AQuAECHR: Attributed Question Answering for European Court of Human Rights

Korbinian Q. Weidinger[*1], **Santosh T.Y.S.S**[*1], **Oana Ichim**[2], **Matthias Grabmair**[1]
[1]School of Computation, Information, and Technology
Technical University of Munich, Germany
[2]Graduate Institute of International and Development Studies, Geneva, Switzerland
{santoshtyss, korbi.weidinger2}@gmail.com, matthias.grabmair@tum.de
oana.ichim@graduateinstitute.ch

Abstract

LLMs have become prevalent tools for information seeking across various fields, including law. However, their generated responses often suffer from hallucinations, hindering their widespread adoption in high stakes domains such as law, which can potentially mislead experts and propagate societal harms. To enhance trustworthiness in these systems, one promising approach is to attribute the answer to an actual source, thereby improving the factuality and verifiability of the response. In pursuit of advancing attributed legal question answering, we introduce AQuAECHR, a benchmark comprising information-seeking questions from ECHR jurisprudence along with attributions to relevant judgments. We present strategies to automatically curate this dataset from ECHR case law guides and utilize an LLM-based filtering pipeline to improve dataset quality, as validated by legal experts. Additionally, we assess several LLMs, including those trained on legal corpora, on this dataset to underscore significant challenges with the current models and strategies dealing with attributed QA, both quantitatively and qualitatively.

1 Introduction

Recent strides in LLMs have prompted legal ecosystem to reexamine their practices, leading to a large number of legal technology startups and law firms adopting LLM-based tools for a variety of tasks, such as sifting through legal case briefs to fast track legal research, formulating litigation strategies, analysing and drafting contracts (Dahl et al., 2024). Despite demonstrating proficiency in a range of law-related tasks such as bar and law school exams (Katz et al., 2024; Martínez, 2024), statutory reasoning and interpretation (Blair-Stanek et al., 2023; Engel and Mcadams, 2024), Issue-Rule-Application-Conclusion framework based legal reasoning (Kang et al., 2023; Guha et al., 2024), lack of trust remains a primary deterrent for adopting such tools due to their propensity to produce hallucinations, generating text that is inconsistent with current case law and doctrines.

In high-stakes domain like law, strict adherence to the source text is of paramount importance and any deviation from them with imprecise interpretations can result in erroneous legal advice or decisions, with potentially harmful consequences. A recent incident underscores this risk when a lawyer relied on ChatGPT for legal research, submitting court filings containing bogus quotes from bogus judicial decisions [*]. To mitigate such negative impacts, the development of models capable of integrating genuine citations to supporting evidence is vital as this enables users to verify the factuality of model output thereby reinforcing the transparency and reliability of LLMs (Bohnet et al., 2022; Kamalloo et al., 2023). To investigate the ability of current LLMs to generate responses along with citations/attributions for legal queries, a high-quality dataset is imperative.

In this work, we curate AQuAECHR, a benchmark of information-seeking questions along with attributions, focusing on the jurisprudence of European Court of Human Rights (ECHR) which adjudicates complaints by individuals against states about alleged violations of their rights as enshrined in the European Convention of Human Rights. Our dataset is constructed from caselaw guides[*], which discusses various legal concepts involved under each convention article by providing references to the paragraphs of the ECHR judgements. We use these discussions along with citations to judgements as responses and use LLMs to automatically construct questions via different strategies, drawing

[*]These authors contributed equally to this work

[*]https://www.nytimes.com/2023/05/27/nyregion/avianca-airline-lawsuit-chatgpt.html

[*]https://ks.echr.coe.int/web/echr-ks/all-case-law-guides

Findings of the Association for Computational Linguistics: ACL 2025, pages 1418–1447
July 27 - August 1, 2025

on the recent works of using LLMs for instruction dataset generation (Köksal et al.; Ushio et al., 2023; Li et al., 2023b). We validate the effectiveness of the curated question-answer dataset with an ECHR expert across different dimensions such as fluency and comprehensiveness. We found that expert-defined legal reasoning paths help LLMs generate better QA pairs, with answers being highly comprehensive and concise through semantic search and sentence-level answer extraction strategies, respectively. Further, we devise an automated low-quality data filtering strategy by prompting LLMs with obtained question-response pairs, based on using LLMs as evaluators (Liu et al., 2023b; Fu et al., 2023; Wang et al., 2023a; Zheng et al., 2024). This filtering step further helps to improve the quality of our curated dataset.

We assess different state-of-the-art LLMs with different attributed question answering strategies such as retrieve-then-generate (vanila, Llatrieval), post-hoc attribution (vanila, RARR) on our curated AQuAECHR dataset. We assess the generated responses and corresponding attributions across answer fluency, answer correctness, citation faithfulness and citation quality. Our findings reveal (i) base models generate decent responses but they do not improve attribution/citation quality when combined with posthoc retrieval method. This could not be mitigated even with posthoc editing mechanisms to improve faithfulness such as RARR. (ii) retrieve-then-generate paradigm with its inductive bias remain faithful to retrieved documents, reflected in its high citation faithfulness scores. Quality of evidence can be moderately improved using LLMs to provide feedback to the retriever in case of some models. (iii) Across models, SaulLM, legally pre-trained and instruction trained observes to be overfitting to legal data, generating good legal terminology on surface level; but eventually loses instruction following capabilities of its base Mistral version, due to domain-specific training. We hope that our benchmark, AQuAECHR, can benefit the legal NLP community in building improved methods for long-form legal question answering and developing automated metrics to assess them.

2 Related Work

Attribution Datasets Early works (Bohnet et al., 2022) utilized short answer datasets such as Natural Questions (Kwiatkowski et al., 2019) to measure attribution evaluation. Subsequent work (Gao et al., 2023) shifted to use long form QA datasets such as ASQA (Stelmakh et al., 2022), QAMPARI (Amouyal et al., 2022) and ELI5 (Fan et al., 2019). These datasets do not have golden citation annotation and hence other works tried to collect references by simulating web environments to collect web behaviour of annotators (Nakano et al., 2021; Qin et al., 2023). Kamalloo et al. (2023) created HAGRID dataset using the question and attribution passages from existing IR dataset, MIRACL (Zhang et al., 2023) and used LLMs to generate golden answers. While above works use unstructured documents like web pages as attribution sources, there have been efforts to use structured knowledge bases as attribtuon sources (Li et al., 2023b; Hu et al., 2024). Rather than complete abstractive answer which makes evaluation challenging, Schuster et al. (2023) has proposed semi-extractive QA task which generates an answer by interleaving verbatim extracted spans from evidence, providing faithful attributions. While most of the works deal with general domain wikipedia, in this work, we explicitly focus on constructing an attributed QA dataset for legal domain, specifically ECHR, with golden citations and responses and use LLMs to craft questions utilizing them. Closest to our work is EXPERTQA (Malaviya et al., 2023), a benchmark of information-seeking questions curated by experts from 32 fields including law. We emphasize the law is jurisdiction specific and not universal, unlike most other fields such as medicine and most of the questions in this dataset are based on the US jurisdiction.

Attribution Generation Different paradigms of systems have been explored to generate text along with attributions. Direct attribution generation involves prompting models to directly generate attributions based on parametric knowledge (Weller et al., 2023; Sun et al., 2022) and they are found to hallucinate references (Zuccon et al., 2023). Retrieve-then-generate method initially retrieve evidence relevant for a query and generate response based on the retrieved evidence (Lewis et al., 2020; Izacard et al., 2022; Jiang et al., 2023b; Borgeaud et al., 2022; Li et al., 2023a). Some systems in this paradigm are trained using feedback from human interactions (Nakano et al., 2021; Glaese et al., 2022; Menick et al., 2022). Other line of work explore post-hoc attribution which initially generate response and then retrieve evidence as attributions (Gao et al., 2022; Huo et al., 2023; He et al., 2022). Some of these systems use these evidence to edit the response to make it grounded to the evidence.

Attribution Evaluation Rashkin et al. (2023) defines that a statement is attributable to source (AIS) if it can be entailed from given source by a generic hearer. Prior works extensively relied on human evaluation to evaluate attributions (Liu et al., 2023a; Bohnet et al., 2022; Nakano et al., 2021; Muller et al., 2023). Subsequent works (Gao et al., 2022; Bohnet et al., 2022) devised automatic AIS metric by tackling it as NLI task between generated response and the attributed source using TRUE model (Honovich et al., 2022). Yue et al. (2023); Li et al. (2024) explored automatic evaluation of attribution by prompting LLMs and fine-tuning smaller LMs.

Other works related to legal question answering and various tasks constructed on ECHR jurisprudence can be found in App. A.

3 AQuAECHR Task & Dataset

We characterize the task of attributable legal question answering as follows: Given a legal query q and a corpus of text passages D obtained from the ECHR judgements, the system is required to generate an output text t that answers the question. The output text consists of a list of n sentences $\{t_1, \ldots, t_n\}$, and each statement t_i can be followed by an in-text citation to a list of passages $C_i = \{c_{i,1}, c_{i,2}, \ldots\}$, where $c_{i,j} \in D$. It should be noted that certain citation styles, such as "According to $[c_{ij}]$, …" are not explicitly covered by this formulation. However, these sentences can be rewritten to adhere to the specified format.

3.1 Data Curation

Obtaining QA Pairs: We use the case-law guides published by the ECHR, accessible on their knowledge-sharing platform maintained by the court's registry, to curate our attributable QA dataset. These guides are designed to inform legal practitioners about fundamental judgments and decisions delivered by the court and they analyze case-law development for each article of the European Convention on Human Rights and its additional Protocols (e.g., Article 4 - Prohibition of slavery and forced labor) as well as transversal themes (e.g., Data Protection, Rights of LGBTI persons), resulting in 28 article-related and 8 theme-related guides. They provide in-depth discussions of each legal concept by offering pinpointed paragraph-level references to the judgments and decisions of the ECHR. These judgments and decisions serve not only to resolve the specific cases brought before the court but also to elucidate, safeguard, and develop the rules instituted by the convention, thereby contributing to their observance by the states and aligning with the doctrine of the convention as a living instrument. Fig. 1 in App. illustrates a snapshot of a case-law guide discussion organized in paragraphs, with citations pointed to paragraph-level references to ECHR judgments.

We regard these paragraphs embedded with attributions to judgments as potential answers and we try to augment them with the corresponding questions to obtain attributed QA pairs. While gold questions could be obtained via legal expert annotations, this effort would be costly and time-consuming. Instead, we use an existing off-the-shelf LLM to generate the corresponding questions using their generative capabilities.

Our QA pair creation strategy consists of three stages. Initially, we instruct `gpt-3.5-turbo-16k-0301` to generate questions that can be answered using the provided multiple paragraphs, selected from the case-law guide sequentially. We collaborate with an ECHR legal expert to formulate the prompts, outlining a predefined reasoning path that involves identifying the legality of domestic measures and restrictions, distinguishing between legal doctrines, and contextually applying those doctrines. The detailed prompts can be found in Appendix B.1. The model is explicitly guided to avoid generating questions that mention specific cases, instead emphasizing the connection of fact patterns to specific legal doctrines, creating a general information-seeking scenario.

Then we use the generated question to perform a semantic similarity search using `text-embedding-3-small`[*] embeddings across all the paragraphs in the respective case-law guide. This helps identify paragraphs in the case-law guide that deal with the same information sought in the query, ensuring the answer is comprehensive. Then we collect the top-k passages from the semantic similarity search along with the original seed passages from the first stage and instruct GPT-3.5 to extract the sentences from these paragraphs that are related to the generated question. While the semantic search step ensures the answer is comprehensive by adding necessary information to the

[*] https://platform.openai.com/docs/guides/embeddings/embedding-models

query across the case-law guide, the sentence-level extraction stage makes it concise by removing information irrelevant to the query.

Passages Corpus Collection: We acquire ECHR judgements collection as an HTML data dump from HUDOC[*], the publicly available database of the ECHR, along with their associated metadata. We retain only the English documents based on their metadata (Document Type: 'HEJUD'). We parsed the judgement into paragraphs using various hand-crafted heuristics to overcome inconsistent HTML structure, the presence of sub-paragraph numbers within each paragraph and the occurrence of spurious paragraph numbers resulting from verbatim text copied from other documents to cross-reference those paragraphs.

3.2 Quality Assessment by Expert

To evaluate the quality of the created QA pairs using our automated strategy, we ask an ECHR legal expert to score each QA pair on a scale of 1-5 for each of the following five dimensions: (i) Question Fluency: whether the generated question is fluent and grammatically sound. (ii) Answer Fluency: whether the selected answer is fluent and coherent. (iii) Answer Comprehensiveness: whether the answer covers all aspects relevant to the question without missing pertinent information. (iv) Answer Conciseness: whether the answer includes the necessary information for the question without redundant content. (v) Question Utility: whether the question is relevant to real-life information-seeking encounters in the expert's professional life.

We compare our strategy to the (i) using single paragraph from case-law guide as potential answer and generating question (ii) using multiple paragraphs as potential answer (iii) using multiple paragraphs with sentence-level answer extraction from multiple paragraphs (iva) using our three step strategy with generating question from multiple paragraphs, searching for additional relevant information and sentence-level answer extraction from all the paragraphs. All these four approaches are prompted to follow chain-of-thought (Wei et al., 2022) where LLM can devise its own reasoning path, contrary to our strategy of pre-defined legal expert curated reasoning path. We create two legal expert defined reasoning paths, but follow rest of the pipeline similar to (iva) and use them as two variants, (ivb, ivc), to generate the question. All the detailed prompts can be found in App. B.1. We thus sample 50 QA pairs from each of the approach, resulting in a total of 300 pairs and are provided to the expert for quality assessment without disclosing the prompting strategy they are obtained from.

We report the averaged scores in Table 1 and the distributions of scores in Appendix B.2. We observe that questions remain fluent across all strategies, highlighting GPT-3.5's robust generative capabilities. However, the single paragraph approach (i) as a potential answer suffers from low comprehensiveness, often missing information relevant to the query. Including multiple paragraphs (ii) improves comprehensiveness but reduces conciseness due to the inclusion of irrelevant information. This issue is partially mitigated by the sentence-level extraction method (iii), which improves conciseness but still lacks comprehensiveness because the relevant information is not always present in a specific location in sequence in the case-law guide. Incorporating a search process to include relevant information across the guide (iva-c) improves comprehensiveness. Using legal specific pre-defined reasoning paths (ivb, ivc) further enhances the extraction of relevant sentences, resulting in improved comprehensiveness and conciseness compared to the LLMs' own reasoning in the chain-of-thought (CoT) approach. Overall, involving a legal expert to craft better prompts helps LLMs generate questions that are perceived to be of higher utility. While extracting individual sentences from paragraphs (iii-ivc) to form the final answer can result in a slight drop in answer fluency and coherence due to breaks in discourse connectivity, this approach is compensated by higher comprehensiveness and conciseness. One way to improve answer fluency could be to use the answer synthesized by the model based on the actual paragraphs or these extractive segments, but we refrain from that as these models are prone to hallucinate and introduce unfaithful information, which would negatively impact the overall QA pair quality.

3.3 Automated Filtering

With recent studies demonstrating that LLMs can act as better reference-free evaluators and obtain better correlation with human quality judgments (Fu et al., 2023; Zheng et al., 2024; Wang et al., 2023a; Liu et al., 2023b), we use G-EVAL (Liu et al., 2023b), a framework of using LLMs with chain-of-thoughts to evaluate the quality of gener-

[*] http://hudoc.echr.coe.int

Strategy	Question Fluency	Answer Fluency	Answer Compr.	Answer Concise.	Question Utility
(i) Single Paragraph	4.92	4.80	3.70	3.92	4.10
(ii) Multiple Paragraphs	4.96	4.80	4.00	3.60	4.70
(iii) Multiple Para. with Sentence level	4.94	4.62	3.82	4.46	4.92
(iva) Sentence level with search - CoT	4.92	4.56	4.26	4.50	4.86
(ivb) Sentence level with search -Legal 1	4.88	4.32	4.08	4.70	4.58
(ivc) Sentence level with search - Legal 2	4.84	4.68	4.40	4.66	4.84

Table 1: Expert assessment of obtained QA pairs using various strategies. The higher the score, the better it is.

	F1	TP	FP	TN	FN
Fluency	96.9	281	14	1	4
Compre.	82.6	195	75	23	7
Concise.	77.7	179	42	18	61
Overall	63.38	156	37	81	26

Table 2: Quality assessment using G-EVAL

ated QA pairs and use it as an automated filtering strategy to improve quality of AQuAECHR. We prompt GPT-3.5 to score the given QA pair based on the answer fluency, comprehensiveness and conciseness and label the ones with less than or equal to 3 to be of low quality. Detailed prompt can be found in App. B.3.

We report F1-score and confusion matrix on binarized value of low and high quality on 300 expert annotated QA pairs in Tab. 2. We observe high F1-scores across each of the dimensions. Among them, conciseness appears challenging, given the subjectivity involved to judge, making it harder. Overall, this automated assessment helps us to some extent in identifying and filtering low-quality samples, improving the quality of our dataset.

3.4 AQuAECHR: Data Analysis

We extend the above generation and filtering strategy across all the guides by applying strategy (ivc) to generate QA pairs and check their quality using our filtering strategy. If we attain a low quality pair, we repeat with strategies (ivb) and (iva) until we attain high-quality pairs; otherwise, we discard the pairs. Next, we identify and filter out duplicate questions based on cosine similarity with `text-embedding-3-small` embeddings at a threshold of 0.8. Additionally, we filter out QA pairs for which the citations in the answers do not refer to English documents, as the cited case-law hyperlinks can point to judgments in either English or French (the two official languages of the Court). We exclude pairs where the citations do not include a pinpointed paragraph number but rather cite an entire case. Finally, this results in 1116 QA pairs, comprising an average of 295 tokens and 5 sentences per answer, 7.8 cited paragraphs on average per answer, 49 tokens per question. Distribution of these statistics are provided in App. B.4.

4 Experiments

4.1 Approaches to Attributed QA

We experiment with the following LLM models: `Mistral-7B-instruct`* (Jiang et al., 2023a), `SaulLM-7B-instruct`* (Colombo et al., 2024) (32k context length) which is initialized with Mistral-7B and further trained on an English legal corpus followed by instruction tuning and the recent `Llama-3-8B-Instruct` and `Llama-3-70B-Instruct`* (Touvron et al., 2023) (8k context length). We generate the response directly from these models by prompting them with questions as closed-book models and hence consequently, they do not not cite any evidences.

Retrieve-then-generate - vanilla (Lewis et al., 2020): In this paradigm, we first retrieve k relevant passages based on the input question. Then LLM is prompted with question and these retrieved passages to produce the answer along with inline citations to these passages.

Retrieve-then-generate - LLatrieval (Li et al., 2023a): LLM fail to generate accurate response if the retrieval cannot retrieve the supporting evidence correctly, overshadowing the LLM's remarkable abilities. To address this, LLatrieval utilizes LLM to update the retrieval result until it verifies that the retrieved documents can sufficiently support answering the question. LLM iteratively generate missing-info query and retrieve a new list of documents that may contain missing information.

*https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

*https://huggingface.co/Equall/Saul-7B-Instruct-v1

*https://huggingface.co/meta-llama/Meta-Llama-3-8B(70B)

Post-hoc retrieval - vanilla: In this paradigm, the models are directly used to generate response as closed book models and attribution is carried out after generation. For each statement in the generated response, we retrieve the best matching passage from the corpus and cite it.

Post-hoc retrieval - RARR (Gao et al., 2022): This method uses research-and-revise strategy where it initially retrieves the evidence based on the generated response and then edits the response based on the retrieved evidence to fix unsupported content and make it consistent with evidence.

Retrievers: We use GTR (Ni et al., 2022), embedding based dense retriever model based on T5-XXL (Raffel et al., 2020). Detailed prompts for all these strategies are provided in App. C.1.

4.2 Evaluation Metrics

We compare the generated responses against the reference response using the following four dimensions automatically.

Answer Fluency: We use MAUVE (Pillutla et al., 2021) to assess the fluency of the output. As LLMs are capable of producing fluent text, we employ this as a sanity check, ensuring that these values are sufficiently high, following Gao et al. (2023).

Answer Correctness: We compare the generated answers to the reference answers using both lexical and semantic similarity metrics. For lexical similarity, we use the ROUGE-L F1 score (Lin, 2004), and for semantic similarity, we use BERTScore (Zhang et al., 2019) computed with the `t5-large` model (Raffel et al., 2020). Additionally, we employ an NLI-based metric at the claim level (i.e., sentence level) following Gao et al. (2023), to validate the presence of claims in the reference answer in the generated answer. Specifically, we use TrueTeacher (Gekhman et al., 2023), a fine-tuned T5-11B model (Raffel et al., 2020), to assess whether each claim in the reference answer (hypothesis) is entailed by the entire generated answer (premise) and report the average entailment score across all claims in each reference answer.

Citation Faithfulness (Text-Citation Attribution): We evaluate the alignment by measuring citation recall (Gao et al., 2023), which determines if the claims in the generated answer are supported by the respective cited passages in the response. Similar to Auto-AIS (Bohnet et al., 2022), we use an NLI framework to assess whether each generated statement with at least one citation is supported by the cited paragraphs. Specifically, for each statement with a citation, we treat the concatenation of all cited paragraphs as the premise and the statement itself as the hypothesis to obtain an entailment score, and then report the recall across all such statements with citations in the generated response. We use the TrueTeacher (Gekhman et al., 2023) model for the NLI assessment.

Evidence Similarity (Citation Quality): We measure the exact match of citations in the generated response compared to the reference response, and report F1-score, following Li et al. (2023b). Due to the potential for multiple cases to contain the same information and selective citation bias in these case-law guides (Santosh et al., 2024b,c), we also compute semantic matching using NLI-based TrueTeacher (Gekhman et al., 2023) between the citation paragraphs in the generated and the reference response. If there are multiple cited paragraphs from the same case, we concatenate these paragraphs to treat them as a single text block of citation and assess whether each reference citation block is covered by the generated citations. Specifically, for each reference citation block, we check if there exists a generated citation that is entailed by the reference citation and report the average recall across all reference citations.

4.3 Results

Fluency: We observe that all models achieve good fluency scores. When utilized as closed book models in Posthoc-vanilla and RARR strategies, `SaulLM-7B` consistently achieves higher MAUVE scores compared to other models. This can be attributed to its legal pre-training, which enables `SaulLM-7B` to generate text that closely resembles legalese answers. However, in Retrieve-then-generate paradigms, other models surpass `SaulLM-7B`, showcasing their adaptability stemming from their diverse instruction tuning phase, allowing these models to quickly adjust to the style based on the retrieved context, with the Llama family showing superiority over Mistral.

Closed Book model + Posthoc Retrieval - vanilla: The base models exhibit competency in addressing legal queries based on their parametric knowledge. Notably, SaulLM achieves a higher ROUGE score owing to its lexical overlap, while other models remain competitive in BERT Score and perform better in fine-grained NLI-based claim recall. This indicates that SaulLM's legal pre-training aids in generating legal terms but may not fully help to cover all the essential aspects required for answer-

Strategy	Model	Fluency	Answer Correctness			Citation Faithful	Evidence Similarity		#Citations per query
		MAUVE	R-L	BERT S	NLI	Recall	EM	NLI	
Closedbook + Posthoc retrieval - vanilla	Mistral-7B	71.2	19.2	*55.6*	5.5	*4.4*	0.3	25.4	3.4
	SaulLM-7B	*78.3*	*21.1*	55.0	4.3	4.3	0.3	19.3	2.4
	Llama-3-8B	63.0	18.9	54.7	4.3	4.0	0.3	*32.3*	4.3
	Llama-3-70B	66.5	19.4	55.3	*5.8*	4.1	0.2	24.4	3.5
Closedbook + Posthoc retrieval - RARR	Mistral-7B	71.2	19.2	*55.6*	5.5	21.5	3.0	70.7	18.0
	SaulLM-7B	*78.3*	*21.1*	55.0	4.3	16.6	*3.1*	58.9	12.7
	Llama-3-8B	63.1	18.9	54.7	3.5	21.2	3.0	59.0	11.9
	Llama-3-70B	66.5	19.4	55.3	*5.8*	*25.6*	*3.1*	***66.6***	15.5
Retrieve-then-generate - vanilla	Mistral-7B	89.6	23.7	58.0	***11.5***	54.2	6.0	*46.0*	6.0
	SaulLM-7B	76.7	21.2	55.2	10.0	14.6	3.6	28.5	3.4
	Llama-3-8B	***91.7***	25.9	58.4	9.6	67.3	6.5	43.8	4.8
	Llama-3-70B	88.1	*26.4*	*58.6*	9.4	*68.5*	*7.3*	45.7	5.7
Retrieve-then-generate - LLatreival	Mistral-7B	79.8	24.7	58.4	*11.2*	69.7	5.7	*49.0*	4.7
	SaulLM-7B	73.0	21.0	54.4	7.9	16.9	2.4	27.2	2.6
	Llama-3-8B	86.3	25.9	58.4	9.7	***72.4***	5.6	39.7	3.8
	Llama-3-70B	*88.4*	***26.7***	***58.9***	9.4	70.3	***7.8***	46.7	4.7

Table 3: Comparison of different LLMs with attributed question answering strategies on AQuAECHR. R-L, BERT S and EM indicate ROUGE-L, Bert Score and Exact Match respectively.

ing the question. Additionally, when these base models are combined with a post hoc retrieval step, they struggle to provide adequate citations, resulting in lower EM scores and moderate NLI matching for evidence similarity (citation quality). This highlights the challenging nature of retrieval in the posthoc step, primarily attributed to a lack of context, leading to a lower number of citations retrieved per query. Moreover, this strategy yields lower citation faithfulness scores as the retrieved citations may not completely encapsulate the text due to the diverse aspects present and the retriever's focus on semantic similarity, which may differ from entailment, particularly given the argumentative nature of legal text. Future work should utilize better reranking strategies based on entailment scores to make the text entailed by the cited text.

Retrieve-then-Generate -vanilla: This strategy, with its inherent bias towards leveraging retrieved evidence for generating responses, yields higher citation faithfulness scores. By retrieving documents solely based on the question, the model gains additional context to produce informative answers, thereby resulting in higher answer correctness scores. However, challenges persist in achieving high evidence NLI scores, primarily due to the difficulty in retrieving all relevant documents solely based on the question. Furthermore, due to the constraint of context length, the selected top documents based on semantic similarity may inadvertently become redundant, failing to cover diverse aspects associated with the question. Thus, there is a need for diverse document re-ranking strategies to encompass multiple aspects in the selected documents. Among these models, SaulLM, with its legal pre-training, exhibits signs of overfitting by inadequately utilizing the retrieved documents to generate responses. This is reflected in its lower citation faithfulness score compared to other models and its tendency to incorporate fewer citations in its responses, resulting in lower evidence similarity scores. This limitation stems from its limited instruction tuning phase with legal-specific tasks, leading to a catastrophic forgetting of instruction following capabilities in its base Mistral.

Closed Book model + Posthoc Retrieval - RARR: We find that fluency and answer correctness remain largely unchanged compared to the vanilla posthoc method. This can be attributed to the minimal number of edits performed by the model, with only 727 sentences out of 22,447 being edited in Llama-8B, accounting for 3.23%, the highest among the models. Detailed statistics regarding the number of edited instances across each model are provided in Appendix 6. We primarily attribute this phenomenon to the agreement confirmation bias of these LLM models which rarely outputs disagreement, thus constraining the editing pipeline. The

Table 4: Average expert ranking and ranking based on automated metrics for Mistral-7B across methods. Lower is better.

Strategy	Ans. Correct.		Cit. Faith.		Cit. Quality	
	Expert	Auto	Expert	Auto	Expert	Auto
RAG	2.45	2.10	2.15	1.70	**2.15**	2.67
LLatrieval	**1.80**	**2.00**	**2.10**	**1.43**	**2.15**	2.20
Post-hoc	2.90	2.95	3.30	3.90	3.35	3.15
RARR	2.85	2.95	2.45	2.98	2.35	**1.98**

Table 5: Spearmann rank correlation between rankings from expert and automated metrics.

Metric	**Corr.**	**p-value**
Rouge-L	0.02	0.87
BERTScore	0.33	0.0025
NLI Claim Rec.	0.16	0.14
Cit. Faith. Rec.	0.45	2.4e-3
Evi. EM	0.22	0.047
Evi. NLI sim.	0.52	6.25e-7

positive agreement bias of the model results in a larger number of citations per query, translating to an increase in citation quality score (both EM and NLI-based) and citation faithfulness score compared to the vanilla posthoc approach.

Retrieve-then-Generate -Llatreival: Overall, this strategy performs well compared to others. It mainly demonstrates utility in selecting important documents by leveraging LLM capabilities to generate missing information as query to retriever facilitating to obtain relevant documents iteratively for Mistral-7B and Llama-3-70B, particularly in NLI-based evidence similarity. This effective selection also results in generating text that is well-supported by its citations, mitigating distractions caused by irrelevant documents in the vanilla case. However, we also observe a decrease in performance for SaulLM and Llama-3-8B, mainly attributed to the models' struggle to formulate effective missing query info and identify relevant documents.

4.4 Human Evaluation

We randomly sample 20 questions with four responses from the four strategies for Mistral-7B model, resulting in total of 80 responses. Each of the four responses for a question are ranked by the legal expert on three criterion of answer correctness/closeness with respect to the target answer, generated text being faithful to the citation and quality of evidence retrieved.

We report the averaged ranks across strategies from expert annotations and automated metrics obtained using NLI based metrics in Table 4. Compared to post-hoc strategies, retrieve-then-generate demonstrate superior performance, with LLatrieval leading in citation faithfulness and answer correctness. Post-hoc methods significantly underperform, even with post-editing, such as in RARR, except in citation quality, where they enumerated a greater number of citations (Tab 3). However, experts did not rank RARR as highly, likely due to their emphasis on precision & authoritativeness as a key factor.

We compute the Spearman rank correlation coefficient (ranging from -1 to 1) (Spearman, 1961) over the list of ranked pairs between expert rankings and the metric-based rankings in Table 5. We observe a positive correlation between the automated metric-based rankings and expert rankings. NLI-based recall for citation faithfulness and evidence demonstrates a strong positive correlation with significant p-values, while for answer correctness, BERTScore exhibits a weak positive correlation.

One key reason for weaker correlations is the difference in how legal experts and automated metrics assess importance. For citation assessment, legal experts may evaluate citations based on their authority or contextual significance, while for answer correctness, they may prioritize certain claims over others. In contrast, automated metrics apply equal weight to all components, failing to capture these nuanced distinctions. Future work should develop metrics that address these complexities in long-form legal QA. Detailed qualitative case study with some examples are presented in App. D.

5 Conclusion

We curate AQuaECHR, a benchmark dataset for legal information seeking questions focusing on ECHR jurisprudence, along with attributions to the evidence, from the caselaw guides automatically using LLMs. We leverage LLMs as reference-free evaluators to filter the low quality data. Overall, the utility and quality of dataset has been examined by legal expert and found to be substantial value. Then we assess different LLMs, including legally trained one, using various strategies for attributed QA. Our evaluation has revealed several shortcomings, including the overfitting nature of domain-specific legal training, the challenging task of semantic similarity in retrieving relevant evidence, and difficulties encountered by LLMs in accurately identifying supporting evidence. We hope that our benchmark, AQuAECHR, benefit the legal NLP community in building improved methods for long-form question answering, providing faithful responses alongside verifiable attribution.

Limitations

AQUAECHR covers only questions and responses in English, with attributions to judgements in English language . While the caselaw guides in other languages do exist and judgements of ECHR are available in their official languages of French and English, we leave non-English languages, either high-resource or low-resource, for future work.

Another potential limitation is the focus on ECHR jurisprudence in our dataset. While ECHR jurisprudence has been an active area of research due to publicly available resources, they represent only a subset of legal contexts globally. Future research could construct attributed QA datasets to include a broader range of legal domains and jurisdictions to enhance its generalizability.

Unlike more general domains such as science or medicine, evaluation in legal settings requires annotators with jurisdiction-specific expertise. For AQUAECHR, we relied on a single human expert with deep knowledge of ECHR jurisprudence due to the difficulty, time, and cost of recruiting highly qualified legal annotators. The evaluation process itself was also particularly demanding, requiring careful examination of detailed legal references to assess evidence quality across multiple dimensions. Future work should aim to involve a larger pool of experts and more diverse annotated samples to improve robustness and reliability of evaluations.

Another important aspect not addressed in our current evaluation framework is the cite-worthiness of claims in responses, which determines the necessity of citations for specific claims. This dimension would provide insights into the proportion of claims in responses that lack supporting citations. Future studies should incorporate this aspect to enhance the comprehensiveness of evaluations.

Furthermore, our evaluation metrics heavily rely on NLI models for assessing different dimensions, where each part of the claim must be verifiable with evidence. However, human judgments often involve implicit world knowledge and may not require substantiating evidence for every aspect of a claim. This discrepancy may lead to a narrower range of NLI values. Future research should explore methods to address this limitation and ensure a more holistic evaluation approach.

Legal retrieval differs significantly from general-purpose retrieval, where semantic similarity dominates. In contrast, legal retrieval must consider factors such as precedential value, temporal relevance (Santosh et al., 2024g) and procedural context (Santosh et al., 2024d). Our retrieval pipeline, while optimized for relevance, does not fully incorporate these constraints. Future retrieval methods should integrate these legal-specific factors to enhance retrieval quality and legal applicability.

Ethics Statement

While we anticipate minimal immediate risks or negative societal consequences associated with our AQuAECHR dataset, it's essential to recognize the potential ethical implications of deploying LLMs for legal information seeking. LLMs have been reported to produce biased outputs, including factual inaccuracies and citations that may not faithfully represent the content. Therefore, caution is warranted in the responsible deployment of such models in real-world applications.

Moreover, LLMs have been shown to reflect and perpetuate biases present in their training data, including societal biases related to race, gender, and socioeconomic status, which may originate from historical legal cases. Addressing these biases is crucial to ensure fair and equitable outcomes in legal scenarios.

Further, the use of LLMs in legal practice raises broader questions about the role of legal professionals and the potential impact of automation on the profession. It's essential to uphold the integrity and professionalism of the legal profession while leveraging LLMs as tools to augment, rather than replace, human expertise. This includes ongoing evaluation and reflection on how LLMs complement legal expertise, while also considering the ethical implications of their integration.

References

Abdelrahman Abdallah, Bhawna Piryani, and Adam Jatowt. 2023. Exploring the state of the art in legal qa systems. *Journal of Big Data*, 10(1):127.

Wasi Uddin Ahmad, Jianfeng Chi, Yuan Tian, and Kai-Wei Chang. 2020. Policyqa: A reading comprehension dataset for privacy policies. *arXiv preprint arXiv:2010.02557*.

Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preoţiuc-Pietro, and Vasileios Lampos. 2016. Predicting judicial decisions of the european court of human rights: A natural language processing perspective. *PeerJ computer science*, 2:e93.

Samuel Joseph Amouyal, Tomer Wolfson, Ohad Rubin, Ori Yoran, Jonathan Herzig, and Jonathan Berant. 2022. Qampari: An open-domain question answering benchmark for questions with many answers from multiple paragraphs. *arXiv preprint arXiv:2205.12665*.

Ngo Xuan Bach, Tran Ha Ngoc Thien, Tu Minh Phuong, et al. 2017. Question analysis for vietnamese legal question answering. In *2017 9th International Conference on Knowledge and Systems Engineering (KSE)*, pages 154–159. IEEE.

Andrew Blair-Stanek, Nils Holzenberger, and Benjamin Van Durme. 2023. Can gpt-3 perform statutory reasoning? In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, pages 22–31.

Bernd Bohnet, Vinh Q Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Massimiliano Ciaramita, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, et al. 2022. Attributed question answering: Evaluation and modeling for attributed large language models. *arXiv preprint arXiv:2212.08037*.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.

Marius Büttner and Ivan Habernal. 2024. Answering legal questions from laymen in german civil law system. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2015–2027.

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. Neural legal judgment prediction in english. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4317–4323.

Ilias Chalkidis, Manos Fergadiotis, Dimitrios Tsarapatsanis, Nikolaos Aletras, Ion Androutsopoulos, and Prodromos Malakasiotis. 2021. Paragraph-level rationale extraction through regularization: A case study on european court of human rights cases. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 226–241.

Andong Chen, Feng Yao, Xinyan Zhao, Yating Zhang, Changlong Sun, Yun Liu, and Weixing Shen. 2023. Equals: A real-world dataset for legal question answering via reading chinese laws. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, pages 71–80.

Pierre Colombo, Telmo Pessoa Pires, Malik Boudiaf, Dominic Culver, Rui Melo, Caio Corro, Andre FT Martins, Fabrizio Esposito, Vera Lúcia Raposo, Sofia Morgado, et al. 2024. Saullm-7b: A pioneering large language model for law. *arXiv preprint arXiv:2403.03883*.

Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. 2024. Large legal fictions: Profiling legal hallucinations in large language models journal of legal analysis (forthcoming).

Xingyi Duan, Baoxin Wang, Ziyue Wang, Wentao Ma, Yiming Cui, Dayong Wu, Shijin Wang, Ting Liu, Tianxiang Huo, Zhen Hu, et al. 2019. Cjrc: A reliable human-annotated benchmark dataset for chinese judicial reading comprehension. In *Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings 18*, pages 439–451. Springer.

Christoph Engel and Richard H Mcadams. 2024. Asking gpt for the ordinary meaning of statutory terms. *MPI Collective Goods Discussion Paper*, (2024/5).

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. Eli5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567.

Erwin Filtz, María Navas-Loro, Cristiana Santos, Axel Polleres, and Sabrina Kirrane. 2020. Events matter: Extraction of events from court decisions. *Legal Knowledge and Information Systems*, pages 33–42.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Y Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, et al. 2022. Rarr: Researching and revising what language models say, using language models. *arXiv preprint arXiv:2210.08726*.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488.

Zorik Gekhman, Jonathan Herzig, Roee Aharoni, Chen Elkind, and Idan Szpektor. 2023. Trueteacher: Learning factual consistency evaluation with large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, et al. 2022. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*.

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.

Ivan Habernal, Daniel Faber, Nicola Recchia, Sebastian Bretthauer, Iryna Gurevych, Indra Spiecker genannt Döhmann, and Christoph Burchard. 2023. Mining legal arguments in court decisions. *Artificial Intelligence and Law*, pages 1–38.

Hangfeng He, Hongming Zhang, and Dan Roth. 2022. Rethinking with retrieval: Faithful large language model inference. *arXiv preprint arXiv:2301.00303*.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021. Cuad: An expert-annotated nlp dataset for legal contract review. *arXiv preprint arXiv:2103.06268*.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. True: Re-evaluating factual consistency evaluation. *arXiv preprint arXiv:2204.04991*.

Christoph Hoppe, David Pelkmann, Nico Migenda, Daniel Hötte, and Wolfram Schenck. 2021. Towards intelligent legal advisors for document retrieval and question-answering in german legal documents. In *2021 IEEE Fourth International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, pages 29–32. IEEE.

Nan Hu, Jiaoyan Chen, Yike Wu, Guilin Qi, Sheng Bi, Tongtong Wu, and Jeff Z Pan. 2024. Benchmarking large language models in complex question answering attribution using knowledge graphs. *arXiv preprint arXiv:2401.14640*.

Weiyi Huang, Jiahao Jiang, Qiang Qu, and Min Yang. 2020. Aila: A question answering system in the legal domain. In *IJCAI*, pages 5258–5260.

Siqing Huo, Negar Arabzadeh, and Charles Clarke. 2023. Retrieving supporting evidence for generative question answering. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, pages 11–20.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Atlas: Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv:2208.03299*.

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. 2023a. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023b. Active retrieval augmented generation. *arXiv preprint arXiv:2305.06983*.

Ehsan Kamalloo, Aref Jafari, Xinyu Zhang, Nandan Thakur, and Jimmy Lin. 2023. Hagrid: A human-llm collaborative dataset for generative information-seeking with attribution. *arXiv preprint arXiv:2307.16883*.

Xiaoxi Kang, Lizhen Qu, Lay-Ki Soon, Adnan Trakic, Terry Yue Zhuo, Patrick Charles Emerton, and Genevieve Grant. 2023. Can chatgpt perform reasoning using the irac method in analyzing legal scenarios like a lawyer? *arXiv preprint arXiv:2310.14880*.

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):20230254.

Phi Manh Kien, Ha-Thanh Nguyen, Ngo Xuan Bach, Vu Tran, Minh Le Nguyen, and Tu Minh Phuong. 2020. Answering legal questions by learning neural attentive text representation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 988–998.

Mi-Young Kim, Randy Goebel, and S Ken. 2015. Coliee-2015: evaluation of legal question answering. In *Ninth International Workshop on Juris-informatics (JURISIN 2015)*.

Abdullatif Köksal, Timo Schick, Anna Korhonen, and Hinrich Schuetze. Longform: Effective instruction tuning with reverse instructions. In *ICLR 2024 Workshop on Navigating and Addressing Data Problems for Foundation Models*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Xiaonan Li, Changtai Zhu, Linyang Li, Zhangyue Yin, Tianxiang Sun, and Xipeng Qiu. 2023a. Llatrieval: Llm-verified retrieval for verifiable generation. *arXiv preprint arXiv:2311.07838*.

Xinze Li, Yixin Cao, Liangming Pan, Yubo Ma, and Aixin Sun. 2023b. Towards verifiable generation: A benchmark for knowledge-aware language model attribution. *arXiv preprint arXiv:2310.05634*.

Yifei Li, Xiang Yue, Zeyi Liao, and Huan Sun. 2024. Attributionbench: How hard is automatic attribution evaluation? *arXiv preprint arXiv:2402.15089*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Nelson F Liu, Tianyi Zhang, and Percy Liang. 2023a. Evaluating verifiability in generative search engines. *arXiv preprint arXiv:2304.09848*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Antoine Louis and Gerasimos Spanakis. 2021. A statutory article retrieval dataset in french. *arXiv preprint arXiv:2108.11792*.

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*, volume 38, pages 22266–22275.

Chaitanya Malaviya, Subin Lee, Sihao Chen, Elizabeth Sieber, Mark Yatskar, and Dan Roth. 2023. Expertqa: Expert-curated questions and attributed answers. *arXiv preprint arXiv:2309.07852*.

Eric Martínez. 2024. Re-evaluating gpt-4's bar exam performance. *Artificial Intelligence and Law*, pages 1–24.

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, et al. 2022. Teaching language models to support answers with verified quotes. *arXiv preprint arXiv:2203.11147*.

Raquel Mochales and Marie-Francine Moens. 2008. Study on the structure of argumentation in case law. In *Proceedings of the 2008 conference on legal knowledge and information systems*, pages 11–20.

Benjamin Muller, John Wieting, Jonathan H Clark, Tom Kwiatkowski, Sebastian Ruder, Livio Baldini Soares, Roee Aharoni, Jonathan Herzig, and Xinyi Wang. 2023. Evaluating and modeling attribution for cross-lingual question answering. *arXiv preprint arXiv:2305.14332*.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

Marıa Navas-Loro and Vıctor Rodriguez-Doncel. 2022. Whenthefact: Extracting events from european legal decisions. In *Legal Knowledge and Information Systems*, pages 219–224. IOS Press.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernandez Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith Hall, Ming-Wei Chang, et al. 2022. Large dual encoders are generalizable retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9844–9855.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. Mauve: Measuring the gap between neural text and human text using divergence frontiers. *Advances in Neural Information Processing Systems*, 34:4816–4828.

Prakash Poudyal, Teresa Gonçalves, and Paulo Quaresma. 2019. Using clustering techniques to identify arguments in legal documents. *ASAIL@ ICAIL*, 2385.

Prakash Poudyal, Jaromír Šavelka, Aagje Ieven, Marie Francine Moens, Teresa Goncalves, and Paulo Quaresma. 2020. Echr: Legal corpus for argument mining. In *Proceedings of the 7th Workshop on Argument Mining*, pages 67–75.

Yujia Qin, Zihan Cai, Dian Jin, Lan Yan, Shihao Liang, Kunlun Zhu, Yankai Lin, Xu Han, Ning Ding, Huadong Wang, et al. 2023. Webcpm: Interactive web search for chinese long-form question answering. *arXiv preprint arXiv:2305.06849*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2023. Measuring attribution in natural language generation models. *Computational Linguistics*, 49(4):777–840.

Abhilasha Ravichander, Alan W Black, Shomir Wilson, Thomas Norton, and Norman Sadeh. 2019. Question answering for privacy policies: Combining computational and legal perspectives. *arXiv preprint arXiv:1911.00841*.

TYS Santosh, Mahmoud Aly, and Matthias Grabmair. 2024a. Lexabsumm: Aspect-based summarization of legal decisions. *arXiv preprint arXiv:2404.00594*.

TYS Santosh, Mahmoud Aly, Oana Ichim, and Matthias Grabmair. 2025a. Lexgenie: Automated generation of structured reports for european court of human rights case law. *arXiv preprint arXiv:2503.03266*.

TYS Santosh, Rashid Gustav Haddad, and Matthias Grabmair. 2024b. Ecthr-pcr: A dataset for precedent understanding and prior case retrieval in the european court of human rights. *arXiv preprint arXiv:2404.00596*.

TYS Santosh, Elvin Quero Hernandez, and Matthias Grabmair. 2024c. Query-driven relevant paragraph extraction from legal judgments. *arXiv preprint arXiv:2404.00595*.

TYSS Santosh, Kevin D Ashley, Katie Atkinson, and Matthias Grabmair. 2024d. Towards supporting legal argumentation with nlp: Is more data really all you need? In *Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 404–421.

TYSS Santosh, Irtiza Chowdhury, Shanshan Xu, and Matthias Grabmair. 2024e. The craft of selective prediction: Towards reliable case outcome classification–an empirical study on european court of human rights cases. *arXiv preprint arXiv:2409.18645*.

TYSS Santosh, Mohamed Hesham Elganayni, Stanisław Sójka, and Matthias Grabmair. 2024f. Incorporating precedents for legal judgement prediction on european court of human rights cases. *arXiv preprint arXiv:2409.18644*.

Tyss Santosh, Oana Ichim, and Matthias Grabmair. 2023a. Zero-shot transfer of article-aware legal outcome classification for european court of human rights cases. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 593–605.

TYSS Santosh, Isaac Misael OlguÃn Nolasco, and Matthias Grabmair. 2025b. Lecopcr: Legal concept-guided prior case retrieval for european court of human rights cases. *arXiv preprint arXiv:2501.14114*.

Tyss Santosh, Marcel Perez San Blas, Phillip Kemper, and Matthias Grabmair. 2023b. Leveraging task dependency and contrastive learning for case outcome classification on european court of human rights cases. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1103–1111.

TYSS Santosh, Tuan-Quang Vuong, and Matthias Grabmair. 2024g. Chronoslex: Time-aware incremental training for temporal generalization of legal classification tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3022–3039.

Tyss Santosh, Shanshan Xu, Oana Ichim, and Matthias Grabmair. 2022. Deconfounding legal judgment prediction for european court of human rights cases towards better alignment with experts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1120–1138.

Tal Schuster, Adam D Lelkes, Haitian Sun, Jai Gupta, Jonathan Berant, William W Cohen, and Donald Metzler. 2023. Semqa: Semi-extractive multi-source question answering. *arXiv preprint arXiv:2311.04886*.

Francesco Sovrano, Monica Palmirani, Biagio Distefano, Salvatore Sapienza, and Fabio Vitali. 2021. A dataset for evaluating legal question answering on private international law. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, pages 230–234.

Charles Spearman. 1961. The proof and measurement of association between two things.

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. 2022. Asqa: Factoid questions meet long-form answers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8273–8288.

Zhiqing Sun, Xuezhi Wang, Yi Tay, Yiming Yang, and Denny Zhou. 2022. Recitation-augmented language models. *arXiv preprint arXiv:2210.01296*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Priyansh Trivedi. 2018. A question answering system on regulatory documents. In *Legal Knowledge and Information Systems: JURIX 2018: The Thirty-first Annual Conference*, volume 313, page 41. IOS Press.

Asahi Ushio, Fernando Alva-Manchego, and Jose Camacho-Collados. 2023. An empirical comparison of lm-based question and answer generation methods. *arXiv preprint arXiv:2305.17002*.

Josef Valvoda, Ryan Cotterell, and Simone Teufel. 2023. On the role of negative precedent in legal outcome prediction. *Transactions of the Association for Computational Linguistics*, 11:34–48.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023a. Is chatgpt a good nlg evaluator? a preliminary study. *arXiv preprint arXiv:2303.04048*.

Steven H Wang, Antoine Scardigli, Leonard Tang, Wei Chen, Dimitry Levkin, Anya Chen, Spencer Ball, Thomas Woodside, Oliver Zhang, and Dan Hendrycks. 2023b. Maud: An expert-annotated legal nlp dataset for merger agreement understanding. *arXiv preprint arXiv:2301.00876*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Orion Weller, Marc Marone, Nathaniel Weir, Dawn Lawrie, Daniel Khashabi, and Benjamin Van Durme. 2023. " according to..." prompting language models improves quoting from pre-training data. *arXiv preprint arXiv:2305.13252*.

Shanshan Xu, Oana Ichim, Isabella Risini, Barbara Plank, Matthias Grabmair, et al. 2023a. From dissonance to insights: Dissecting disagreements in rationale construction for case outcome classification. *arXiv preprint arXiv:2310.11878*.

Shanshan Xu, Leon Staufer, Oana Ichim, Corina Heri, Matthias Grabmair, et al. 2023b. Vechr: A dataset for explainable and robust classification of vulnerability type in the european court of human rights. *arXiv preprint arXiv:2310.11368*.

Xiang Yue, Boshi Wang, Kai Zhang, Ziru Chen, Yu Su, and Huan Sun. 2023. Automatic evaluation of attribution by large language models. *arXiv preprint arXiv:2305.06311*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2023. Miracl: A multilingual retrieval dataset covering 18 diverse languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

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*, volume 34, pages 9701–9708.

Guido Zuccon, Bevan Koopman, and Razia Shaik. 2023. Chatgpt hallucinates when attributing answers. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, pages 46–51.

A Related Work

Legal Question Answering (LQA) It encompasses several components including identifying the existing law, interpreting legal statues and regulations and applying legal principles, doctrines, and precedents to specific factual situations (Abdallah et al., 2023). Different LQA datasets have been created on various legal sources such as privacy policies (Ahmad et al., 2020; Ravichander et al., 2019), regulatory documents (Trivedi, 2018), Belgian statutes (Louis and Spanakis, 2021; Louis et al., 2024), Vietnamese legal advice (Kien et al., 2020; Bach et al., 2017), Japanese bar exam (Kim et al., 2015), Chinese judicial exams and legal advice (Zhong et al., 2020; Huang et al., 2020; Duan et al., 2019; Chen et al., 2023), contract reviews (Hendrycks et al., 2021; Wang et al., 2023b), German legal advice (Hoppe et al., 2021; Büttner and Habernal, 2024), international private law (Sovrano et al., 2021). To ensure reliability of system generated answers in such specialized domains, it necessitates supporting evidence to gain the trust of users, which motivates us to curate AQuAECHR to assess LQA along with attribution, different from prior works which focus on answer correctness.

Tasks related to ECHR Jurisdiction Previous works involving ECtHR corpus has dealt with judgement prediction (Aletras et al., 2016; Chalkidis et al., 2019, 2021; Santosh et al., 2022, 2023a,b; Valvoda et al., 2023; Xu et al., 2023a; Santosh et al., 2024e,f), argument mining (Mochales and Moens, 2008; Habernal et al., 2023; Poudyal et al., 2019, 2020), event extraction (Filtz et al., 2020; Navas-Loro and Rodriguez-Doncel, 2022), summarization (Santosh et al., 2024a, 2025a), query based relevant paragraph retrieval (Santosh et al., 2024c), vulnerability identification (Xu et al., 2023b), prior case retrieval (Santosh et al., 2024b, 2025b). In this work, we leverage the case law guides maintained by registry of ECHR to derive attributed question answering dataset.

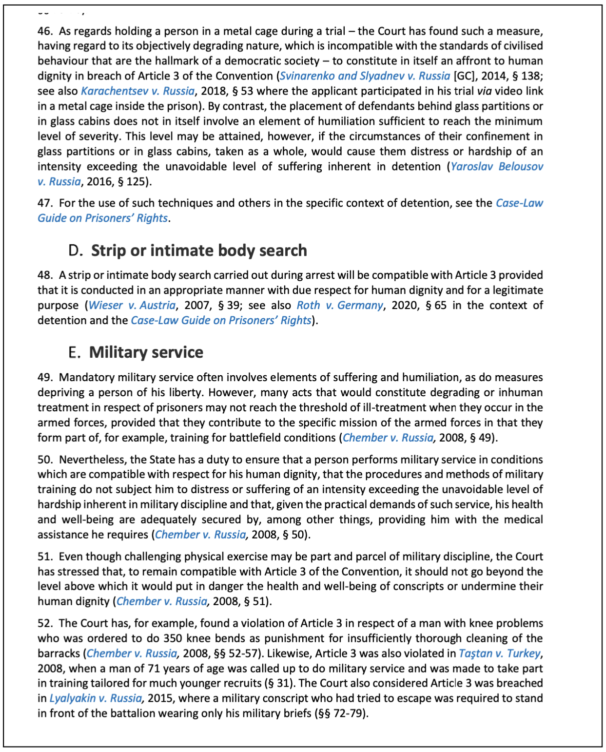
B AQuaECHR dataset

Figure 1 presents a snapshot of a case-law guide, illustrating how discussions of various legal concepts are organized into paragraphs with explicit citations to specific paragraphs in ECHR case-law judgments. This structure forms the foundation for curating our attribution-based QA dataset.

B.1 Prompts for dataset curation

Prompt 2 corresponds to the Single Paragraph strategy. Prompt 3 refers to the Multiple Paragraphs strategy, while Prompt 4 represents the Multiple Paragraphs with Sentence-Level strategy. Prompts 5 and 6 correspond to the Sentence-Level with Search – CoT strategy, detailing the question generation and sentence-level answer extraction steps, respectively. Additionally, Prompts 7 and 8 describe question generation steps performed using legal expert-provided reasoning structures, referred to as

[*] https://ks.echr.coe.int/documents/d/echr-ks/guide_art_3_eng



46. As regards holding a person in a metal cage during a trial – the Court has found such a measure, having regard to its objectively degrading nature, which is incompatible with the standards of civilised behaviour that are the hallmark of a democratic society – to constitute in itself an affront to human dignity in breach of Article 3 of the Convention (*Svinarenko and Slyadnev v. Russia* [GC], 2014, § 138; see also *Karachentsev v. Russia*, 2018, § 53 where the applicant participated in his trial *via* video link in a metal cage inside the prison). By contrast, the placement of defendants behind glass partitions or in glass cabins does not in itself involve an element of humiliation sufficient to reach the minimum level of severity. This level may be attained, however, if the circumstances of their confinement in glass partitions or in glass cabins, taken as a whole, would cause them distress or hardship of an intensity exceeding the unavoidable level of suffering inherent in detention (*Yaroslav Belousov v. Russia*, 2016, § 125).

47. For the use of such techniques and others in the specific context of detention, see the *Case-Law Guide on Prisoners' Rights*.

D. **Strip or intimate body search**

48. A strip or intimate body search carried out during arrest will be compatible with Article 3 provided that it is conducted in an appropriate manner with due respect for human dignity and for a legitimate purpose (*Wieser v. Austria*, 2007, § 39; see also *Roth v. Germany*, 2020, § 65 in the context of detention and the *Case-Law Guide on Prisoners' Rights*).

E. **Military service**

49. Mandatory military service often involves elements of suffering and humiliation, as do measures depriving a person of his liberty. However, many acts that would constitute degrading or inhuman treatment in respect of prisoners may not reach the threshold of ill-treatment when they occur in the armed forces, provided that they contribute to the specific mission of the armed forces in that they form part of, for example, training for battlefield conditions (*Chember v. Russia*, 2008, § 49).

50. Nevertheless, the State has a duty to ensure that a person performs military service in conditions which are compatible with respect for his human dignity, that the procedures and methods of military training do not subject him to distress or suffering of an intensity exceeding the unavoidable level of hardship inherent in military discipline and that, given the practical demands of such service, his health and well-being are adequately secured by, among other things, providing him with the medical assistance he requires (*Chember v. Russia*, 2008, § 50).

51. Even though challenging physical exercise may be part and parcel of military discipline, the Court has stressed that, to remain compatible with Article 3 of the Convention, it should not go beyond the level above which it would put in danger the health and well-being of conscripts or undermine their human dignity (*Chember v. Russia*, 2008, § 51).

52. The Court has, for example, found a violation of Article 3 in respect of a man with knee problems who was ordered to do 350 knee bends as punishment for insufficiently thorough cleaning of the barracks (*Chember v. Russia*, 2008, §§ 52-57). Likewise, Article 3 was also violated in *Taştan v. Turkey*, 2008, when a man of 71 years of age was called up to do military service and was made to take part in training tailored for much younger recruits (§ 31). The Court also considered Article 3 was breached in *Lyalyakin v. Russia*, 2015, where a military conscript who had tried to escape was required to stand in front of the battalion wearing only his military briefs (§§ 72-79).

Figure 1: Snapshot of a case-law guide titled *Article 3: Prohibition of Torture*[*].

Legal-1 and Legal-2, respectively. Both of these legal-based reasoning strategies use Prompt 9 for answer extraction, incorporating expert-provided reasoning structures. For all these prompts, paragraphs from the case-law guides were included without paragraph numbers and were separated by spaces. In sentence-level strategies, sentences `sentences` were prepended with numbers 1 to n and separated by newline characters.

B.2 Quality Assessment by Expert

Fig. 10 illustrates the distribution of Quality Scores across various dimensions, including question fluency, answer fluency, answer comprehensiveness, answer conciseness, and question utility. These scores were provided by experts for 50 generated QA pairs evaluated using different strategies.

B.3 Prompts for dataset filtering

Prompt 11 corresponds to automated evaluation of QA pairs.

B.4 Dataset Analysis

Fig. 12 shows the distribution of the number of tokens in questions, the number of tokens in answers, the number of sentences in answers, and the number of cited passages in answers.

C Experiments

C.1 Prompts

Prompt 13 refers to closed book strategy. Prompt 14 corresponds to the retrieve-then-generate - vanilla strategy. Prompts 15, 16 and 17 correspond to the retrieve-then-generate - LLatrieval for assessing or scoring the retrieved documents, selecting the documents and query formulation for missing information. Prompts 18, 19, and 20 correspond to post-hoc retrieval - RARR strategy for evidence requirement assessment, evidence scoring and editing based on evidence. Evidence in all the prompts is provided in the format of `[Doc i]: document text`.

Model	%Edits	# Edits/#Sentences
Mistral-7B	0.1	18/17479
Saul-7B	0.5	55/11559
Llama-3-8B	3	727/22447
Llama-3-70B	0.09	22/23880

Table 6: Number of edits performed by different models in RARR framework

D Case Study

We present several examples to the legal expert, along with outputs from various models, to analyze the characteristics of the generated responses.

Example 1: Table 7 provides an example question related to the open-ended topic of data protection, along with the target answer and responses from the Mistral-7B model using base (closed book strategy), as well as retrieve-then-generate strategies such as vanilla and Llatrieval.

The base model's response outlines general principles applicable to disclosing electronic information, elaborating on various circumstances encompassed by these principles. However, it fails to provide a precise answer to the core question regarding the defense's involvement in setting criteria for judicial proceedings. The retrieve-then-generate approach supplements the base answer by detailing specific circumstances related to the defense's role in defining disclosure criteria, such as ensuring individuals' ability to defend themselves is not impeded by disclosure. Nevertheless, it lacks precise information about safeguards against arbitrariness in cases requiring a balance of interests.

Llatrieval provides concise text but does not offer new insights into the defense's actual role,

```
Task:
Define a single question based on the provided paragraph from the ECHR case law guides.
The paragraph must answer the question precisely and completely.
Think about the context of the paragraph and how it can be used to create a challenging legal
question.
Use the language from the paragraph in the question.

Paragraph:
{PARAGRAPH}

Answer Template:
Thoughts: {{thoughts on the paragraph and what question could be asked}}
Question: {{the question that can be answered exactly by the provided paragraph}}
```

Prompt 2: Question Generation using *Single Paragraph* strategy.

```
Task:
Define a single question based on the provided paragraphs from the ECHR case law guides.
The paragraphs must answer the question precisely and completely.
Think about the context of each paragraph and what challenging legal question it can answer.
Use the language from the paragraphs in the question.

Paragraphs:
{PARAGRAPHS[5]}

Answer Template:
Thoughts: {{thoughts on the paragraphs and what question could be asked}}
Question: {{the question that can be answered exactly by the provided paragraphs}}
Paragraphs: {{Array of paragraph numbers needed to answer the question. Example: [1,2,6,7]}}
```

Prompt 3: Question Generation using *Multiple Paragraphs* strategy.

which is central to the question. It also overlooks specifying problematic data aspects for the defense and the issues disclosure may raise, unlike vanilla retrieve-then-generate. Both models rightly emphasize the defense's need to challenge, impacting procedure quality. While Llatrieval tends towards more thorough arguments, vanilla retrieve-then-generate leans towards general principles.

In terms of language use, base models employ less legalese, as evidenced by their use of non-legal language for some legal technical terms such as 'help streamline the disclosure' would actually be referred to 'provides additional guarantees' in legal jargon. In contrast, retrieval approaches incorporate legal jargon and logical steps, reflecting overall reasoning aided by retrieved evidence.

Example 2: Table 8 provides an example question addressing the legal issue of determining the applicable Article to a situation where a foreign prisoner is denied contact with their family, along with generated responses using the retrieve-then-generate strategy from four different LLMs.

An ideal response would specify Article 8 would be the applicable because contact with family while in prison was ruled to be part of a prisoner's private life guaranteed under that Article. The Court analyses each limitation on access to family as an interference and that interference is then considered in light of several considerations that the Court deems relevant for that specific situation, forming the core of the Court's reasoning. There are some common considerations for all Articles of the Convention and there also are some specific considerations for each Article. Specific considerations for Article 8 in dealing with right of prisoners to have contact with family and access to the outside world are: legislation, practical conditions of transportation, geographic location and possibilities for the prisoner's family to visit. In addition, when analyzing restrictions on prisoners, the Court has established that States have an obligation to assist the detainee to keep contact with his family.

Generally, for this type of questions, the greater the number of considerations as to why Article 8 is a candidate, the better the support to justify the application of that Article. Even, whenever lawyers and judges look for reasons justifying the application of an Article to a new situation they validate already existing considerations applicable to a situation in order to check whether these considerations

```
Scenario:
Assume the role of an experienced lawyer specializing in ECHR case law.
Your objective is to develop educational and challenging questions for trainee lawyers.
Each question should be based on the provided paragraphs from the ECHR case law guides.
When formulating a question avoid over-contextualizing (referring to specific countries or asking
about specific cases) and avoid asking questions about broad concepts for which the provided
sentences only give partial answers.

Sentences:
{PARAGRAPH_SENTENCES[5]}

Steps:
1. Come up with a question that can be answered exactly with a subset of the provided sentences from
the case law guide.
2. For each sentence reason about if it is required to answer the question.
3. Formulate an exact answer by combining the provided sentences.

Answer Template:
Question: {{Question that can be answered exactly with a subset of the provided sentences}}
Sentence [1]: {{Reason about why this sentence is (not) required for the answer}}
Sentence [2]: {{Reason about why this sentence is (not) required for the answer}}
... (reason about all sentences 1 to k)
Sentence [k]: {{Reason about why this sentence is (not) required for the answer}}
Chosen Sentences: {{Array of chosen sentences. Example: [1,2,4,5,7]}}
```

Prompt 4: Question Generation using *Multiple Paragraphs with Sentence level* strategy.

```
Case law paragraphs:
{PARAGRAPHS[3]}

Task:
Define a single challenging legal question that can be answered with the given case law paragraphs.
Reuse the language from the case law in the question.

Question:
```

Prompt 5: Question Generation using *Sentence level with search - CoT* strategy.

may or may not apply to their situation.

However, none of the models provided comprehensive details of all involved legal aspects. Some aspects considered by the Court, such as conjugal visits and access to social media, were not mentioned. Mistral provided the best answer with a brief overview of why Article 8 applies to the given situation by going through all the steps of the Court's core reasoning structure (essential aspect hardship, interests, normal hardships and restrictions) and identifying the most relevant circumstances that influence that core reasoning (geographical citation, interests of both prisoner and his family, distance, financial hardship).

On the other hand, Llama-3-8b includes general statements of why Article 8 applies, but only very briefly mentions some special circumstances at the end (like prisoner's young age and financial hardship) before concluding abruptly on the possible application. Llama-3-70B response was overall better than Llama-3-8B even though Llama-3-8B had an additional details such as whether the prisoner has young children and financial hardships. This mainly because of the conclusion it finally provides the Court will likely analyze if the effects on the private life go beyond the normal hardships and restrictions which enter in the very concept of imprisonment, whereas Llama-3-8B concludes that the court would likely conclude that the denial of any contact with the family would be a disproportionate interference, giving thus a straightforward answer. It leaves the impression that this is what the Court would say but in deciding whether the interference is disproportionate the Court first analyzes circumstances which go beyond or below the normal hardships, involving a kind of sequential reasoning steps.

Legally pre-trained SaulLM uses specific legal language but provides the conclusion for granted. For instance, it provides general statements about

```
Case law sentences:
{PARAGRAPH_SENTENCES[n]}

Task: Answer the following question with the provided sentences.
At the end of each sentence in your answer cite the used sentences in square brackets.

Question:
{QUESTION}
Answer:
```

Prompt 6: Answer Extraction step using *Sentence level with search - CoT* strategy.

what a hardship is and does not explain why in the case of denial of access this would always amount to a violation. Legal professionals typically anticipate pros and cons for hypotheses rather than straightforward answers, especially for this kind of questions. SaulLM skipped intermediary steps of the decision-making process. It has to strike a balance between like general statements and principles and the specific circumstances from the question to make it legally informative and to be perceived useful by legal professionals.

Example 3: To highlight the utility of editing phenomenon in RARR, we pick up some examples where the model performed edits. We present the question with the response from base and RARR highlighting the sentences that underwent the edit.

In this context of question provided in Tab 9, Article 12 applies to transgender individuals wishing to marry a person of the opposite sex (i.e. opposite to her or his newly assigned sex), as well as to same-sex couples wishing to marry or are already married. However, only a total ban on the former constitutes a violation of Article 12, and a total ban on the latter is to date Convention compliant. We observe the response from SaulLM that identifies the wrongly presented information and have successfully edited to make it faithful and accurate.

In the second example response from Llama-3-8B (Tab 10), the model performs two edits. The former edit sounds faithful and accurate as the base model mentions that revocation is proportionate to the legitimate aim of maintaining trust and safety, but usually restrictions affecting professional license are not considered in light of that aim but rather in light of the aim of applicant's right and reputation. The latter is redundant and just a stylistic change, highlighting the LLMs incapability to perform entailment, marking it as disagreement requiring edit but even with the edit, it preserves the content correctly.

Tab 11 presents another response from Llama-3-8B. The first two edits presented do not actually add anything new, but undergo a rephrase of the base. While the last edit actually says the contrary to the base, but also contrary to the decision of the Court and the retrieved text highlighting the vulnerabilities of model to introduce unfaithful content in this process. The Court arrived to the conclusion that the restrictions were not proportionate to the aim sought by the restriction and the edit says the contrary, that they were proportionate. Moreover, the base correctly says that measures lacked clear criteria for application and were arbitrary and the edit says that measures were in accordance with the law and had no impact on the right to private life, which stands completely inaccurate.

Takeaways: Overall, the legal expert's analysis reveals significant limitations in the LLMs' understanding of legal language and subtleties of legal judgments. These models struggle to discern norms and the factors that shape them, often failing to adapt arguments based on subtle details. They lack the ability to connect general descriptions of legal Articles with specific circumstances that inform their application, missing the crucial relationship between facts and legal precedent.

Moreover, the models demonstrate errors such as an inability to extract concepts linking precedents, leading to challenges in understanding how different aspects of an Article relate to each other or apply to other Articles. They also struggle to differentiate between various legal components such as situations, standards, criteria, conditions, reasons, and restrictions. Consequently, their responses either lack necessary details about the situation presented in the question or provide overly generalized answers that overlook the specific context, rendering them ineffective for practical legal analysis.

In essence, these limitations highlight the need for further development in LLMs to better grasp the complexities of legal language and reasoning, enabling them to produce more nuanced and con-

```
Your objective is to develop educational and challenging questions for lawyers working with ECHR case
 law and for judges who want to draft judgments based on ECHR case law.
Each question should be based on the provided paragraphs from the ECHR case law guides.
When formulating a question reuse the language from the ECHR case law and match legal doctrines to
specific facts.
Emphasize the patterns that link facts to specific legal doctrines.

Doctrines and facts:
{PARAGRAPHS[3]}

Steps:
1. Identify how the margin of appreciation and positive obligations apply in relation to the State's
discretion
2. Identify the reasons that justify necessity and pressing social needs
3. Identify the reasons that command that rights be effective in their application
4. Identify how reasonable measures apply in relation to the State's discretion and to restrictions
imposed by States or private individuals
5. Identify the reasons set forth by the Court to defer to domestic reasons provided by domestic
authorities
6. Define a question that can be answered exactly by the given legal doctrines and applicable facts
to those doctrines

Answer Template:
Margin of appreciation: {{how do the margin of appreciation and positive obligations apply in
relation to the State's discretion}}
Necessities: {{reasons that justify necessity and pressing social needs}}
Effectivity: {{reasons that command that rights be effective in their application}}
Reasonable Measures: {{how do reasonable measures apply in relation to the State's discretion and to
restrictions imposed by States or private individuals?}}
Domestic Reasons: {{the reasons set forth by the Court to defer to domestic reasons provided by
domestic authorities}}
Question: {{define a single question that can be answered exactly by the given legal doctrines and
applicable facts reusing the language from the ECHR case law}}
```

Prompt 7: Question Generation using *Sentence level with search - Legal 1* strategy.

textually relevant responses in legal applications.

```
Your objective is to develop educational and challenging questions for lawyers working with ECHR case
 law and for judges who want to draft judgments based on ECHR case law.
Each question should be based on the provided paragraphs from the ECHR case law guides.
When formulating a question reuse the language from the ECHR case law and match legal doctrines to
specific facts.
Emphasize the patterns that link facts to specific legal doctrines.

Doctrines and facts:
{PARAGRAPHS[3]}

Steps:
1. Identify what are the criteria under the Convention for applying the rights enshrined therein.
2. Identify the conditions that the Court sets forth with view to analyse the legality of domestic
measures and restrictions
3. Identify the reasons provided by the Court to protect applicants and victims and to differentiate
between them.
4. Identify the reasons set forth by the Court to distinguish between legal doctrines and contextual
application of those doctrines.
5. Assign a set of facts to its corresponding Article and identify a sequence of reasons that justify
 the application of the Article to those facts.
6. Explain why analogies and comparisons between Article-fact pair are pertinent
7. Identify separately reasons that are linked to margin of appreciation of the State from those
linked to the Court's appreciation of facts
8. Define a question that can be answered exactly by the given Article-facts correspondence

Answer Template:
Criteria for rights: {{how does the Court define the criteria for applying rights}}
Legality of domestic measures and restrictions: {{conditions that determine that domestic measures
are compliant with the Convention}}
Protection and differentiation of applicants and victims: {{circumstances and conditions that limit
or allow applicants and victims to present their case}}
Distinction between legal doctrines and contextual application: {{how and why and in what
circumstances legal doctrines apply to specific facts}}
Article-facts correspondence: {{the reasons set forth by the Court to justify the application of the
Article/Articles to those facts}}
Analogies and comparisons between Article-fact pair {{the reasons set forth by the Court to justify
why articles and facts differ from one another}}
margin of appreciation of States and Court's appreciation {{the reasons, circumstances and conditions
 set forth by the Court to explain margin of appreciation of States and its own appreciation}}
Question: {{define a single question that can be answered exactly by the given legal doctrines and
applicable facts and by the Article-fact pair, reusing the language from the ECHR case law and
adjusting the question and the answer depending on how fact-Article correspondence is better
addressed with What? How? Why?}}
```

Prompt 8: Question Generation using *Sentence level with search - Legal 2* strategy.

```
Your objective is to develop educational and challenging question-answer pairs for lawyers working
with ECHR case law and for judges who want to draft judgments based on ECHR case law.

Doctrines and facts:
{PARAGRAPH_SENTENCES[n]}

Task: Answer the following question based on the provided doctrines and facts from the ECHR case law
guides.

Use the provided doctrines and facts to answer the question.
Use citations! At the end of each sentence in your answer add all the numbers of the used facts and
doctrines in square brackets.

Question: {QUESTION}
Answer:
```

Prompt 9: Answer Extraction step using *Sentence level with search - Legal 1/2* strategy.

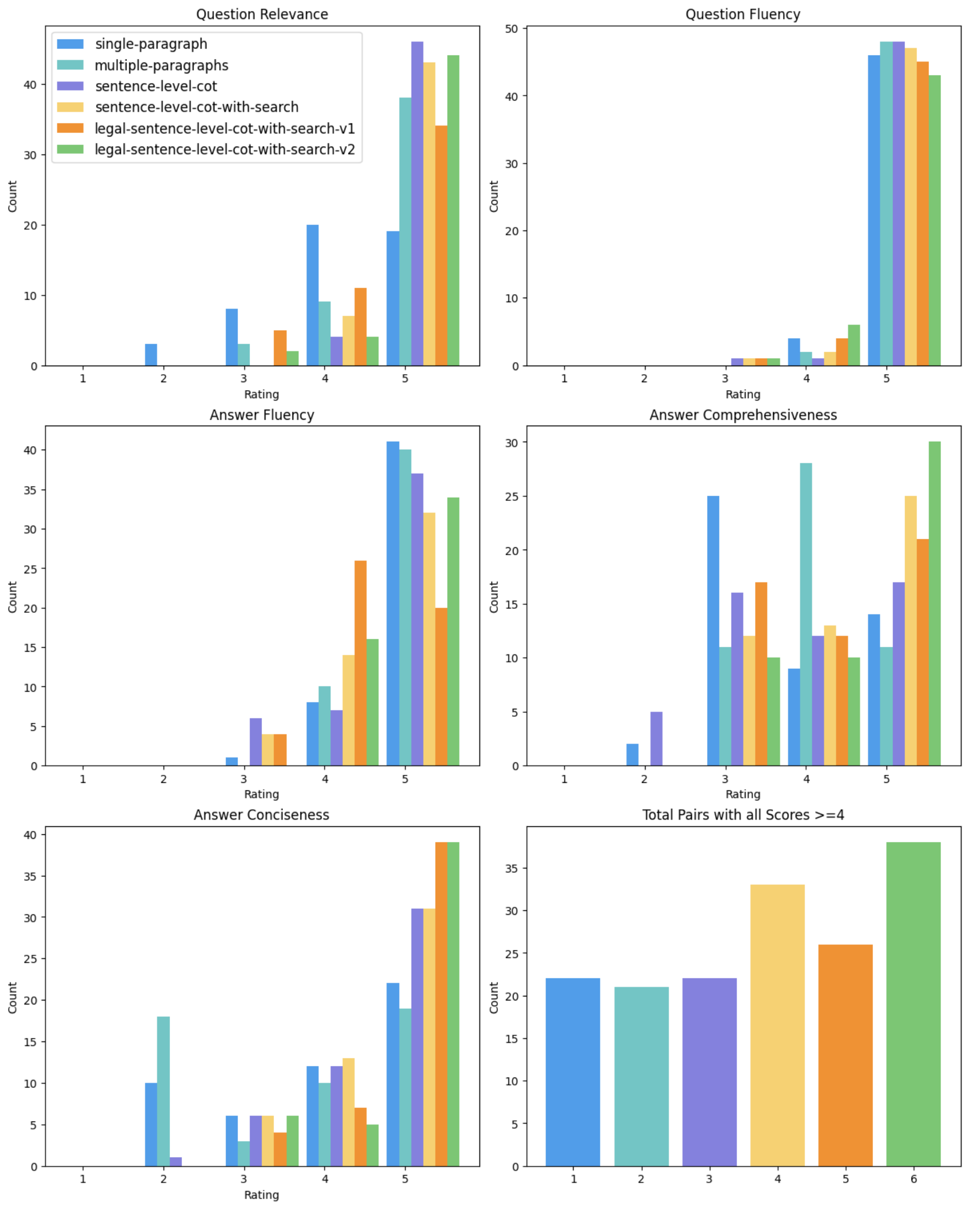


Figure 10: Distribution of Quality Scores provided by experts for 50 generated QA pairs across various strategies.

```
You are a strict legal expert judging ECHR legal question-answer pairs. The answer might be bad, so
be strict!

Question: {QUESTION}
Potential Answer: {ANSWER}

You MUST answer each question in full sentences!

The response MUST follow this template:
Comprehensiveness Analysis: {{Go through the answer and analyze how well it answers the question.
Does it cover all angles of the question? If the question is not a proper question or not a generic
question (mentions a specific case), give a score of 1.}}
Comprehensiveness Score: {{A score from 1 (not comprehensive at all) to 5 (extremely comprehensive)}}
Conciseness: {{Is there any part in the answer irrelevant/unrelated to the question? If so, what is
unneeded?}}
Conciseness Score: {{A score from 1 (not concise at all) to 5 (extremely concise)}}
Answer Fluency: {{Are there any bad sentence transitions in the answer? Are the sentences ordered
correctly? Does the answer start with text clearly continuing the previous text that is not there?}}
Answer Fluency Score: {{A score from 1 (not fluent) to 5 (perfectly fluent)}}
```

Prompt 11: Automated QA quality evaluation used for filtering generated QA pairs.

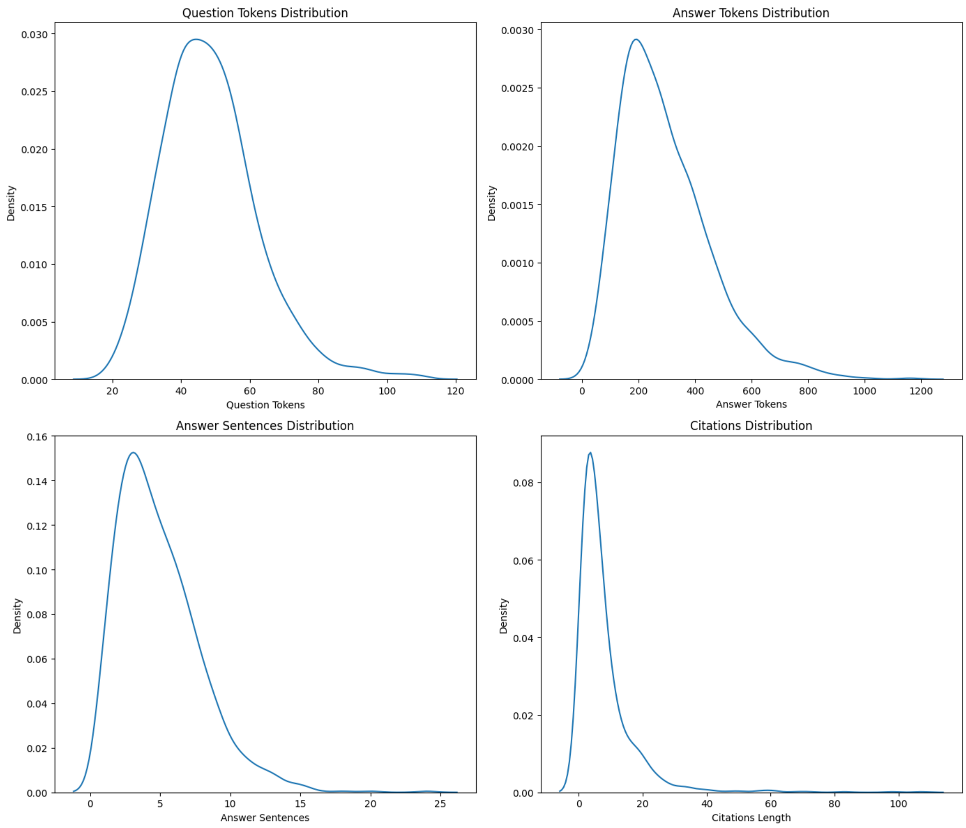


Figure 12: Analysis of AQuAECHR Data: Distributions of tokens in questions and answers, sentences in answers, and cited passages in answers.

```
You are an ECHR legal expert tasked to answer the following question.
Question: {question}
Answer:
```

Prompt 13: Closedbook Strategy

```
You are an ECHR legal expert tasked to answer a question.
The following documents were retrieved and should help you answer the question:
{documents}

Instructions:
Use the retrieved documents to answer the question.
Reuse the language from the documents!
Cite relevant documents at the end of a sentence!
Accepted formats: sentence [citation(s)].
Valid citation formats: [Doc 1] or [Doc 1, Doc 2, Doc 3]
You must follow the [Doc i] format! Do NOT use case names or paragraph numbers to cite documents!
You should NOT provide a list of all used citations at the end of your response!

Question: {question}
Answer:
```

Prompt 14: Retrieve-then-generate - vanilla

```
You are ScoreGPT as introduced below.
You are ScoreGPT, capable of scoring candidate documents based on their level of support for the
corresponding question, with a rating range from 0 to 10.

Input:
- Question: The specific question.
- Candidate Documents: Documents whose combination may maximally support the corresponding question.

Skill:
1. Analyzing the given question(s) and understanding the required information.
2. Searching through documents to score them based on their level of support for the corresponding
question(s),
with a rating range from 0 to 10.

Output:
- A score ranging from 0 to 10, where a higher score indicates greater support of the candidate
documents for the corresponding question, and a lower score indicates lesser support.

Candidate Documents:
{documents}

Question:
{question}

Output Format: (You MUST follow this output format!)
Thoughts: [Your thoughts about how well the candidate documents support the question]
Score: [SCORE]
```

Prompt 15: Retrieve-then-generate - LLatrieval: Document Scoring.

```
You are DocSelectorGPT, capable of selecting a specified number of documents for answering the user's
 specific question.

Input:
- Question: The specific question
- Candidate Documents: Documents contain supporting documents which can support answering the given
questions. Candidate documents will have their own identifiers for FactRetrieverGPT to cite.

Skill:
1. Analyzing the given question and understanding the required information.
2. Searching through candidate documents to select k supporting documents whose combination can
maximally support giving a direct, accurate, clear and engaging answer to the question and make the
answer and is closely related to the core of the question.

Workflow:
1. Read and understand the questions posed by the user.
2. Browse through candidate documents to select k documents whose combination can maximally support
giving a direct, accurate, clear and engaging answer to the question(s) and make the answer and is
closely related to the core of the question.
3. List all selected documents.

Output:
- Selected Documents: The identifiers of selected supporting documents whose combination can
maximally support giving an accurate and engaging answer to the question and make the answer and is
closely related to the core of the question.

Output Example:
Selected Documents: Doc 2, Doc 6, Doc 8 (You MUST follow this format!)

Max number of selectable documents: {k}
- You can only select a maximum of {k} documents!

Candidate Documents:
{documents}

Question:
{question}

Output Format (You MUST follow this output format!)
Thoughts: [Your thoughts about which candidate documents support the question well and why]
Selected Documents: [document identifiers]
```

Prompt 16: Retrieve-then-generate - LLatrieval: Document Selection

```
You are a PassageRetriever,
capable of identifying missing content that answers the given question but does not exist in the
given possible answering
passages and then using your own knowledge to generate correct answering passages using missing
content you identify.

Input:
- Question: The specific question.
- Answering Passages: Possible answering passages.

Output:
- Correct answering passages generated using missing content you identify based on your own knowledge
.

Rules:
1. You have to use your own knowledge to generate correct answering passages using missing content
you identify.
2. Only generate the required correct answering passages. Do not output anything else.
3. Directly use your own knowledge to generate correct answering passages if you think the given
possible answering passages do not answer to the given question.

Workflow:
1. Read and understand the question and possible answering passages.
2. Identify missing content that answers the given question but does not exist in the given possible
answering passages.
3. Directly use your own knowledge to generate correct answering passages if you think the given
possible answering passages do not answer to the given question(s).
4. Use your own knowledge to generate correct answering passages using missing content you identify.

Answering Passages:
{documents}

Question:
{question}

Output Format: (You MUST follow this output format!)
Correct Answering Passages: [correct answering passages]
Missing Passages: [missing passages]
```

Prompt 17: Retrieve-then-generate - LLatrieval: Query formulation for missing information.

```
You will be determining if a sentence of the answer should be supported by a case law citation.

## Context
Question: {question}
Answer: {answer}

## Now we will analyze if the following sentence should have a citation

The sentence for which to decide if it should have a citation:
<sentence-start>{sentence}</sentence-end>

First, analyze the sentence and decide if it should be supported by a case law citation.
Note:
- General knowledge, headers, and other non-sentences do not require citations.
- Legal arguments, facts, examples, ... should be supported by citations.

The format of your response MUST look like this:
Thoughts: [Reason why the sentence <sentence-start>{sentence}</sentence-end> should have a citation]
Should have a supporting citation: [Yes/No]
```

Prompt 18: Post-hoc retrieval - RARR: Checking requirement of Evidence.

```
You will be determining if a piece of evidence agrees with, disagrees with, or is irrelevant to a
sentence for a given question.

## Context
Question: {question}
Answer: {answer}

## Now we will analyze the evidence for the following sentence in the answer

Evidence:
<evidence-start>{evidence}</evidence-end>

The sentence for which to decide if the evidence agrees, disagrees, or is irrelevant:
<sentence-start>{sentence}</sentence-end>

Carefully analyze the evidence and explain in a reasoning step whether it agrees, contradicts, or is
irrelevant to the sentence in <sentence-start>...</sentence-end> tags.
Then, based on your reasoning, provide your final classification, which MUST be one of the following:
- Agrees
- Disagrees
- Irrelevant

The format of your response MUST look like this:
Thoughts: [Reason whether the evidence agrees, disagrees, or is irrelevant to the given sentence]
Final classification: [Your final classification here: Agrees/Disagrees/Irrelevant]
```

Prompt 19: Post-hoc retrieval - RARR: Checking whether Evidence is relevant.

```
You will be editing a sentence based on the disagreement with the evidence.

## Context
Question: {question}
Answer: {answer}

## Now we will edit the following sentence in the answer based on the disagreement with the evidence

Evidence:
<evidence-start>{evidence}</evidence-end>

The sentence in the answer the evidence disagrees with:
<sentence-start>{sentence}</sentence-end>

First, carefully analyze the sentence and identify the part that contains the disagreement with the
evidence.
Then, rewrite the sentence with MINIMAL modification to resolve the disagreement.
We will not accept drastic changes to the sentence!

The format of your response MUST look like this:
Thoughts: [Reason why the sentence in <sentence-start>...</sentence-end> should be edited]
Fix with minimal edit: [The corrected entire sentence with MINIMAL modification to resolve the
disagreement enclosed by <fixed-sentence-start>...<fixed-sentence-end>]
```

Prompt 20: Post-hoc retrieval - RARR: Editing answer based on evidence.

Table 7: Example Mistral Answers

Question: In the context of disclosure of electronic data, what criteria should be considered to determine what might be relevant for disclosure, and why is the defense's involvement in laying down these criteria important?
Target Answer: In the context of disclosure of evidence, complex issues may arise concerning the disclosure of electronic data, which may constitute a certain mass of information in hands of the prosecution. In such a case, an important safeguard in the sifting process is to ensure that the defence is provided with an opportunity to be involved in the laying-down of the criteria for determining what might be relevant for disclosure (Sigurður Einarsson and Others v. Iceland, 2019, § 90; see also Rook v. Germany, §§ 67 and 72).
Base Answer: In the context of disclosure of electronic data before the European Court of Human Rights (ECtHR) or any other court, the relevance of the data is a key consideration. Relevance refers to the data's ability to make a fact more or less probable as to an allegation or issue in the case. When it comes to electronic data, several criteria should be considered to determine its relevance for disclosure: 1. Proximity: Is there a logical connection between the data and the facts or issues at hand? For example, emails or messages exchanged between parties may be highly relevant if they discuss the alleged misconduct or contain important information. 2. Authenticity: Does the data origin from a trusted or reliable source? Has its integrity been maintained during collection and storage? The authenticity of electronic data is crucial for it to be considered admissible and relevant. 3. Admissibility: Does the data possess probative value, or can it be used to prove or disprove a particular fact? For example, text messages, emails, or social media postings can be admissible as evidence provided they meet the authenticity and relevance requirements. 4. Quantity: Is the volume of data manageable and worth the resources required for its processing? The sheer amount of digital data can sometimes be overwhelming, and it should not be disproportionate to the issue at hand or result in an undue burden on the parties involved. 5. Format: Is the format of the data easily accessible and understandable to the court, as well as the parties involved? This criterion is crucial since data can be available in various formats, making it imperative to understand the capabilities and limitations of each format and the means required to access and interpret it. The defense's involvement in laying down these criteria is essential for several reasons: 1. Protection of their rights: The defense has a fundamental right to a fair trial and must be provided with evidence that could reasonably be expected to aid their defense. Therefore, they have a vested interest in ensuring that the appropriate criteria for determining relevance and admissibility of electronic data are met. 2. Efficiency: The defense's input can help streamline the disclosure process by focusing on data that is truly relevant and admissible. This ensures that the trial does not become unnecessarily prolonged, saving resources and time for all parties. 3. Transparency: The defense's input in the disclosure process can promote transparency and ensure that all evidence considered in the case is accessible and understandable to all parties. This is crucial to maintaining the integrity and impartiality of the judicial process.
Retrieve-then-generate - vanilla: In the context of disclosure of electronic data, the criteria to determine what might be relevant for disclosure include evidence directly relevant to the facts of the case and other evidence that might relate to the admissibility, reliability, and completeness of the former (CASE OF YAKUBA v. UKRAINE § 43, CASE OF MATANOVIĆ v. CROATIA § 161). The material the disclosure of which the applicant sought may fall into the latter category (CASE OF YEVDOKIMOV v. UKRAINE § 46). The defense's involvement in laying down these criteria is important as they can ensure that the impact of the relevant material on the safety of the conviction is considered in the light of detailed and informed arguments (CASE OF MATANOVIĆ v. CROATIA § 155). The Court must satisfy itself that the domestic procedure allows for the consideration of the relevance and impact of the undisclosed material in an effective manner. Furthermore, in the context of communications data, it should be noted that these data are often subjected to automated processing and queries, and applying the same safeguards as for content is not practicable due to the large volume and frequent temporal nature of the data (CASE OF BIG BROTHER WATCH AND OTHERS v. THE UNITED KINGDOM § 296). However, the defense's role in laying down these criteria is crucial in ensuring that any limitations on the disclosure of communications data do not compromise the requirements of Article 6. Overall, the defense's involvement in determining what is relevant for disclosure is essential in ensuring the fairness and transparency of criminal proceedings, as they can help ensure that all material evidence is disclosed to the defense, enabling them to effectively challenge the prosecution's case and protect their client's rights (CASE OF YAKUBA v. UKRAINE § 43).
Retrieve-then-generate - LLatrieval: In the context of disclosure of electronic data, the criteria that should be considered to determine what might be relevant for disclosure include the opportunity for the defense to challenge the authenticity of the evidence and to oppose its use (CASE OF PAVLENKO v. RUSSIA § 116, CASE OF ILGAR MAMMADOV v. AZERBAIJAN (No. 2) § 209, CASE OF GÄFGEN v. GERMANY § 95, CASE OF VUKOTA-BOJIC v. SWITZERLAND § 95, CASE OF SVETINA v. SLOVENIA § 44). The quality of the evidence obtained, including the circumstances in which it was obtained, must also be taken into consideration, as these factors may cast doubts on its reliability or accuracy (CASE OF PAVLENKO v. RUSSIA § 116, CASE OF ILGAR MAMMADOV v. AZERBAIJAN (No. 2) § 209, CASE OF GÄFGEN v. GERMANY § 95, CASE OF VUKOTA-BOJIC v. SWITZERLAND § 95, CASE OF SVETINA v. SLOVENIA § 44). The Court has previously held that where the reliability of evidence is in dispute, the existence of fair procedures to examine its admissibility takes on an even greater importance (CASE OF PAVLENKO v. RUSSIA § 116). The defense's involvement in laying down these criteria is important because it ensures that the rights of the defense are respected throughout the proceedings (CASE OF PAVLENKO v. RUSSIA § 116, CASE OF ILGAR MAMMADOV v. AZERBAIJAN (No. 2) § 209, CASE OF GÄFGEN v. GERMANY § 95, CASE OF VUKOTA-BOJIC v. SWITZERLAND § 95, CASE OF SVETINA v. SLOVENIA § 44). The defense should be given the opportunity to challenge the authenticity of the electronic data and to oppose its use in order to maintain a fair trial (CASE OF PAVLENKO v. RUSSIA § 116, CASE OF ILGAR MAMMADOV v. AZERBAIJAN (No. 2) § 209, CASE OF GÄFGEN v. GERMANY § 95, CASE OF VUKOTA-BOJIC v. SWITZERLAND § 95, CASE OF SVETINA v. SLOVENIA § 44). The quality of the electronic data and the circumstances in which it was obtained are also crucial factors that the defense should be allowed to examine and challenge, as they may impact the reliability and accuracy of the evidence (CASE OF PAVLENKO v. RUSSIA § 116, CASE OF ILGAR MAMMADOV v. AZERBAIJAN (No. 2) § 209, CASE OF GÄFGEN v. GERMANY § 95, CASE OF VUKOTA-BOJIC v. SWITZERLAND § 95, CASE OF SVETINA v. SLOVENIA § 44). By involving the defense in the determination of what is relevant for disclosure, the fairness of the proceedings as a whole is upheld.

Table 8: Example Retrieve-then-generate - Vanilla Answers

Question: Which article of the Convention would be applicable to a situation where a foreign prisoner is denied any contact with their family, including through telephone conversations or occasional visits, and what reasons would the Court provide to justify the application of that article to these facts?
Target Answer: As the Convention does not guarantee as such the right to an inter-state transfer or the right of a detainee to be allocated to a particular prison, it is important to ensure that foreign prisoners maintain some contact with their families, at least through telephone conversations or occasional visits (Labaca Larrea and Others v as. France (dec.), 2017, § 44). Moreover, the authorities may be required under Article 8 to make concessions for allowing a prisoner to contact and speak to his or her family members in their own language (Nusret Kaya and Others v. Turkey, 2014, §§ 60-61).
Llama 8b: Based on the facts provided in the paragraphs, the applicable article of the Convention would be Article 8, which guarantees the right to respect for private and family life. This is because the prisoner's denial of any contact with their family, including through telephone conversations or occasional visits, would likely have a significant impact on their private and family life. The Court would likely justify the application of Article 8 to these facts by highlighting the importance of maintaining contact with one's family, particularly in the context of imprisonment. The Court has already established that it is an essential part of a prisoner's right to respect for family life that the authorities enable him or her, or if need be, assist him or her, to maintain contact with his or her close family (CASE OF YUDIN AND OTHERS v. RUSSIA § 3, CASE OF RAZVOZZHAYEV v. RUSSIA AND UKRAINE AND UDALTSOV v. RUSSIA § 270, CASE OF VOYNOV v. RUSSIA § 49, CASE OF POLYAKOVA AND OTHERS v. RUSSIA § 81, CASE OF SAIDOV AND OTHERS v. RUSSIA § 7, CASE OF KHAMASTKHANOVA AND OTHERS v. RUSSIA § 10, CASE OF YEVSYUKOV AND OTHERS v. RUSSIA § 7). The Court would likely emphasize that the denial of any contact with the family would go beyond the "normal" hardships and restrictions inherent in the very concept of imprisonment, particularly in view of the geographical situation of remote penal facilities and the realities of the transport system (CASE OF RAZVOZZHAYEV v. RUSSIA AND UKRAINE AND UDALTSOV v. RUSSIA § 270, CASE OF POLYAKOVA AND OTHERS v. RUSSIA § 81, CASE OF SAIDOV AND OTHERS v. RUSSIA § 7, CASE OF KHAMASTKHANOVA AND OTHERS v. RUSSIA § 10, CASE OF YEVSYUKOV AND OTHERS v. RUSSIA § 7). In addition, the Court would likely consider the specific circumstances of the case, such as the prisoner's young child and the financial hardship of the relatives, to justify the application of Article 8 (CASE OF DANILEVICH v. RUSSIA § 9). The Court would likely conclude that the denial of any contact with the family would be a disproportionate interference with the prisoner's right to respect for private and family life, and would therefore be in violation of Article 8 of the Convention.
Llama 70b: The applicable article of the Convention would be Article 8, which guarantees the right to respect for private and family life (CASE OF YUDIN AND OTHERS v. RUSSIA § 3). The Court would justify the application of Article 8 to these facts by stating that it is an essential part of a prisoner's right to respect for family life that the authorities enable him or her, or if need be assist him or her, to maintain contact with his or her close family (CASE OF RAZVOZZHAYEV v. RUSSIA AND UKRAINE AND UDALTSOV v. RUSSIA § 270, CASE OF VOYNOV v. RUSSIA § 49, CASE OF POLYAKOVA AND OTHERS v. RUSSIA § 81, CASE OF SAIDOV AND OTHERS v. RUSSIA § 7, CASE OF KHAMASTKHANOVA AND OTHERS v. RUSSIA § 10, CASE OF YEVSYUKOV AND OTHERS v. RUSSIA § 7). The Court would also emphasize that, on the issue of family visits, Article 8 of the Convention requires States to take into account the interests of the prisoner and his or her relatives and family members (CASE OF RAZVOZZHAYEV v. RUSSIA AND UKRAINE AND UDALTSOV v. RUSSIA § 270, CASE OF VOYNOV v. RUSSIA § 49, CASE OF POLYAKOVA AND OTHERS v. RUSSIA § 81, CASE OF SAIDOV AND OTHERS v. RUSSIA § 7, CASE OF KHAMASTKHANOVA AND OTHERS v. RUSSIA § 10, CASE OF YEVSYUKOV AND OTHERS v. RUSSIA § 7). Furthermore, the Court would consider that placing a prisoner in a particular penal facility may raise an issue under Article 8 of the Convention if its effects on his or her private and family life go beyond the "normal" hardships and restrictions inherent in the very concept of imprisonment (CASE OF RAZVOZZHAYEV v. RUSSIA AND UKRAINE AND UDALTSOV v. RUSSIA § 270, CASE OF VOYNOV v. RUSSIA § 49, CASE OF POLYAKOVA AND OTHERS v. RUSSIA § 81, CASE OF SAIDOV AND OTHERS v. RUSSIA § 7, CASE OF KHAMASTKHANOVA AND OTHERS v. RUSSIA § 10, CASE OF YEVSYUKOV AND OTHERS v. RUSSIA § 7).
Mistral: The applicable article of the Convention would be Article 8, which guarantees the right to respect for private and family life (CASE OF YUDIN AND OTHERS v. RUSSIA § 3). The Court would justify the application of Article 8 to these facts by stating that it is an essential part of a prisoner's right to respect for family life that the authorities enable him or her, or if need be assist him or her, to maintain contact with his or her close family (CASE OF RAZVOZZHAYEV v. RUSSIA AND UKRAINE AND UDALTSOV v. RUSSIA § 270, CASE OF VOYNOV v. RUSSIA § 49, CASE OF POLYAKOVA AND OTHERS v. RUSSIA § 81, CASE OF SAIDOV AND OTHERS v. RUSSIA § 7, CASE OF KHAMASTKHANOVA AND OTHERS v. RUSSIA § 10, CASE OF YEVSYUKOV AND OTHERS v. RUSSIA § 7). The Court would also emphasize that, on the issue of family visits, Article 8 of the Convention requires States to take into account the interests of the prisoner and his or her relatives and family members (CASE OF RAZVOZZHAYEV v. RUSSIA AND UKRAINE AND UDALTSOV v. RUSSIA § 270, CASE OF VOYNOV v. RUSSIA § 49, CASE OF POLYAKOVA AND OTHERS v. RUSSIA § 81, CASE OF SAIDOV AND OTHERS v. RUSSIA § 7, CASE OF KHAMASTKHANOVA AND OTHERS v. RUSSIA § 10, CASE OF YEVSYUKOV AND OTHERS v. RUSSIA § 7). Furthermore, the Court would consider that placing a prisoner in a particular penal facility may raise an issue under Article 8 of the Convention if its effects on his or her private and family life go beyond the "normal" hardships and restrictions inherent in the very concept of imprisonment (CASE OF RAZVOZZHAYEV v. RUSSIA AND UKRAINE AND UDALTSOV v. RUSSIA § 270, CASE OF VOYNOV v. RUSSIA § 49, CASE OF POLYAKOVA AND OTHERS v. RUSSIA § 81, CASE OF SAIDOV AND OTHERS v. RUSSIA § 7, CASE OF KHAMASTKHANOVA AND OTHERS v. RUSSIA § 10, CASE OF YEVSYUKOV AND OTHERS v. RUSSIA § 7).
Saul: The Convention article that would be applicable to a situation where a foreign prisoner is denied any contact with their family, including through telephone conversations or occasional visits, would be Article 8 (CASE OF POLYAKOVA AND OTHERS v. RUSSIA § 81, CASE OF SAIDOV AND OTHERS v. RUSSIA § 7). The Court justifies the application of Article 8 to these facts because it establishes that it is an essential part of a prisoner's right to respect for family life that the authorities enable them, or if need be, assist them, to maintain contact with their close family.

Table 9: Example Edit by Saul using Post-hoc Retrieval - RARR

Question: How does the Court differentiate between the rights of transgender individuals and same-sex couples in relation to the right to marry under Article 12, and what justifies the distinction between a total ban on marriage for transgender individuals and a total ban on same-sex marriage?

Answer: The European Court of Human Rights (ECtHR) acknowledges that the rights of transgender individuals and same-sex couples are distinct under the ambit of Article 12, which provides the right to marry. In evaluating the rights of transgender individuals, the ECtHR understands that their legal status is that of a person of the sex assigned at birth and their requests to get married relate to their acquired gender. **<base>**In contrast, same-sex couples' requests to marry involve the right to form a different-sex marriage, which is not their legal status.**</base>** **<edit>**Same-sex couples' requests to marry involve the right to form a same-sex marriage, which is not their legal status.**</edit>** The Court also differentiates between the circumstances surrounding the two groups' right to marry. Transgender individuals may suffer from legal impediments such as statutory bans on marriages between persons of the same sex, making it challenging to marry after undergoing gender reassignment surgery. In contrast, same-sex couples do not face legal obstacles in most European States when they try to get married, and their request is to change the legal description of marriage rather than facilitate marriage for the same-sex couple. Furthermore, the ECtHR recognizes the difference in the justifications behind the ban on marriage for transgender individuals and a ban on same-sex marriage. A ban on marriage for transgender individuals, specifically when a transgender person has undergone gender reassignment surgery, infringes on the right to marry under Article 12, as it hinders their ability to marry. In this context, a total ban on marriage of transgender individuals is discriminatory and unlawful. On the other hand, a total ban on same-sex marriage stems from the aim of protecting the traditional institution of marriage being established between a man and a woman (e.g. Obergefell, cited above). The ECtHR acknowledges that this difference in justifications means that while it might be acceptable to impose a total ban on same-sex marriage for the protection of the traditional form of marriage, it is unlawful to impose a total ban on transgender individuals' right to marry due to the disparate nature of the justifications behind the restrictions.

Table 10: Example Edit by Llama 8b using Post-hoc Retrieval - RARR

Question: In a case involving the revocation of a professional license due to pending criminal proceedings, under what conditions would the Court find that a State has failed to fulfill its positive obligations under Article 8 of the Convention?

Answer: A great question! To determine whether a State has failed to fulfill its positive obligations under Article 8 of the European Convention on Human Rights, we need to carefully examine the circumstances of the case. Article 8 of the Convention, which protects the right to respect for private and family life, has an "interface" with the concept of a professional license, which is essential for a person's livelihood and career. The Court has developed a nuanced approach to assess whether a State has breached its positive obligations under Article 8 when revoking a professional license due to pending criminal proceedings. Here are the key conditions:
1. **Interference with the right to respect for private and family life**: The Court will first assess whether the revocation of the professional license interferes with the individual's right to respect for private and family life, including their professional life. This may involve considering the gravity of the pending criminal charges and the impact of the revocation on the individual's ability to maintain their livelihood and contribute to their family's well-being. 2. ** Proportionality * *: **<base>**The Court will then consider whether the revocation of the professional license is proportionate to the legitimate aim sought to be achieved, such as maintaining public trust and safety.**</base>** **<edit>**The Court will then consider whether the revocation of the professional license is proportionate to the legitimate aim sought to be achieved, such as protecting the applicant's private life and reputation.**</edit>**
<base>The Court will evaluate the relationship between the measure and the aim, considering factors like the seriousness of the criminal charges, the individual's professional track record, and the potential risks to the public.**</base>** **<edit>**The Court will evaluate the proportionality of the measure, considering factors like the seriousness of the criminal charges, the individual's professional track record, and the potential risks to the public.**</edit>**
3. **Alternative measures**: The Court will examine whether the State had available alternative measures to achieve its aim, such as suspending the license instead of revoking it, imposing administrative sanctions, or providing support to the individual to prevent recidivism. 4. **Reasonable relationship**: The Court will assess whether there is a reasonable relationship between the measures taken and the legitimate aim sought to be achieved. The measures should be suitable, necessary, and proportionate to the aim.
If the State fails to demonstrate that its measures are reasonable and proportionate, and that alternative measures were not available or considered, the Court may find that the State has failed to fulfill its positive obligations under Article 8. In conclusion, for the Court to find that a State has failed to fulfill its positive obligations under Article 8, the State's measures to revoke a professional license due to pending criminal proceedings must be disproportionate, unreasonable, and lacking in proportionality. The Court will scrutinize the measures taken, considering the individual's right to respect for private and family life, the proportionality of the measures, and the availability of alternative measures.

Table 11: Example Edit by Llama 8b using Post-hoc Retrieval - RARR

Question: In the context of employment restrictions, why did the Court find that the imposition of such restrictions on former KGB employees in the private sector cannot be justified under Article 14 of the Convention taken in conjunction with Article 8?
Answer: A great question! In the landmark case of E.B. v. France, the European Court of Human Rights (ECtHR) ruled that the imposition of employment restrictions on former KGB employees in the private sector cannot be justified under Article 14 of the Convention taken in conjunction with Article 8. In essence, the Court held that the restrictions violated the applicants' right to respect for their private and family life (Article 8) and their rights under the non-discrimination principle (Article 14). Here's why: 1. **Lack of proportionality**: The restrictions were deemed to be disproportionate to the legitimate aim of preventing the dissemination of classified information. The Court found that the sanctions were too severe and did not take into account the applicants' individual circumstances or the nature of their work. 2. **Arbitrary and unpredictable application**: **<base>**The Court criticized the lack of clear and stable criteria for applying the employment restrictions.**</base>** **<edit>**The Court criticized the lack of a clear legal basis for the domestic authorities' decisions in applying the employment restrictions.**</edit>** **<base>**This led to arbitrary and unpredictable decision-making, which undermined the applicants' trust in the system.**</base>** **<edit>**The Court criticized the lack of clear and transparent decision-making, which undermined the applicants' trust in the system.**</edit>** 3. **Disproportionate impact on certain groups**: The restrictions disproportionately affected certain groups, such as former KGB employees in the private sector, who had already undergone strict screening and had been vetted by their employers. The Court found that these individuals were being punished twice for the same "offense," first during their time as government employees and again in their private sector careers. 4. **No effective domestic remedy**: The applicants had no effective domestic remedy to challenge the employment restrictions. The Court held that the applicants' right to an effective remedy (Article 13) was also violated. In conclusion, the ECtHR ruled that the employment restrictions on former KGB employees in the private sector were not justified under Article 14 taken in conjunction with Article 8 because they were arbitrary, disproportionate, and lacked clear criteria for application. **<base>**The restrictions were deemed to be a disproportionate interference with the applicants' rights, including their right to respect for their private and family life.**</base>** **<edit>**The restrictions were deemed to be in accordance with the law and necessary in a democratic society, including their impact on the applicants' right to respect for their private and family life.**</edit>**