictions, by explicitly searching housing law in each jurisdiction. Similar to the annotation procedure for Bar Exam QA, the annotation process is modeled off of the legal research process [23]. The database provides questions, answers, and citations to statutes which support the answer.

The original database contains a mixture of free-response, multi-answer multiple-choice, and Y/N questions. Prior work has observed that evaluating LLM responses for non-Y/N responses can be challenging [8]. Thus, we restrict Housing Statute QA to only contain Yes/No questions. We do so by first using all the Y/N questions contained in the original LSC database. Next, we convert the multi-answer multiple-choice questions into new Y/N questions. For each answer-choice in multiple-choice answer space, we create a new Y/N question asking if that answer is true. Thus, from a single multiple-choice question with five answer choices, we derive five new Y/N questions. In Appendix B, Table 6, we provide an example of an original question from the database and our reformulated Y/N question. In Figure 1, we provide a histogram illustrating the distribution of the number of gold passages (statutes) per transformed example in the dataset.

The LSC Database annotates each question with citations to state laws which contain information relevant for answering the question. We build a corpus from Justia's available state statutes from the year 2021.[3] If the 2021 data from the jurisdiction was not available, the most recently published set of statutes was used. We use statute citations on the original questions to identify relevant statutes in this corpora. We note that Justia's coverage of state law is incomplete, and some state statutes are not available via Justia.

Our released version of Housing Statutes QA consists of two splits. The first split (rc_questions)—which we study here—contains 6,853 question-answer pair examples with labeled supporting statutes.
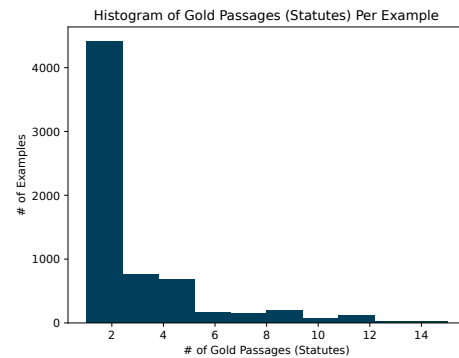
Figure 1: Histogram of number of gold passages (statutes) per example in the Housing Statue QA dataset.

This can be used as an evaluation set for the retrieval and downstream question answering tasks. The second split (knowledge_qa) is larger (9,297 examples), and contains question-answer pairs for which we could not identify a labeled supporting statute. While the lack of a statute annotation prevents these questions being used for retrieval evaluation, we believe they may be of independent interest to researchers. The retrieval passage pool contains ~2M passages.

4 Comparison to Existing Tasks

In typical open-domain question answering tasks, the relevant passages restates or closely restates a significant portion of the question and the answer is, by construction, a substring of the gold passage. Therefore, the relationship between the question, gold passage, and answer for such tasks can often be recovered by comparing lexical similarities of the texts. We find that this is also the case for existing legal IR datasets derived from case citation relationships [6]. In contrast, our tasks require a greater degree of reasoning to connect the question to the gold passage and answer and the gold passage is more lexically distant from both the question and the answer.

Dataset	$LS_{retrieval}$	LS_{QA}
Bar Exam QA	0.07 ± 0.00	0.14 ± 0.01
Housing Statute QA	0.09 ± 0.00	0.00 ± 0.00
Natural Questions	0.27 ± 0.01	0.25 ± 0.01
HotPotQA	0.26 ± 0.00	0.24 ± 0.00
COLIEE	0.27 ± 0.00	-
CLERC	0.26 ± 0.01	-

Table 2: Lexical similarity task score for the retrieval task and downstream QA task for each dataset. $LS_{retrieval}$ is the mean lexical similarity score between (query, gold passage) and LS_{QA} is the mean lexical similarity score between (gold passage, answer) over dataset examples, reported with 95% confidence intervals reported. In Appendix G, we report results for the t-test for the difference of means, which provide evidence of statistical significant differences between our datasets and other representative datasets at $\alpha = 0.05$.

To analyze this property of the tasks, we compare the task complexity of our datasets against two popular general domain IR tasks: Natural Questions (NQ) [17], HotpotQA [41], and two existing legal domain IR tasks: COLIEE (2024) [6], CLERC [11]. We use the same metrics to compare task complexity as those used in BIRCO [40], lexical similarity and baseline performance of existing IR methods. We compare the lexical similarities between the (query, gold passage) pair and (gold passage, answer) pair for each dataset example.

We use TF-IDF cosine similarity as the lexical similarity metric because it is a closely related metric to BM25, a strong lexical baseline ranking function for retrieval [30]. For NQ and HotpotQA, we report metrics over the BEIR benchmark test sets [35], since these subsets are commonly used to evaluate retrieval performance, the train set of COLIEE (2024)[4], since the test set labels are not released, and the CLERC test set.[5]

Figure 2 shows that the lexical similarities between query and gold passage for NQ, HotpotQA, COLIEE, and CLERC are distributed normally around a mean of 0.25 - 0.27 (Table 2), while those distributions for Bar Exam QA and Housing Statute QA are heavily skewed towards similarities < 0.10, with mean similarities of 0.07 and 0.08 (Table 2). Lexical similarities are also lower for the gold passages and answers in our tasks, since additional inference is typically needed to conclude the correct legal outcome in the answer from the gold passage rules and the facts in the query.[6]

[4]We report metrics on Task 1.1, the retrieval task derived from Canadian case law, since these laws were originally written in English, though we find that the metrics on Task 2.1 on the Japanese Civil Code (translated to English) are similar.

[5]CLERC presents two settings for gold passage selection. The reference passage is either (1) the full case text of the central citation in the query (document) or (2) a set of sampled analysis paragraphs from the case (paragraph). Since it is not clear that an arbitrary section of the full case text supports the specific citing context in a given query, we report scores for the paragraph with maximal lexical similarity.

[6]We note that for Housing Statute QA, question types with categorical answers are transformed to Yes/No answers to standardize downstream evaluation, so the lexical similarity of the original answer is likely higher for this subset of questions.
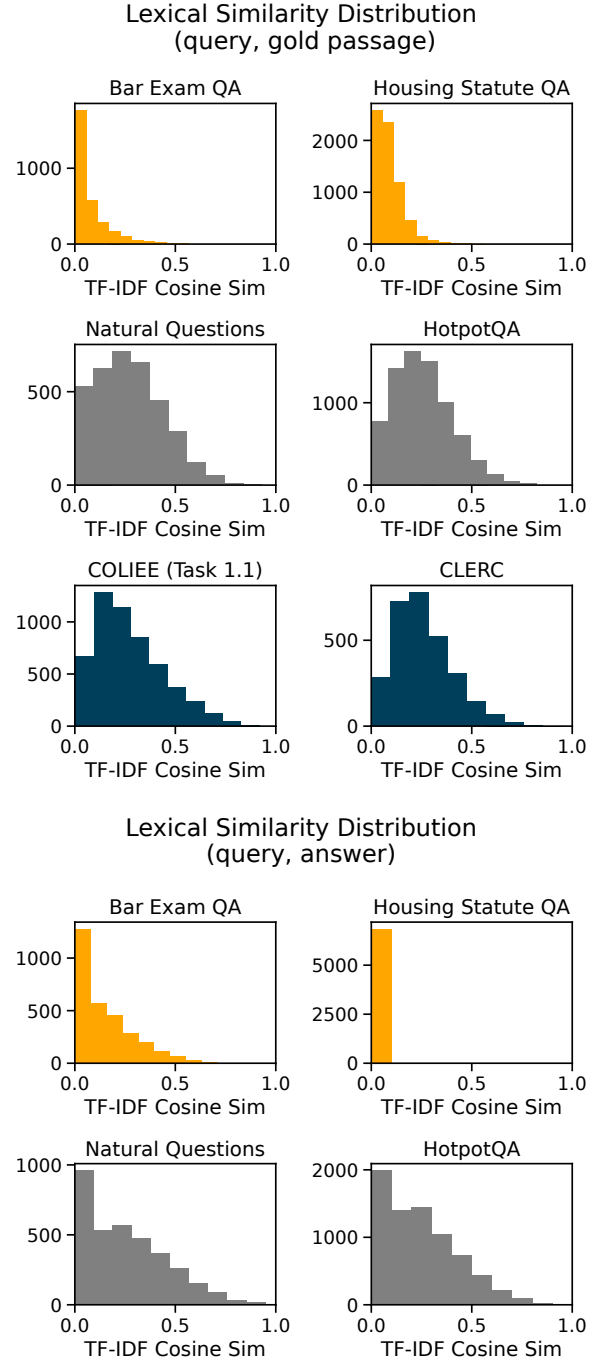
Figure 2: Histograms of the example lexical similarity of (query, gold passage) and (gold passage, answer) over the following datasets: Bar Exam QA and Housing Statute QA (row 1), NQ and HotpotQA (row 2), COLIEE and CLERC (row 3). In Appendix G, we report results for the Kolmogorov-Smirnov test for distributional equivalence between the task similarity distributions, which provide evidence of statistical significant differences between our datasets and other representative datasets at $\alpha = 0.05$.

Dataset	BM25	E5-large-v2
Bar Exam QA	5.03	7.00
Housing Statute QA (lower)	18.3	24.4
Housing Statute QA (upper)	40.8	50.6
Natural Questions	40.4	68.7
HotpotQA	32.7	56.2
COLIEE	38.1	32.7
CLERC	11.8	6.80

Table 3: Baseline retrieval performance (Recall@10) of BM25 (lexical) and E5-large-v2 (dense) retrieval methods on Bar Exam QA (aggregate), Housing Statute QA, NQ, HotpotQA, COLIEE, and CLERC. We report the recall lower/upper bound for Housing Statute QA, see Section 5.2 for details.

5 Evaluation

5.1 Baseline Retrievers

We evaluate a number of baselines, including BM25 [30], which has been shown to be a robust lexical retrieval baseline [20, 31, 35], and the E5 family of retrieval models: E5-small-v2, E5-base-v2, E5-large-v2, E5-mistral-7b-instruct, a series of dense embedding models available in a range of sizes that have been shown to perform well on a broad suite of tasks and complex retrieval tasks in particular [37, 38, 40]. The E5 models are trained from MiniLM [39], BERT base and large (uncased) [5], and Mistral 7B Instruct [14] models.

5.2 Experimental Setup

For Bar Exam QA, we evaluate retrieval performance on the full passage corpus. In the body of the paper, we report results for the aggregate dataset. We report disaggregated results for the Historical MBE and Barbri subsets in Appendix H.

For Housing Statute QA, the dataset includes information about the jurisdiction (U.S. state or territory) of each query and passage, so for each query, we retrieve from a candidate passage pool of the statute passages for the given jurisdiction. The candidate passage pools for each jurisdiction range in size from 10,676 to 155,974 passages. Due to the transformation of the original database questions to Y/N questions described in Section 3.2, not all of the gold passages for the original question may be relevant to each Y/N question. However, we show in Figure 1 that the vast majority of the transformed examples have only 1-2 gold passage labels. We report recall as the retrieval of at least one gold passage for a given query (upper bound) in Results (Section 6). We include full retrieval results computing recall as the retrieval of all the gold passages for a given query (lower bound) in Appendix H.

We also evaluate the comparison tasks on the same baselines.

5.3 Query Expansion

As discussed in Jia et al. [13], Wang et al. [40], simple retrieval methods can fail to capture the correct search intent or task objective when the retrieval request itself requires reasoning—often the case in legal retrieval-augmented QA. However, recent work on query expansion [12, 19, 40] may provide some path forward. To that end, we test several retrieval methods for query expansion in addition to baseline retrieval methods. We use GPT-3.5 (gpt-3.5-turbo-0613) as the generative model for our query expansion experiments.[7]

Paraphrasing. We evaluate query expansion using a prompt to the generative model to paraphrase the query. As illustrated in Table 7, the queries for Bar Exam QA are often quite long. We use this prompting method to study whether simplifying the language in the query helps the retriever. The queries in Housing Statute QA are short, so we do not test this query expansion method for that dataset.

Chain-of-Thought (CoT).. Jagerman et al. [12] use CoT prompting to expand user queries and find performance increases on BEIR [35], we compare against their method. We also evaluate the comparison tasks on this reasoning query expansion method.

Structured Legal Reasoning. We build on these prior approaches and also test a modified query expansion method tailored to the legal setting. Our query expansion method prompts a generative model to perform structured reasoning about the relevant higher-order knowledge hierarchy of the legal task (e.g., the higher-level rules implicated by the query facts) and expands the query with the generated reasoning rollout. The closest prompt-based query expansion approach to this is that of Jagerman et al. [12], however, we note that ours encodes the legal reasoning process by explicitly prompting the generative model to perform legal issue spotting and brainstorm potential legal rules that address the issue. In some ways, this is also related to the prompt-based task-specific re-ranking method by Wang et al. [40], since it adds task-specific prompting and domain knowledge to the retrieval mechanism, but their approach focuses on re-ranking rather than query expansion.

The exact prompts for the query expansion methods on Bar Exam QA and Housing Statute QA are available in Appendix E. In Appendix F, we show an example of the generated query expansion for the same question with the different prompting methods, to illustrate how the structured reasoning rollouts encode the implicit steps required for the legal retrieval tasks by capturing the latent issues and enumerating potential rules addressing the issues that match the language of the statements of law (or primary sources of law) in the passage corpora.

5.4 Retrieval-Augmented Question Answering

We evaluate downstream QA performance for Llama 3 8B Instruct[8] and GPT-4o-mini (gpt-4o-mini-2024-07-18) on Bar Exam QA and Housing Statute QA. We evaluate baseline performance with no passage, performance with retrieved passages using each baseline retriever (with and without each query expansion method), performance with the generative reasoning rollout from the structured legal reasoning query expansion method as a pseudo-passage, and performance with the annotated gold passages.

For Llama 3 8B Instruct, we predict the answer by taking the maximum likelihood prediction over the answer choice letters (e.g., A, B, C, D for Bar Exam QA, Y or N for Housing Statute QA).[9] For

---

[7]We set maximum length = 1024 and temperature = 1 and use the default hyperparameters otherwise (top p = 1, frequency penalty = 0, presence penalty = 0, best of = 1).

[8]https://ai.meta.com/blog/meta-llama-3/

[9]As in the implementation here: https://github.com/artidoro/qlora/blob/main/qlora.py
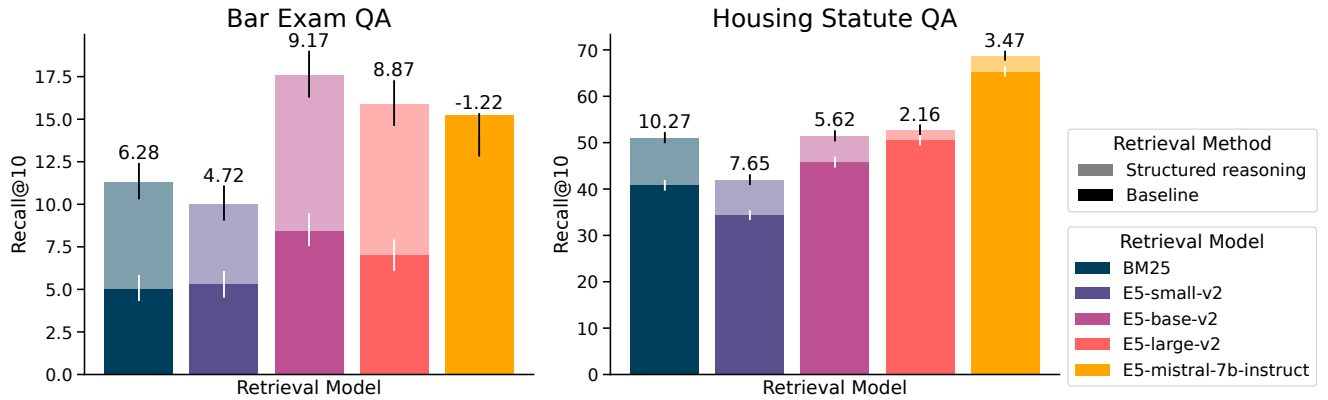
Figure 3: Recall of baseline and structured reasoning rollout query expansion retrieval for lexical (BM25) and dense models (E5 family), evaluated on our legal retrieval benchmark tasks. The gain in Recall@10 with the structured reasoning rollout method is labeled above each bar. 95% confidence intervals are estimated with a percentile bootstrap ($n = 1000$). For Housing Statute QA, recall is reported as retrieval of at least one gold passage (upper bound).

GPT-4o-mini, we predict the answer with an open-ended generative setup, using the default hyperparameters set by the API.

6 Results

Bar Exam QA and Housing Statute QA are challenging for baseline retrievers, especially lexical retrievers like BM25 and smaller models. As noted by Wang et al. [40], it is difficult to directly compare retrieval metrics across datasets due to differences in the relevance scale and number of relevant passages per query across tasks, but we report baseline retrieval performance on our tasks and the comparison tasks in Table 3 and Appendix H. In particular, BM25, a strong lexical baseline [30], and E5-large-v2, a dense embedding retrieval model that BIRCO evaluation focuses on due to its strong performance on complex tasks [37, 40], achieve significantly lower recall on Bar Exam QA than other tasks. In comparison, BM25 performs well for NQ and HotpotQA and can be a surprisingly strong baseline for some other legal IR tasks like COLIEE, due to high lexical similarity between queries and gold passages (Table 2).[10]

Query expansion, particularly our structured query expansion, helps lexical retrievers and small models recover some performance on these legal tasks. Figure 3 shows baseline retrieval performance without query expansion and retrieval performance with structured legal reasoning query expansion. For our legal tasks, characterized by reasoning-focused retrieval task objectives and low (query, relevant passage) lexical similarity, our method of generative query expansion achieves statistically significantly improvement on baseline retrieval performance across all models evaluated, with the largest gains in recall for the more lexically-focused retrieval models. On Bar Exam QA, the gain in Recall@10 for structured reasoning query expansion over baseline retrieval is 6.28 ± 0.99 for BM25 and 8.86 ± 1.16 for E5-large-v2. On

Housing Statute QA, the gain in Recall@10 for structured reasoning query expansion over baseline retrieval is 10.27 ± 1.08 for BM25 and 2.16 ± 1.15 for E5-large-v2. The gains in performance are statistically significant for BM25 and the three smaller E5 models. We do not observe gains for E5-mistral-7b-instruct, but this model is also already highly performant at baseline. We believe this is likely because E5-mistral-7b-instruct is fine-tuned on instruction-following data for retrieval tasks [38], so it may have greater learned retrieval task objective awareness than BM25 or smaller dense embedding models.

Structured reasoning rollouts outperform other prompting techniques for generative query expansion that increase the verbosity of the query. Paraphrasing does not improve and can hurt retrieval performance slightly compared to baseline on Bar Exam QA, suggesting that summarizing long queries and introducing synonymous language may not be sufficient for improving retrieval performance on legal retrieval tasks that necessitate additional reasoning about the query.

We find that CoT reasoning query expansion improves retrieval performance over baseline on all models for Bar Exam QA. For Bar Exam QA, where CoT is more effective, the difference between CoT and structured reasoning is smaller, but still significant for most models at Recall@10 (3.09 ± 1.15 for E5-large-v2). For Housing Statute QA, where CoT is less effective, the difference beween CoT and structured reasoning in larger and significant for all models at Recall@10 (8.97 ± 1.01 for E5-large-v2).

For datasets with higher lexical similarity between questions and gold passages, reasoning rollouts for query expansion are less helpful. We hypothesize that this is because less reasoning is required to complete the retrieval task, so existing retrieval models at baseline already achieve strong performance. For Bar Exam QA, the gain in Recall@10 for CoT over baseline retrieval for E5-base-v2 is 7.37 ± 1.15, while on NQ and HotpotQA, the difference is 1.59 ± 1.45 and −2.79 ± 1.12 respectively. This suggests the expected gain of generative reasoning rollouts for query expansion on retrieval tasks

---

[10]For CLERC, we evaluate the document-level setting. In general, in a legal citation, the full cited case document is not necessarily relevant to the citing context. Citing contexts typically refer to a specific section of the cited case as support. Baseline models would likely yield higher performance on more granular section-level annotations.

may be greater for low lexical similarity tasks compared to high lexical similarity tasks. For some high lexical similarity retrieval benchmarks, such as NQ, current state of the art retrieval models at the ~100-300M parameter size approach performance saturation on the benchmark. Retrieval tasks with lower lexical similarity that necessitate greater reasoning, such as ours, are more challenging for these models, and we observe the greatest gains from appending more reasoning tokens through query expansion in these cases. Additionally, performance may depend on the quality of the generative reasoning rollout. Though we use one generative model, GPT-3.5, in our experiments, we expect that improvement will be correlated with the quality of the generative model used for reasoning rollout.

Hard retrieval task examples help distinguish more capable retrieval models. Figure 4 shows the relationship between baseline retrieval performance and example query and gold passage lexical similarity for the Housing Statute QA dataset, across the five retrieval models. We observe that for examples with high query and gold passage similarity, models perform similarly well. The E5-mistral-7b-instruct model outperforms the other models most significantly on the set of hard examples with low query and gold passage similarity. The relationship illustrates the importance of more complex retrieval tasks for benchmarking retrieval model performance.
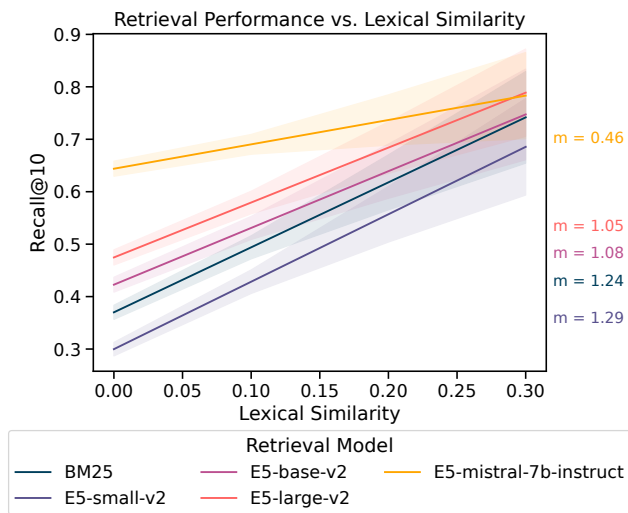

Figure 4: Baseline retrieval performance vs. lexical similarity (query, gold passage) line of best fit over Housing Statute QA task. Recall@10 is averaged over examples bucketed by intervals of lexical similarity scores (bucket sizes of 0.1). 95% confidence intervals are estimated over each bucket. m is the slope of the line of best fit.

Improvement in retrieval performance (Recall@10) doesn't always translate to significant downstream improvements in QA. One challenge is that the retriever must be able to reason about which passages to retrieve, but the downstream LLM must also be able to reason about the retrieved passages. We report full results for downstream evaluation with no passage, retrieved passages, the generative reasoning rollout from the structured reasoning query

expansion method as a pseudo-passage, and the gold passage on Llama 3 8B Instruct and GPT-4o-mini in Appendix I. We find that improvements are upper-bounded by how well models can make use of the gold passage, with only a 20% gain for Llama 3 8B Instruct. However, because the maximum improvement from finding the optimal passage is 20%, even a 10% gain on retrieval can only lead to a 2% improvement on downstream task performance in theory, consistent with the trends that we see (Appendix I).

On GPT-4o-mini (gpt-4o-mini-2024-07-18), we find that the model still struggles to reason over and apply both the retrieved and gold passages to the more challenging Bar Exam QA questions. Using the generative reasoning rollout as a pseudo-passage seems to confuse the model, resulting in a reduction in accuracy, likely from distractors in the generated information. On the other hand, GPT-4o-mini achieve significant improvement in accuracy with retrieved passages over no passage on the downstream task for Housing Statute QA (23.53 percentage point gain). On this case, the generative reasoning rollout as a pseudo-passage also helps improve downstream QA performance, but not as much as using the retrieved passage (68.51% vs. 71.71% accuracy). This suggests that a more capable downstream reasoning model can be helpful for evaluating the utility of retrieved passages to difficult reasoning QA tasks and distinguishing between the quality of various passage sources.

To improve legal retrieval augmented LLMs, future work should focus on improving the reasoning abilities of retrievers as well as the ability of downstream models to reason about retrieved passages.

7 Conclusion

In this work, we introduce two new benchmark datasets for evaluating retrieval-augmented question answering in the legal setting: Bar Exam QA and Housing Statute QA. This benchmark provides ~10K paired, query, gold passage, answer examples, with high-quality, human-annotated gold passages. These datasets contain substantive legal questions as queries and supporting law as passages, simulating reasoning-intensive real-world legal retrieval tasks.

We note that Bar Exam QA and Housing Statute QA do not represent the full distribution of legal questions legal practitioners are likely to encounter in practice, since they cover only the areas tested on the Bar Exam and statutory housing law. However, they provide a closer setting to real-world legal tasks than many other legal retrieval datasets. Unlike legal retrieval datasets with extractive constructions, our query, gold passage pairs were hand-annotated by law students, who were instructed to use legal research tools to find supporting law for a given legal question justifying the answer.

Our benchmarks serve to help researchers, practioners, and policymakers better understand the suitability of retrieval approaches for different legal retrieval tasks over time, as models' performance on general-domain retrieval benchmarks don't necessarily appear to generalize well to law. In our evaluations, we show that the tasks are challenging for lexically-focused retrievers, but generative query expansion techniques that roll out reasoning can help improve retrieval performance. These findings suggest that retrievers must themselves be reasoners too. And that certain legal tasks may be particularly well suited to exposing limitations of current

retrieval models on reasoning-intensive retrieval tasks. This conclusion comports with discussions among legal scholars that determining what law is relevant to addressing a legal question is itself a nontrivial problem and is a separate reasoning skill from the reasoning skills required to apply relevant law to novel scenarios [3]. We hope that our datasets and evaluations can serve as a resource for future work on reasoning-focused retrieval-augmented LLM tasks.

Acknowledgments

This work is dedicated to Andrea Vallebueno, in loving memory. Andrea was a dear friend and labmate. She had a brilliant, warm spirit with a special gift for research and teaching others. Her light overflowed on to each person in her life and inspired so so many, close and far. We remember and hope to carry on the legacy of her life, the dignity and respect with which she treated every person she encountered, her welcoming and inclusive nature, and her passion for her research on computational methods for addressing socially impactful problems.

References

[1] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).

[2] Ilias Chalkidis, Manos Fergadiotis, Nikolaos Manginas, Eva Katakalou, and Prodromos Malakasiotis. 2021. Regulatory Compliance through Doc2Doc Information Retrieval: A case study in EU/UK legislation where text similarity has limitations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty (Eds.). Association for Computational Linguistics, Online, 3498–3511. https://doi.org/10.18653/v1/2021.eacl-main.305

[3] Andrea Anne Curcio, Carol L Chomsky, and Eileen R Kaufman. 2018. How to Build a Better Bar Exam. *New York State Bar Association Journal* (2018), 37–41.

[4] Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. 2024. Large legal fictions: Profiling legal hallucinations in large language models. *arXiv preprint arXiv:2401.01301* (2024).

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[6] Randy Goebel, Yoshinobu Kano, Mi-Young Kim, Juliano Rabelo, Ken Satoh, and Masaharu Yoshioka. 2024. Overview of Benchmark Datasets and Methods for the Legal Information Extraction/Entailment Competition (COLIEE) 2024. In *New Frontiers in Artificial Intelligence*, Toyotaro Suzumura and Mayumi Bono (Eds.). Springer Nature Singapore, Singapore, 109–124.

[7] Matthias Grabmair, Kevin D. Ashley, Ran Chen, Preethi Sureshkumar, Chen Wang, Eric Nyberg, and Vern R. Walker. 2015. Introducing LUIMA: an experiment in legal conceptual retrieval of vaccine injury decisions using a UIMA type system and tools. In *Proceedings of the 15th International Conference on Artificial Intelligence and Law* (San Diego, California) *(ICAIL '15)*. Association for Computing Machinery, New York, NY, USA, 69–78. https://doi.org/10.1145/2746090.2746096

[8] Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, et al. 2024. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems* 36 (2024).

[9] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. *Proceedings of the International Conference on Learning Representations (ICLR)* (2021).

[10] Nils Holzenberger, Andrew Blair-Stanek, and Benjamin Van Durme. 2020. A dataset for statutory reasoning in tax law entailment and question answering. *arXiv preprint arXiv:2005.05257* (2020).

[11] Abe Bohan Hou, Orion Weller, Guanghui Qin, Eugene Yang, Dawn Lawrie, Nils Holzenberger, Andrew Blair-Stanek, and Benjamin Van Durme. 2024. CLERC: A Dataset for Legal Case Retrieval and Retrieval-Augmented Analysis Generation. *arXiv preprint arXiv:2406.17186* (2024).

[12] Rolf Jagerman, Honglei Zhuang, Zhen Qin, Xuanhui Wang, and Michael Bendersky. 2023. Query expansion by prompting large language models. *arXiv preprint arXiv:2305.03653* (2023).

[13] Pengyue Jia, Yiding Liu, Xiangyu Zhao, Xiaopeng Li, Changying Hao, Shuaiqiang Wang, and Dawei Yin. 2023. Mill: Mutual verification with large language models for zero-shot query expansion. *arXiv preprint arXiv:2310.19056* (2023).

[14] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825* (2023).

[15] Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Vancouver, Canada.

[16] Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. 2024. Gpt-4 passes the bar exam. *Philosophical Transactions of the Royal Society A* 382, 2270 (2024), 20230254.

[17] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: a Benchmark for Question Answering Research. *Transactions of the Association of Computational Linguistics* (2019).

[18] Angeliki Lazaridou, Adhi Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Tomas Kocisky, Sebastian Ruder, Dani Yogatama, Kris Cao, Susannah Young, and Phil Blunsom. 2021. Mind the Gap: Assessing Temporal Generalization in Neural Language Models. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 29348–29363. https://proceedings.neurips.cc/paper_files/paper/2021/file/f5bf0ba0a17ef18f9607774722f5698c-Paper.pdf

[19] Yibin Lei, Yu Cao, Tianyi Zhou, Tao Shen, and Andrew Yates. 2024. Corpus-Steered Query Expansion with Large Language Models. *arXiv preprint arXiv:2402.18031* (2024).

[20] Jimmy Lin. 2019. The Neural Hype and Comparisons Against Weak Baselines. *SIGIR Forum* 52, 2 (jan 2019), 40–51. https://doi.org/10.1145/3308774.3308781

[21] Antoine Louis and Gerasimos Spanakis. 2022. A Statutory Article Retrieval Dataset in French. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Dublin, Ireland, 6789–6803. https://aclanthology.org/2022.acl-long.468

[22] Antoine Louis, Gijs van Dijck, and Gerasimos Spanakis. 2024. Interpretable long-form legal question answering with retrieval-augmented large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 22266–22275.

[23] LSC. 2021. Eviction Laws Database: Local Dataset. Prepared by the Center for Public Health Law Research at Temple University's Beasley School of Law for Legal Services Corporation. https://www.lsc.gov/initiatives/effect-state-local-laws-evictions/lsc-eviction-laws-database.

[24] Megan Ma, Aparna Sinha, Ankit Tandon, and Jennifer Richards. 2024. *Generative AI Legal Landscape 2024*. Technical Report. Technical report.

[25] 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. arXiv:2405.20362

[26] Robert Mahari, Dominik Stammbach, Elliott Ash, and AlexSandy' Pentland. 2023. LePaRD: A Large-Scale Dataset of Judges Citing Precedents. *arXiv preprint arXiv:2311.09356* (2023).

[27] Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. WWW'18 Open Challenge: Financial Opinion Mining and Question Answering. In *Companion Proceedings of the The Web Conference 2018* (, Lyon, France,) *(WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1941–1942. https://doi.org/10.1145/3184558.3192301

[28] Timothy McFarlin. 2023. A More Realistic Bar Exam Will Benefit Legal Education. *The Bar Examiner* 92, 2 (2023).

[29] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. (2019).

[30] Stephen E Robertson and K Sparck Jones. 1976. Relevance weighting of search terms. *Journal of the American Society for Information science* 27, 3 (1976), 129–146.

[31] Guilherme Moraes Rosa, Ruan Chaves Rodrigues, Roberto Lotufo, and Rodrigo Nogueira. 2021. Yes, bm25 is a strong baseline for legal case retrieval. *arXiv preprint arXiv:2105.05686* (2021).

[32] 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* (2024).

[33] Jaromír Šavelka and Kevin D Ashley. 2022. Legal information retrieval for understanding statutory terms. *Artificial Intelligence and Law* (2022), 1–45.

[34] Hongjin Su, Howard Yen, Mengzhou Xia, Weijia Shi, Niklas Muennighoff, Han-yu Wang, Haisu Liu, Quan Shi, Zachary S Siegel, Michael Tang, et al. 2024. BRIGHT: A Realistic and Challenging Benchmark for Reasoning-Intensive Retrieval. *arXiv preprint arXiv:2407.12883* (2024).

[35] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*. https://openreview.net/forum?id=wCu6T5xFjeJ

[36] Santosh T.y.s.s., 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)*, Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (Eds.). ELRA and ICCL, Torino, Italia, 5473–5483. https://aclanthology.org/2024.lrec-main.486

[37] 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 preprint arXiv:2212.03533* (2022).

[38] 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* (2023).

[39] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 5776–5788. https://proceedings.neurips.cc/paper_files/paper/2020/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[40] Xiaoyue Wang, Jianyou Wang, Weili Cao, Kaicheng Wang, Ramamohan Paturi, and Leon Bergen. 2024. BIRCO: A Benchmark of Information Retrieval Tasks with Complex Objectives. *arXiv preprint arXiv:2402.14151* (2024).

[41] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[42] Lucia Zheng, Neel Guha, Brandon R Anderson, Peter Henderson, and Daniel E Ho. 2021. When does pretraining help? assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings. In *Proceedings of the eighteenth international conference on artificial intelligence and law*. 159–168.

[43] Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. JEC-QA: a legal-domain question answering dataset. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 9701–9708.

A Bar Exam QA Dataset Construction

The gold passage annotation process for Bar Exam QA is modeled off of the legal research process. For each example, the law student has access to the general area of law of the question, the question, the answer choices, and the correct answer. This annotation effort took roughly 9 months to complete.

(1) First, the law students identify the rule of law relevant to the example. They consider the Barbri answer key bank (from the set of examples with explanation passages), bar exam study guides and practice guides (e.g., *The Ultimate Guide to the MBE*), and other secondary sources (e.g., American Law Reports). They identify general rules of law that could help answer the question and a list of general and specific

search terms and legal concepts based on keywords from these secondary sources.

(2) Next, the law students compose a Westlaw Terms and Connector search query for relevant cases stating the rule of law. These search queries are written by hand by the law students, without AI assistance. In Table 4, we provide examples of Terms and Connector search queries constructed by law student annotators in the search process. From the search results, the law students read the descriptions of cases for one that appears on point and review the case headnotes to identify a case with a statement of the rule of law that mirrors the rule of law identified.

(3) Lastly, the law students find a succinct, generalizable statement of the rule of law in the identified case text and annotates the example with this text as the gold passage label.

"due process" /p "termination"
"unreasonable burden on interstate commerce"
controvers! /p "declaratory judgment"
"Presidential pardon power"
"Due process" +5 "balancing"

Table 4: Examples of Westlaw Terms and Connector search queries

B Housing Statute QA Dataset Construction

Table 5 shows a sample of original questions from the LSC Eviction Database [23]. Table 6 shows examples of reformatted original questions to Y/N questions.

C Dataset Release and Licenses

Our datasets are publicly available on HuggingFace at the following links.

- Bar Exam QA: https://huggingface.co/datasets/reglab/barexam_qa
- Housing Statute QA: https://huggingface.co/datasets/reglab/housing_qa

The passage pool comes from 3 sources with permissive licenses: Cornell LII (CC BY-NC-SA 2.5), Case Law (Public Domain), Justia (Public Domain). Our compilation of these sources follows these licenses.

For Bar Exam QA, for the historical MBE subset, we release the queries and gold passages, annotated by our research team, which we believe are a transformative fair use of the queries, and release them under a CC-BY-NC-SA license. We also believe that release of the historical bar exam multiple choice options and answers in full would be a transformative fair use, since it is for public interest educational purposes, unlikely to affect markets for exams (since they are older and no longer for sale), and much of the data can already be found in other fair use compilations like MMLU professional law auxiliary training set [9], Common Crawl, and others. However, in the interest of responsible practice, we release the multiple choice options and answers only to researchers through a gated release mechanism with a restrictive license to the compilation. Some of

these can nonetheless already be found in the MMLU auxiliary training set [9] released under an MIT license, or can be acquired from publicly available sources on the web, available in Common-Crawl and Archive.org. For the Barbri subset, we do not release the dataset due to copyright concerns and treat the subset as a private, held-out test set, which we report separate evaluation results on.

For Housing Statute QA, we release under CC-BY-SA. LSC allows download and redistribution.

D Dataset Examples

Tables 7, 8, 9, 10, 11, and 12 show representative examples from Bar Exam QA, Housing Statute QA, Natural Questions, HotpotQA, COLIEE (Task 1.1), and CLERC.

E Query Expansion Prompts

Table 13 shows the query expansion prompts for Bar Exam QA. Table 14 shows the query expansion prompts for Housing Statute QA.

F Structured Reasoning Prompt Query Expansion Examples

For the question from Bar Exam QA in Table 15, we provide the query expansions with the paraphrasing prompt, chain-of-thought prompt, and structured reasoning prompt, and the gold passage for comparison in Table 16, to illustrate how the structured reasoning prompt expansion encodes implicit retrieval task steps and captures latent legal issues.

G Lexical Similarity Statistical Test Results

In Tables 17, 18, 19, 20, we report the statistical test results comparing the lexical similarity distribution over Bar Exam QA and Housing Statute QA to other popular general and legal domain retrieval tasks.

H Retrieval Results

For Bar Exam QA, Table 21, 22, and 23 report full retrieval performance evaluation results for the aggregate, historical MBE, and Barbri subsets.

For Housing Statute QA, Table 24 reports full retrieval performance evaluation results. Since we transform the original questions to binary classification (Y/N) questions for the downstream task, some gold statute passages are relevant for the original question, but not the derived Y/N question. In Table 24, recall is computed as the retrieval of at least one gold passage (upper bound). For completeness, we also provide Table 25, where recall is computed as the retrieval of all gold passages (lower bound); as an example has a maximum of 10 gold passages, we report this value from retrieval depth of $k = 10$.

For NQ, Table 26 reports retrieval performance evaluation results.

For HotpotQA, Table 27 reports retrieval performance evaluation results. We compute recall as the retrieval of the two gold passages, since both gold passages are required to answer each question; we report this value from retrieval depth of $k = 2$.

For COLIEE, Table 28 reports retrieval performance results on Task 1.1 For CLERC, Table 29 reports retrieval performance results for the document setting.

I Downstream QA Results

Tables 30, 31, and 33 show downstream QA results for Llama 3 8B Instruct and LLama 3 70B Instruct on Bar Exam QA and for LLama 3 8B Instruct on Housing Statute QA. Tables 32 and 34 show downstream QA results for GPT-4o-mini (gpt-4o-mini-2024-07-18) on Bar Exam QA and Housing Statute QA.

Can a landlord evict a tenant for endangering property? This includes where the law refers to situations that could result in damage to the property; or a pet capable of causing damage to persons or property.
Secondary methods of service are defined as those methods that may be used if the primary method is unsuccessful. Are certified mail and regular mail a permitted secondary methods of service for an eviction action?
Are court records for eviction cases by default publicly available?
Does the law require landlords to provide information on how to cure when giving tenants notice to vacate the property?
Can a landlord evict a tenant for committing or failing to dispose of waste?

Table 5: Sample questions for Housing Statutes QA

Original question	Original answer choices	Reformatted Y/N questions
What type(s) of landlord(s) does state/territory eviction law explicitly regulate?	Residential landlords generally	Are residential landlords explicitly regulated by eviction law?
	Mobile/manufactured home landlords	Are mobile/manufactured home landlords explicitly regulated by eviction law?
	Corporate landlords	Are corporate landlords explicitly regulated by eviction law?
	Floating home landlords	Are floating home landlords explicitly regulated by eviction law?
	Landlords with minimal rental properties	Are landlords with minimal rental properties explicitly regulated by eviction law?

Table 6: Example of reformatted original questions to Y/N questions

Query	Under an aid-to-education statute passed by the state legislature a few years ago, many private schools receive state benefits. One private school receives: (i) free textbooks from the state, (ii) an exemption from state taxes, and (iii) 20% of its operating budget in the form of state grants. The remaining 80% of the school's-budget is covered by tuition fees and by donations from alumni and others. The school is licensed by the state, but the state has no requirement for certification and licensure of teachers in private schools.
A teacher was hired to teach history at the school. The teacher was given the standard three-year contract given to teachers in their first stint at the school. In the fall term of his second year, the teacher gave a lecture to his students criticizing the school's use of school uniforms and encouraging the students to organize a protest against the uniform policy. After the speech, the teacher was called to the administrative office by the headmaster and fired on the spot, despite the teacher's protests that he had almost two years left on his contract. The teacher requested a hearing and was told to leave the premises of the school immediately.
If the teacher files suit in federal district court alleging that his constitutional rights have been violated, the teacher will:
Gold Passage	The Fourteenth Amendment Due Process Clause, which makes many of the provisions of the Bill of Rights applicable to the states, does not apply to purely private conduct that interferes with these rights. Thus, unless the private individual (i) was performing exclusively public functions, or (ii) took actions with significant state involvement, the individual's action is not unconstitutional. In this case, the school is a private institution performing a function-education-that has never been considered to be an exclusively public function. [See Pierce v. Society of Sisters (1925)] Furthermore, its licensing by the state and receipt of state funds do not constitute significant state involvement with regard to its personnel matters. [See Rendell-Baker v. Kohn (1982)]
Answer	Fail, because assistance and involvement by the state did not result in the private school's action being conduct by the state.

Table 7: Example from Bar Exam QA

Jurisdiction	Alabama
Query	Does the law specify rebuttals available to tenants subject to eviction proceedings?
Gold Passage(s)	(b) If a landlord acts in violation of subsection (a), the tenant is entitled to the remedies provided in Section 35-9A-407 and has a defense in any retaliatory action against the tenant for possession.
	(a) In an action for possession or in an action for rent when the tenant is in possession, the tenant may counterclaim for any amount the tenant may recover under the rental agreement or this chapter. It is in the court's discretion whether the tenant is to remain in possession. The tenant shall pay into court rent accrued and thereafter accruing as it comes due. The court shall determine the amount due to each party. The party to whom a net amount is owed shall be paid first from the money paid into court, and the balance by the other party. If no rent remains due after application of this section, judgment shall be entered for the tenant in the action for possession. If the defense or counterclaim by the tenant is without merit and is not raised in good faith, the landlord may recover reasonable attorney's fees.
	Acceptance of rent with knowledge of a default by the tenant or acceptance of performance by the tenant that varies from the terms of the rental agreement constitutes a waiver of the landlord's right to terminate the rental agreement for that breach, unless otherwise agreed after the breach has occurred.
	(b) If contrary to the rental agreement or Section 35-9A-204, after receiving notice of the breach from the tenant, the landlord willfully or negligently fails to promptly make available heat, running water, hot water, electric, gas, or other essential service, the tenant may:(1) send a written notice specifying the date of termination not less than 14 days after receipt of notice and upon vacation of the premises, the rental agreement shall be rightfully terminated without further obligation or penalty. If the rental agreement is terminated pursuant to this section, the landlord shall return all security recoverable by the tenant under Section 35-9A-201 and all unearned prepaid rent; or(2) recover damages based upon the diminution in the fair rental value of the dwelling unit.
	(a) Except as provided in this chapter, if there is a material noncompliance by the landlord with the rental agreement or a noncompliance with Section 35-9A-204 materially affecting health and safety, the tenant may deliver a written notice to the landlord specifying the acts and omissions constituting the breach and that the rental agreement will terminate upon a date not less than 14 days after receipt of the notice if the breach is not remedied within that period, and the rental agreement shall terminate as provided in the notice subject to the following:
Answer	Yes

Table 8: Example from Housing Statute QA

Query	Where is the bowling hall of fame located?
Gold Passage	The World Bowling Writers (WBW) International Bowling Hall of Fame was established in 1993 and is located in the International Bowling Museum and Hall of Fame , on the International Bowling Campus in Arlington , Texas.
Answer	Arlington , Texas

Table 9: Example from Natural Questions [17]

Query	Peter Marc Jacobson is best known as the co-creator of the popular sitcom "The Nanny", which he created and wrote with his then wife an actress born in which year?
Gold Passage (s)	Peter Marc Jacobson (born October 27, 1957) is an American television writer, director and producer, and actor. He is best known as the co-creator of the popular sitcom "The Nanny", which he created and wrote with his then wife actress Fran Drescher, who was the star of the series. He was often credited as Peter Marc in his early acting roles.
	Francine Joy "Fran" Drescher (born September 30, 1957) is an American actress and activist. She is best known for her role as Fran Fine in the hit TV series "The Nanny" (1993–99), and for her nasal voice and thick New York accent.
Answer	1957

Table 10: Example from HotpotQA [41]

Query	Summary:
	The applicant, a citizen of Afghanistan, claimed refugee protection. In 2010, he was accepted for resettlement to Canada as a member of a humanitarian-protected person abroad class (country of asylum). In 2013, the Minister of Public Safety and Emergency Preparedness and the Minister of Citizenship and Immigration applied for an order that the applicant's refugee status cease on the basis that he had reavailed himself of the protection of his country of nationality (Immigration and Refugee Protection Act, s. 108(1)(a)). The Refugee Protection Division allowed the application. The applicant applied for judicial review.
	The Federal Court dismissed the application and certified the following question: "In a cessation application pursuant to paragraph 108(1)(a) of IRPA, do the same or substantially the same legal considerations, precedents, and analysis apply to persons found to be Convention refugees as to persons found to be in need of protection as members of the Country of asylum class?" ...
Gold Passage(s)	[30] Ample case law from the Immigration Appeal Division in the 1980s was directly concerned with a child's intent in a context where a child's parents had abandoned permanent residence. Although these decisions are not binding on the Court (<FRAGMENT_SUPPRESSED> (QL/Lexis) at paragraph 14), they emphasize the importance of considering the intention of minors when they reach the age to form it, on the basis that they could not form it upon the departure of their parents because of their young age.
	[31] The male applicant was three years old at the time of the initial departure to Mexico and was therefore not able not form an intention to reavail himself of the protection of Mexico. This could have been different at eleven years of age, his age at the time of the hearing. At that point, there should have been further analysis in order to find that an 11-year-old child cannot form an intention that differs from that of his parents.
	[32] However, nothing in the evidence or in the submissions made by the parties makes it possible to determine whether the intention of the child could have been different from that of his mother.
	IX. Conclusion [33] In the circumstances of this case and in light of the foregoing, the Court cannot intervene because the decision does not go beyond the range of reasonableness...
	[22] However, the Court finds that the Board erred in its consideration of the applicant's explanation relating to his business activities in Thailand. As outlined in <FRAGMENT_SUPPRESSED> , a review on the standard of reasonableness is concerned with the "existence of justification, transparency and intelligibility" in the decision. With respect, the Court finds a justification lacking in the present case. It is unclear to the Court why the Board believed that the applicant's explanation with respect to why he obtained a Congolese passport was insufficient. This conclusion may have been open to the Board to make; however, the Court finds it unreasonable that the Board failed to indicate why this explanation was insufficient. If the Board did not believe the applicant's explanation and found him not to be credible then it should have said so. If it had another reason for not finding the explanation sufficient, it should have stated so as well, especially with the type of explanations provided here by the applicant to rebut his presumed intention "to avail himself of the protection of the country of his nationality". [23] True the burden was on the applicant to rebut this presumption, and he tried. But here his explanations as a whole were not discarded by the Board because they were not credible; on the contrary the decision seems to imply that, the simple fact of possessing a Congolese passport that the applicant refused for a very specific reason to return to the Congolese authorities when requested by them to do so, constitutes proof of his intention to reavail himself of the protection of his country of nationality. The Court cannot accept such implied finding in the present affair in view of the inexistence of any credibility finding in the decision with respect to the applicant's explanations.
	[24] For the foregoing reasons the Court finds the Board's decision to be unreasonable.
	[25] The Court agrees with the parties that there is no question of general interest to certify...

Table 11: Example from COLIEE (Task 1.1) [6]

Query	to constitute clear error. See United States v. Sullivan, 75 F.3d 297, 302-03 (7th Cir.1996). Throughout his briefs, Siegler attempts to portray the August 31 letter as a solicitation rather than a threat, in effect trying to challenge his conviction f or violating 18 U.S.C. § 876. By pleading guilty, however, Siegler admitted both of the elements of Count II (mailing a threatening communication). See McCarthy v. Unit ed States, 394 U.S. 459, 466, 89 S.Ct. 1166, 22 L.Ed.2d 418 (1969) ("[A] guilty plea is an admission of all the elements of a formal criminal charge."); United States v . Gilliam, 255 F.3d 428, 433 (7th Cir.2001) (same). In the written plea agreement an d during the plea hearing, Siegler admitted that on August 31, 1999, he wrote and ma iled to Hester a letter threatening Hauger; no more was required for a conviction un der 18 U.S.C. § 876. See REDACTED .C. § 876 requires proof of two elements: (1) a t hreatening communication (2) was sent through the mail); United States v. Khorrami, 895 F.2d 1186, 1192 (7th Cir.1990) (conviction under 18 U.S.C. § 876 does not requir e proof that defendant intended to carry out threat). By admitting that the letter h e sent contained a threat within the meaning of 18 U.S.C. § 876, Siegler waived any subsequent argument about the nature of the threat. See United States v. Newman, 148 F.3d 871, 876 (7th Cir.1998) (defendant's stipulation to conduct in plea agreement conclusively admitted facts and waived subsequent challenge to them). Accordingly, S iegler's argument that the letter did not contain a "true threat" is irrelevant to h is appeal of his sentence. Siegler also argues that because he did not send the lett er to Hauger or directly communicate the threat to her, there
Gold Passage(s)	FLAUM, Circuit Judge. For a period of more than four years, Richard Geisler was involved in a romantic relationship with Tena Camille DeAck len. During this time, the couple shared a joint bank account. Their relationship ended in early 1992, and Geisler thereafter contended tha t DeAcklen improperly withdrew $1,280 of his money from their joint account. DeAck-len refused to repay this money, whereupon Geisler — who is white — began sending racially-charged, threatening letters to DeAck-len — who is African-American. In the end, Geisler sent six of the se hateful letters between September 1994 and January 1996. The district court convicted Geisler of six counts of mailing threatening commu nications with the intent to extort money in violation of 18 U.S.C. § 876. We affirm his convictions. Geisler stipulated at trial that he a uthored the letters that formed the basis for the charged offenses. There was similarly no dispute that he had sent the letters through the mails. Finally, Geisler did not— nor could he — challenge that the threats of injury and death (along with references to his "friends" aff iliated with the Ku Klux Klan who might assist him in carrying out these threats) contained in these letters constituted threats sufficient to trigger § 876. Rather, his challenge on appeal focuses on the fact that DeAcklen did not read all of the threatening letters that he se nt through the mails. Indeed, she testified that she read one letter in January 1995, as well as one or two others (she could not remember precisely), but that she turned over the other letters directly to the FBI without opening them. Geisler contends that, because DeAcklen ne ver received the threats contained in some of his letters, he did not violate § 876 on those counts. This argument reflects a patently inco rrect interpretation of, the requirements of § 876 and our Circuit's precedent, and Geisler recognizes as much. Under thé plain language o f the statute, the Government only needed to prove that Geisler sent a communication through the mails that contained a threat to injure De Acklen; Geisler's proposed "receipt" requirement is nowhere to be found in the statute. For this reason, we have stated repeatedly that the only two elements of a § 876 violation are (1) a threatening communication (2) sent through the mails. See, e.g., United States v. Sulliva n, 75 F.3d 297, 302 (7th Cir.1996) ("The sending of threatening communications is a crime quite apart from any intent to carry out the thre ats."); United States v. Aman, 31 F.3d 550, 551 (7th Cir.1994) (stating that § 876 "prohibits the mailing of threatening communications"); United States v. Johnson, 965 F.2d 460, 467 (7th Cir.1992) (noting that § 876 "simply re-quirfes] that a defendant knowingly cause to be de livered a threatening letter in the U.S. mails"). In light of the plain language of the statute, it is not surprising that other Circuits s hare our view that there are only two required elements of a § 876 violation. See, e.g., United States v. Turner, 960 F.2d 461, 463 n. 2 (5 th Cir.1992); United States v. Davis, 926 F.2d 969, 971 (10th Cir.), cert. denied, 500 U.S. 926, 111 S.Ct. 2036, 114 L.Ed.2d 121 (1991); Un ited States v. Davis, 876 F.2d 71, 73 (9th Cir.), cert. denied, 493 U.S. 866, 110 S.Ct. 188, 107 L.Ed.2d 143 (1989); United States v. Linco ln, 589 F.2d 379, 381 (8th Cir.1979); United States v. Chatman, 584 F.2d 1358, 1361 (4th Cir.1978). We reject Geisler's attempt to create a new element of the offense...

Table 12: Example from CLERC [11]

Paraphrasing	Given a legal question, paraphrase the question in "Paraphrase:".
Chain-of-Thought	Given a legal question in "Question:", answer the question in "Answer:". Explain your reasoning in "Explanation:". Think step by step.
Structured Reasoning	Given a set of facts about a legal scenario in "Question:", identify the key legal issue that arises from the facts and provide the applicable legal rule in "Rule:".

Table 13: Query expansion prompts for Bar Exam QA query expansion

Chain-of-Thought	Consider the housing statute for {jurisdiction} in the year 2021. Given a legal question in "Question:", answer the question in "Answer:". Explain your reasoning in "Explanation:". Think step by step.
Structured Reasoning	Consider the housing statute for {jurisdiction} in the year 2021. The question given in "Question:" is a legal question about housing and eviction law in {jurisdiction}. Provide the applicable legal rule in "Rule:". If you do not know the state law, provide governing rules that address the question under typical eviction law.

Table 14: Query expansion prompts for Housing Statute QA

Question	Pauline, an unmarried female, was prominent in the women's liberation movement. She recently gave birth to a baby and publicly announced that she had no intention of marrying the father or disclosing his identity. The local newspaper, Journal, decided to do a series of articles on Pauline, entitled "The Perils of Pauline." The first article about Pauline discussed her parents. The article correctly stated that Mary, her mother, had died recently and Frank, her father, is still living. The article referred to the fact that at the time of Pauline's birth there were rumors that she had been born six months after the marriage of Mary and Frank, that Frank was not in fact her father, and that a person identified as Albert, who had played minor roles in two motion pictures, was her real father. Albert has lived in retirement for the last ten years. If Pauline asserts a claim based on invasion of privacy against Journal for the statements in the first article about her birth and it is established that the statements are true, the most likely result is that Pauline will

Table 15: Example question from Bar Exam QA for prompt expansion comparison in Table 16.

Paraphrasing Expansion	If Pauline files a lawsuit against the newspaper Journal for invasion of privacy regarding the information published about her birth, and it is proven that the statements made in the article are accurate, what is the most probable outcome for Pauline?
Chain-of-Thought Expansion	Pauline will not succeed in her claim based on invasion of privacy against Journal. In order to establish a claim for invasion of privacy, Pauline would need to show that Journal's statements about her birth were false and that they invaded her reasonable expectation of privacy. However, in this case, it is mentioned that the statements made by Journal in the first article about Pauline's birth are true. It is established that there were rumors about the circumstances of her birth, including the identity of her father. Since the statements are true, Pauline cannot argue that Journal invaded her privacy by publishing this information. The truth of the statements serves as a valid defense for Journal, making it unlikely for Pauline to succeed in her claim.
Structured Reasoning Expansion	The applicable legal rule in this case would likely be related to the tort of invasion of privacy. In the United States, the specific elements and standards for invasion of privacy can vary depending on the jurisdiction. However, a common framework for invasion of privacy claims involves four distinct categories: 1. Intrusion upon seclusion: This refers to the unauthorized invasion into a person's private affairs or physical space in a way that would be highly offensive to a reasonable person. 2. Public disclosure of private facts: This involves the public dissemination of private and confidential information about an individual that would be highly offensive to a reasonable person and is not of legitimate public concern. 3. False light: This occurs when false or misleading information is publicly attributed to an individual, portraying them in a highly offensive and false manner. 4. Appropriation of name or likeness: This refers to the unauthorized use of a person's name, likeness, or identity for commercial purposes, without their consent. Based on the facts provided, it seems that the issue relevant to Pauline's claim against the Journal would fall under the category of ""public disclosure of private facts."" If the statements made in the first article about Pauline's birth are true (i.e., that there were rumors regarding her parentage), it may be more difficult for Pauline to establish a claim for invasion of privacy. The truth of the statements could potentially be a defense for the Journal, as disclosing true information may carry more weight in the context of a public figure like Pauline.
Gold Passage	There are a number of defenses to a public disclosure claim. Below, you'll find explanations of some of the most common defenses: Legitimate public interest: Whether the public has a legitimate interest in the facts-at-issue is a question that depends on the context of the case, and one in which there is no particular formula for the courts to follow. Whether this defense can be effectively asserted will depend largely on whether the person involved has made him or herself - in a temporary newsworthy capacity or a more permanent celebrity capacity - something of a public figure. In such cases, details of their private lives are more likely to be considered items of legitimate public interest. The passage of time may lessen the public interest in a given fact (the newsworthiness of it), which may weaken this defense. Consent: Consent is a total defense. If the plaintiff has consented in some way to the disclosure, whether through a release form or through accepting an interview, then he or she cannot pursue a claim for public disclosure of private fact. Public Record: Matters of public record, such as birth date, military service records, and others, are exempted. The defendant may claim this defense by showing that the disclosed fact was actually a matter of public record. However, it should be noted that, unlike defamation actions, truth is no defense to a claim for public disclosure of private facts. This means that a defendant cannot refute a claim by showing that the disclosed fact was actually true or accurate.

Table 16: Example of query expansions with different prompting methods and gold passage for the question from Bar Exam QA in Table 15. The structured reasoning expansion most clearly identifies the legal issue and statement of the applicable legal rule that resembles the gold passage.

Distribution 1	Distribution 2	Test statistic (D)	p-value
Bar Exam QA	NQ	0.590	<0.001
Bar Exam QA	HotpotQA	0.610	<0.001
Bar Exam QA	COLIEE (Task 1.1)	0.607	<0.001
Bar Exam QA	CLERC	0.630	<0.001
Housing Statute QA	NQ	0.596	<0.001
Housing Statute QA	HotpotQA	0.593	<0.001
Housing Statute QA	COLIEE (Task 1.1)	0.585	<0.001
Housing Statute QA	CLERC	0.614	<0.001

Table 17: Kolmogorov-Smirnov test results comparing the lexical similarity (query, gold passage) distribution of Bar Exam QA and Housing Statute QA to the distribution of other general and legal domain IR tasks.

Distribution 1	Distribution 2	Test statistic (D)	p-value
Bar Exam QA	NQ	0.251	<0.001
Bar Exam QA	HotpotQA	0.249	<0.001
Housing Statute QA	NQ	0.770	<0.001
Housing Statute QA	HotpotQA	0.771	<0.001

Table 18: Kolmogorov-Smirnov test results comparing the lexical similarity (gold passage, answer) distribution of Bar Exam QA and Housing Statute QA to the distribution of other general domain IR tasks.

Distribution 1	Distribution 2	Test statistic (t)	p-value
Bar Exam QA	NQ	-58.1	<0.001
Bar Exam QA	HotpotQA	-63.8	<0.001
Bar Exam QA	COLIEE (Task 1.1)	-60.4	<0.001
Bar Exam QA	CLERC	-60.1	<0.001
Housing Statute QA	NQ	-79.4	<0.001
Housing Statute QA	HotpotQA	-87.7	<0.001
Housing Statute QA	COLIEE (Task 1.1)	-83.6	<0.001
Housing Statute QA	CLERC	-80.5	<0.001

Table 19: t-test results comparing the mean lexical similarity (query, gold passage) of Bar Exam QA and Housing Statute QA to the mean lexical similarity of other general and legal domain IR tasks.

Distribution 1	Distribution 2	Test statistic (t)	p-value
Bar Exam QA	NQ	-23.2	<0.001
Bar Exam QA	HotpotQA	-24.1	<0.001
Housing Statute QA	NQ	-99.0	<0.001
Housing Statute QA	HotpotQA	-102	<0.001

Table 20: t-test results comparing the mean lexical similarity (gold passage, answer) distributions of Bar Exam QA and Housing Statute QA to the mean lexical similarity of other general domain IR tasks.

Method	Recall@1	Recall@10	MRR@10	Recall@100	Recall@1000
BM25					
Baseline	1.83	5.03	2.68	9.99	20.52
Paraphrase	1.77	5.03	2.77	10.60	22.45
CoT	3.70	9.65	5.28	22.18	42.76
Structured reasoning	3.43	11.31	5.51	26.39	47.86
$E5_{small-v2}$					
Baseline	2.51	5.30	3.39	11.38	21.88
Paraphrase	1.90	5.26	2.88	10.50	21.26
CoT	3.09	8.93	4.72	18.48	33.56
Structured reasoning	3.53	10.02	5.25	23.03	38.25
$E5_{base-v2}$					
Baseline	3.33	8.42	4.71	16.34	29.14
Paraphrase	2.89	7.74	4.17	15.15	29.69
CoT	5.88	15.79	8.55	31.32	49.01
Structured reasoning	6.25	17.60	9.52	33.42	50.34
$E5_{large-v2}$					
Baseline	3.13	7.00	4.25	15.35	27.00
Paraphrase	2.48	6.11	3.46	13.49	24.49
CoT	4.82	12.77	6.91	26.80	43.75
Structured reasoning	5.84	15.86	8.69	32.34	49.01
$E5_{mistral-7b}$					
Baseline	5.71	15.25	8.19	34.10	56.28
Paraphrase	4.96	13.08	7.15	31.32	53.06
CoT	3.46	13.32	5.88	35.29	58.51
Structured reasoning	2.81	14.03	5.64	37.40	60.73

Table 21: Retrieval performance on Bar Exam QA, aggregated.

Method	Recall@1	Recall@10	MRR@10	Recall@100	Recall@1000
BM25					
Baseline	0.25	0.75	0.37	2.26	8.79
Paraphrase	0.33	0.84	0.45	2.85	10.71
CoT	0.75	2.68	1.2	8.87	27.53
Structured reasoning	0.59	3.1	1.21	12.55	32.05
E5$_{small-v2}$					
Baseline	0.08	0.59	0.18	2.68	9.29
Paraphrase	0.08	0.92	0.27	3.26	9.54
CoT	0.25	2.34	0.81	8.95	20.5
Structured reasoning	0.42	2.34	0.94	10.96	23.6
E5$_{base-v2}$					
Baseline	0.25	0.84	0.39	3.51	11.21
Paraphrase	0.17	0.75	0.35	3.85	14.39
CoT	0.75	4.1	1.55	13.56	27.7
Structured reasoning	1.0	4.1	1.74	12.8	29.37
E5$_{large-v2}$					
Baseline	0.17	0.92	0.34	4.27	12.3
Paraphrase	0.08	0.84	0.23	3.6	10.88
CoT	0.67	4.18	1.56	14.23	28.7
Structured reasoning	1.34	5.19	2.42	16.49	31.8
E5$_{mistral-7b}$					
Baseline	0.84	3.26	1.45	9.71	26.36
Paraphrase	0.5	1.76	0.84	7.11	18.41
CoT	1.26	5.86	2.39	20.33	42.43
Structured reasoning	0.67	6.95	2.26	23.51	42.85

Table 22: Retrieval performance on Bar Exam QA, disaggregated (Historical MBE subset).

Method	Recall@1	Recall@10	MRR@10	Recall@100	Recall@1000
BM25					
Baseline	2.81	7.66	4.11	14.71	27.49
Paraphrase	2.75	7.05	3.98	14.6	28.21
CoT	5.51	13.88	7.77	30.14	51.24
Structured reasoning	5.18	16.31	8.14	34.55	56.53
E5$_{small-v2}$					
Baseline	3.09	7.66	4.43	14.55	27.38
Paraphrase	2.81	6.94	4.05	14.27	28.48
CoT	4.35	12.12	6.54	23.2	40.22
Structured reasoning	4.9	12.89	7.06	28.21	44.79
E5$_{base-v2}$					
Baseline	3.86	10.3	5.71	20.17	37.47
Paraphrase	3.31	7.99	4.64	17.47	32.45
CoT	7.49	18.62	10.49	38.07	58.46
Structured reasoning	8.1	23.25	12.52	44.35	62.87
E5$_{large-v2}$					
Baseline	4.52	10.03	6.13	20.61	34.99
Paraphrase	3.42	7.49	4.52	15.65	29.2
CoT	6.78	17.63	9.54	33.5	51.85
Structured reasoning	7.16	19.83	10.72	39.23	57.13
E5$_{mistral-7b}$					
Baseline	5.29	11.85	7.05	25.79	41.82
Paraphrase	2.09	5.51	3.05	13.44	26.61
CoT	3.14	11.85	5.38	27.55	47.93
Structured reasoning	1.93	10.96	4.2	28.1	48.54

Table 23: Retrieval performance on Bar Exam QA, disaggregated (Barbri subset).

Method	Recall@1	Recall@10	MRR@10	Recall@100	Recall@1000
BM25					
Baseline	14.72	40.81	21.99	62.41	76.19
CoT	14.07	43.43	22.11	64.22	77.0
Structured reasoning	18.68	51.09	27.74	71.76	81.69
E5$_{small-v2}$					
Baseline	8.97	34.35	15.95	64.42	81.63
CoT	8.86	33.71	15.7	57.9	76.62
Structured reasoning	13.38	42.0	21.34	69.02	83.41
E5$_{base-v2}$					
Baseline	13.56	45.75	22.38	74.61	86.65
CoT	13.24	40.83	21.19	68.89	84.63
Structured reasoning	16.56	51.36	26.45	76.21	87.25
E5$_{large-v2}$					
Baseline	16.08	50.58	26.02	78.83	87.73
CoT	12.91	43.76	21.61	70.67	84.9
Structured reasoning	17.86	52.74	28.01	78.11	87.48
E5$_{mistral-7b}$					
Baseline	25.36	65.31	37.7	84.31	88.87
CoT	26.25	64.54	38.13	82.58	88.68
Structured reasoning	30.21	68.79	42.54	85.04	89.1

Table 24: Retrieval performance on Housing Statute QA. Recall is computed as retrieval of at least one gold passage for a given query (upper bound).

Method	Recall@10	MRR@10	Recall@100	Recall@1000
BM25				
Baseline	18.31	10.35	39.89	59.68
CoT	18.85	9.81	42.04	60.44
Structured reasoning	23.83	13.73	49.53	65.1
E5$_{small-v2}$				
Baseline	13.32	6.53	37.57	66.32
CoT	13.63	7.12	32.23	58.66
Structured reasoning	18.12	9.35	43.11	68.52
E5$_{base-v2}$				
Baseline	20.3	10.4	51.61	75.53
CoT	17.57	9.54	44.89	70.22
Structured reasoning	23.97	12.25	53.67	75.75
E5$_{large-v2}$				
Baseline	24.4	13.22	57.2	78.36
CoT	20.28	10.22	45.05	70.06
Structured reasoning	26.0	13.68	54.44	77.0
E5$_{mistral-7b}$				
Baseline	35.77	21.22	68.96	81.02
CoT	35.4	21.11	65.77	79.79
Structured reasoning	39.33	24.72	69.18	81.12

Table 25: Retrieval performance on Housing Statute QA. Recall is computed as retrieval of all gold passages for a given query (lower bound). Since a query has a maximum of 10 potentially relevant gold passages, this metric is computed for $k \geq 10$.

Method	Recall@1	Recall@10	MRR@10	Recall@100	Recall@1000
BM25					
Baseline	13.04	40.44	21.13	67.06	81.69
CoT	22.83	57.44	33.62	79.87	89.77
E5$_{small-v2}$					
Baseline	27.38	64.54	39.12	85.14	94.5
CoT	31.89	68.37	43.65	86.96	94.24
E5$_{base-v2}$					
Baseline	29.11	66.66	40.73	86.85	95.37
CoT	31.4	68.25	43.04	87.46	95.22
E5$_{large-v2}$					
Baseline	30.45	68.68	42.59	88.62	95.97
CoT	32.73	70.45	44.8	88.59	95.71
E5$_{mistral-7b}$					
Baseline	3.36	11.15	5.49	22.33	38.21
CoT	28.13	58.14	37.57	76.36	87.8

Table 26: Retrieval performance on Natural Questions.

Method	Recall@2	Recall@10	MRR@10	Recall@100	Recall@1000
BM25					
Baseline	13.1	32.72	29.26	56.89	74.73
CoT	14.41	37.03	32.72	58.69	74.41
$E5_{small-v2}$					
Baseline	19.27	45.0	40.52	66.47	81.62
CoT	21.36	46.97	42.41	65.29	79.26
$E5_{base-v2}$					
Baseline	23.28	49.55	45.36	70.68	86.01
CoT	21.35	46.75	42.3	65.24	79.7
$E5_{large-v2}$					
Baseline	27.1	56.18	51.73	75.75	89.01
CoT	23.98	50.18	45.57	68.32	81.0
$E5_{mistral-7b}$					
Baseline	5.55	17.25	14.68	35.58	57.47
CoT	17.57	39.73	35.72	58.42	74.58

Table 27: Retrieval performance on HotpotQA.

Method	Recall@1	Recall@10	MRR@10	Recall@100	Recall@1000
BM25					
Baseline	0.0	38.11	10.92	71.6	92.33
CoT	0.08	40.77	11.56	75.27	94.76
$E5_{small-v2}$					
Baseline	0.0	27.15	8.53	59.94	87.72
CoT	1.8	30.13	10.0	63.93	89.36
$E5_{base-v2}$					
Baseline	0.0	28.09	8.55	59.62	88.58
CoT	3.05	32.47	11.53	65.96	91.47
$E5_{large-v2}$					
Baseline	0.0	32.71	9.97	63.15	88.89
CoT	2.27	33.8	11.42	68.62	92.1
$E5_{mistral-7b}$					
Baseline	0.0	52.11	16.42	83.02	96.56
CoT	0.86	38.97	12.46	78.09	95.46

Table 28: Retrieval performance on COLIEE.

Method	Recall@1	Recall@10	MRR@10	Recall@100	Recall@1000
BM25					
Baseline	0.11	11.75	3.43	27.57	48.26
CoT	0.63	8.52	2.78	22.2	40.06
$E5_{small-v2}$					
Baseline	1.16	5.02	2.28	10.94	21.89
CoT	0.84	3.65	1.64	8.31	17.33
$E5_{base-v2}$					
Baseline	1.33	5.3	2.4	11.29	22.9
CoT	0.84	3.89	1.68	8.59	18.41
$E5_{large-v2}$					
Baseline	1.3	6.8	2.9	14.42	25.68
CoT	0.84	5.4	2.1	12.35	23.82
$E5_{mistral-7b}$					
Baseline	1.37	8.49	3.35	19.22	32.3
CoT	0.81	4.91	1.93	12.63	23.33

Table 29: Retrieval performance on CLERC.

Retrieval Method	Accuracy (Top 1)	Accuracy (Top 10)
No passage	37.84	-
BM25	39.10	41.14
BM25 + reasoning	42.82	44.42
$E5_{small-v2}$	39.24	41.06
$E5_{small-v2}$ + reasoning	41.69	44.72
$E5_{base-v2}$	39.67	41.99
$E5_{base-v2}$ + reasoning	42.46	44.98
$E5_{large-v2}$	40.63	41.83
$E5_{large-v2}$ + reasoning	42.09	44.75
$E5_{mistral-7b}$	41.33	43.52
$E5_{mistral-7b}$ + reasoning	42.46	44.39
Reasoning rollout as pseudo-passage	42.69	
Gold passage	57.38	-

Table 30: Downstream task performance on Bar Exam QA on Llama-3-8B-Instruct. We perform coarse retrieval of the top k passages per query using the retrieval method in the first column of the table. + reasoning indicates the structured reasoning query expansion method. Then, we rerank the top k retrieved passages using the predicted answer confidence from the Llama-3-8B-Instruct model. Accuracy (Top k) is calculated using the passage that gives the maximum confidence answer prediction from the top k retrieved passages.

Retrieval Method	Accuracy
No passage	52.22
Gold passage	68.80

Table 31: Downstream task performance on Bar Exam QA on Llama-3-70B-Instruct. We evaluate for no passage and gold passage to show that the gold passage quality is high, but the Llama-3-8B-Instruct struggles to apply them to the question.

Retrieval Method	Accuracy
No passage	49.73
Reasoning rollout as pseudo-passage	47.94
Retrieved passage	50.23
Gold passage	62.76

Table 32: Downstream task performance on Bar Exam QA on GPT-4o-mini (gpt-4o-mini-2024-07-18). We evaluate for no passage, retrieved passage (E5$_{\text{mistral-7b}}$ + structured reasoning), the generative reasoning rollout from the structured reasoning query expansion method as a pseudo-passage, and the gold passage.

Retrieval Method	Accuracy (Top 1)	Accuracy (Top 10)
No passage	58.81	-
BM25	68.22	66.43
BM25 + reasoning	67.43	66.18
E5$_{\text{small-v2}}$	67.59	65.83
E5$_{\text{small-v2}}$ + reasoning	67.30	65.65
E5$_{\text{base-v2}}$	68.10	66.14
E5$_{\text{base-v2}}$ + reasoning	68.19	65.72
E5$_{\text{large-v2}}$	68.43	66.14
E5$_{\text{large-v2}}$ + reasoning	67.95	65.72
E5$_{\text{mistral-7b}}$	70.26	66.81
E5$_{\text{mistral-7b}}$ + reasoning	69.03	66.58
Reasining rollout as pseudo-passage	70.23	
Gold passage	75.27	-

Table 33: Downstream task performance on Housing Statute QA on Llama-3-8B-Instruct. We perform coarse retrieval of the top k passages per query using the retrieval method in the first column of the table. + reasoning indicates the structured reasoning query expansion method. Then, we rerank the top k retrieved passages using the predicted answer confidence from the Llama-3-8B-Instruct model. Accuracy (Top k) is calculated using the passage that gives the maximum confidence answer prediction from the top k retrieved passages.

Retrieval Method	Accuracy
No passage	48.18
Reasoning rollout as pseudo-passage	68.51
Retrieved passage	71.71
Gold passage	77.98

Table 34: Downstream task performance on Housing Statute QA on GPT-4o-mini (gpt-4o-mini-2024-07-18). We evaluate for no passage, retrieved passage (E5$_{\text{mistral-7b}}$ + structured reasoning), the generative reasoning rollout from the structured reasoning query expansion method as a pseudo-passage, and the gold passage.