Interpretable Long-Form Legal Question Answering with Retrieval-Augmented Large Language Models

Interpretable Long-Form Legal Question Answering with Retrieval-Augmented Large Language Models

Antoine Louis, Gijs van Dijck, Gerasimos Spanakis

Law & Tech Lab, Maastricht University
{a.louis, gijs.vandijck, jerry.spanakis}@maastrichtuniversity.nl

Abstract

Many individuals are likely to face a legal dispute at some point in their lives, but their lack of understanding of how to navigate these complex issues often renders them vulnerable. The advancement of natural language processing opens new avenues for bridging this legal literacy gap through the development of automated legal aid systems. However, existing legal question answering (LQA) approaches often suffer from a narrow scope, being either confined to specific legal domains or limited to brief, uninformative responses. In this work, we propose an end-to-end methodology designed to generate *long-form* answers to *any* statutory law questions, utilizing a "retrieve-then-read" pipeline. To support this approach, we introduce and release the Long-form Legal Question Answering (LLeQA) dataset, comprising 1,868 expert-annotated legal questions in the French language, complete with detailed answers rooted in pertinent legal provisions. Our experimental results demonstrate promising performance on automatic evaluation metrics, but a qualitative analysis uncovers areas for refinement. As one of the only comprehensive, expert-annotated long-form LQA dataset, LLeQA has the potential to not only accelerate research towards resolving a significant real-world issue, but also act as a rigorous benchmark for evaluating NLP models in specialized domains. We publicly release our code, data, and models.

Introduction

Legal disputes are an inevitable part of everyday life, with many individuals finding themselves entangled in issues related to marriage, debts, or employment (Farrow et al. 2016; Ponce et al. 2019). However, most people have little to no knowledge about their rights and fundamental legal processes (Balmer et al. 2010). As a result, they either take no action or turn to the internet for advice (Denvir 2016). Unfortunately, the latter often directs users towards commercial websites that prioritize their own marketing efforts over providing thorough, useful legal guidance (Hagan and Li 2020). While invaluable, expert legal assistance is often prohibitively expensive, which results in a considerable number of vulnerable individuals being left unprotected or exploited due to their inability to afford it. This barrier to accessing legal information fosters a significant imbalance within the legal system, impeding the universal right to equal access to justice for all. The global implications of this issue are significant: an estimated 1.5 billion individuals wrestle with unresolved legal challenges, and 4.5 billion may be excluded from the protective measures that the law provides (Garavano et al. 2019). In light of these circumstances, there is growing consensus that improved access to legal information could dramatically enhance the outcomes of legal disputes for many people (Currie 2009).

The rapid progress in natural language processing and the growing availability of digitized legal data present unprecedented opportunities to bridge the gap between people and the law. For instance, legal text summarization (Bhattacharya et al. 2019; Shukla et al. 2022) holds the potential to simplify complex legal documents for the layperson, while legal judgment prediction (Chalkidis et al. 2019; Trautmann et al. 2022) could unveil insightful correlations between an individual's situation and the probable legal outcome. Similarly, legal question answering (LQA) could offer affordable, expert-like assistance to the masses, thereby empowering marginalized parties when utilized for public welfare. However, existing research on LQA tends to exhibit a constrained scope, often concentrating on specialized legal domains, such as tax law (Holzenberger et al. 2020) or privacy policies (Ravichander et al. 2019), or limiting the responses to uninformative brief answers like yes/no replies (Rabelo et al. 2022) or few-word spans (Duan et al. 2019).

In this paper, we present an end-to-end approach aimed at generating *long-form* responses to *any* statutory law questions. Our methodology harnesses the popular "retrieve-then-read" pipeline, which first leverages a retriever over a large evidence corpus to fetch a set of relevant legislative articles, and then employs a reader to peruse these articles and formulate a comprehensive, interpretable answer. Our retriever relies on a lightweight bi-encoder model, wich enables fast and effective retrieval. For our reader, we use an instruction-tuned large language model (LLM) that we adapt to our task via two distinct learning strategies: in-context learning, wherein the model learns from instructions and a set of contextually provided examples; and parameter-efficient finetuning, where a small number of extra parameters are optimized on a downstream dataset while the base model's weights are quantized and remain unchanged.

To support training and evaluating such systems, we collect and release the Long-form Legal Question Answering

(LLeQA) dataset. LLeQA builds upon BSARD (Louis and Spanakis 2022), an information retrieval dataset in French comprising 1,108 legal questions labeled with relevant provisions from a corpus of 22,633 Belgian law articles, and enhance it in two ways. First, we introduce 760 new legal questions (+69%) and 5,308 additional statutory articles (+23%). Second, we supplement the data with new types of annotations, including an exhaustive taxonomy for the question, the jurisdictions concerned, the exact paragraph-level references within the relevant articles, and a comprehensive answer written by seasoned legal professionals. Owing to the rich variety of its annotations, LLeQA serves as a multifaceted resource that extends its utility beyond legal question answering and has the potential to catalyze significant progress in various legal tasks, such as legal inquiry classification, legal topic modeling, and legal information retrieval.

Our experimental results show that retrieval-augmented LLMs exhibit commendable performance on automatic evaluation metrics, measuring alignment with target answers. Yet, a deeper qualitative analysis reveals that these syntactically correct responses, despite seemingly covering the intended topics, frequently harbor inaccuracies and erroneous information. This discrepancy underscores the limitations inherent in relying solely on automatic metrics for assessing such systems, and indicates substantial room for improvement both in terms of modeling and evaluation.

In summary, our main contributions are:

1. A novel dataset for long-form question answering (LFQA) in the legal domain and French language, comprising 1,868 legal questions, meticulously annotated by legal professionals, with detailed answers and references to relevant legal provisions, drawn from a substantial knowledge corpus containing 27,942 statutory articles.
2. A comprehensive evaluation of the retrieve-then-read framework in the context of legal LFQA, while emphasizing interpretability and exploring various learning strategies for the reader.
3. A public release of our code, dataset, and checkpoints to facilitate future research on interpretable legal LFQA.[1]

Related Work

Legal question answering. Addressing legal questions has long posed intricate challenges within the legal NLP community, stemming from the inherent complexities of legal texts, including specialized terminology, complex structure, and nuanced temporal and logical connections. To stimulate advancement in this field, an array of datasets and benchmarks has emerged. Duan et al. (2019) craft a judicial reading comprehension dataset in the Chinese language, aimed at fostering the development of systems capable of mining fine-grained elements from judgment documents. Ravichander et al. (2019) present a corpus of questions about privacy policies of mobile applications with the objective of empowering users to comprehend and selectively investigate privacy matters. Holzenberger et al. (2020) introduce a dataset for statutory reasoning in tax law. Zhong et al. (2020) present a multi-choice question answering dataset designed to asses professional legal expertise. Rabelo et al. (2022) hold a competition wherein a task consists in answering "yes" or "no" given a legal bar exam problem related to Japanese civil law. Lastly, both Mansouri and Campos (2023) and Chen et al. (2023a) offer a corpus featuring question-answer pairs in English and Chinese, respectively, sourced from online law-oriented forums.

Knowledge-grounded question answering. Mainstream approaches to tackling knowledge-intensive QA tasks commonly rely on external knowledge sources to enhance the answer prediction, such as collected documents (Voorhees 1999), web-pages (Kwok et al. 2001), or structured knowledge bases (Berant et al. 2013; Yu et al. 2017). These *open-book* models (Roberts et al. 2020) typically index the knowledge corpus before employing a retrieve-then-read pipeline to predict a response based on multiple supporting documents (Chen et al. 2017). To an extent, this paradigm can be likened to query-based multi-document summarization (Tombros and Sanderson 1998), where the objective lies in providing users with a succinct and precise overview of the top-ranked documents related to their queries. Query-driven summarization may adopt different methodologies, manifesting either in an extractive form, where specific portions of evidence text are selected (Otterbacher et al. 2009; Litvak and Vanetik 2017), or in an abstractive form, where the information is synthesized into new expressions (Nema et al. 2017; Baumel et al. 2018; Ishigaki et al. 2020).

Rationale generation. To gain insights into model predictions (Lei et al. 2016), recent advancements have explored the generation of abstractive textual explanations in areas such as commonsense reasoning (Rajani et al. 2019) and natural language inference (Kumar and Talukdar 2020). Alternatively, Lakhotia et al. (2021) proposed the extractive generation of predefined evidence markers instead of decoding raw explanations. Complementing generation, studies have concentrated on extracting rationales from evidence input segments (Bastings et al. 2019; Chalkidis et al. 2021), as well as analyzing saliency maps to underscore key input tokens instrumental to each prediction (Ribeiro et al. 2016).

The LLeQA Dataset

Dataset Construction

In this section, we describe our process to create LLeQA, which involves three main stages. First, we gather and refine annotated legal questions. Then, we build an expansive corpus of supportive statutory articles drawn from Belgian legislation. Finally, we enrich the question annotations by generating paragraph-level references within relevant articles. We elaborate upon each of these steps below.

Collecting question-answer pairs. The data construction process starts with collecting high-quality question-answer pairs on a legal matter. Echoing Louis and Spanakis (2022), we partner with Droits Quotidiens, a Belgian non-profit organization that endeavors to make the law comprehensible and accessible to the most vulnerable. To this end, the organization maintains a rich website featuring thousands of

[1] https://github.com/maastrichtlawtech/lleqa

Dataset	Average # of words			# Pairs	Answer type	Domain	Source	Lang.
	Ques.	Evid.	Ans.					
JEC-QA (Zhong et al. 2020)	47	58	15	26,365	Multi-choice	Statutory law	Law exam	zh
SARA (Holzenberger et al. 2020)	46	489	1	376	Binary, numeric	Tax law	Jurists	en
PrivacyQA (Ravichander et al. 2019)	8	3,237	140	1,750	Multi-span	Privacy policy	Jurists	en
CRJC (Duan et al. 2019)	*unk.*	*unk.*	*unk.*	51,333	Binary, span	Case law	Jurists	zh
FALQU (Mansouri et al. 2023)	144	-	244	9,880	Long-form	Statutory law	Web forum	en
COLIEE-21 (Rabelo et al. 2022)	41	94	1	887	Binary	Civil law	Law exam	ja, en
EQUALS (Chen et al. 2023a)	32	252	69	6,914	Long-form	Statutory law	Web forum	zh
LLeQA (ours)	15	1,857	264	1,868	Long-form	Statutory law	Jurists	fr

Table 1: Comparison of public legal question answering (LQA) datasets. LLeQA has answers an order of magnitude longer and is the only expert-annotated long-form LQA dataset to cover any statutory law subjects.

legal questions commonly posed by Belgian citizens. Each question comes with its own individual page, encompassing one or more categorizations, references to relevant legislative statutes, and a detailed answer written in layman's terms by experienced jurists. With their help, we collect approximately 2,550 legal questions. We then filter out questions that are unsuitable for retrieval-based question answering. Specifically, we discard questions whose references are either missing, too vague (e.g., an entire law or book), or from a statute not collected in our knowledge corpus. Additionally, we group duplicate questions found across different subcategories on the website. This yields a final number of 1,868 question-answer pairs, each with legal references.

Mining supporting information. Next, we build the knowledge corpus of statutory articles used to provide evidence that a system can draw upon when generating an answer. We start by extracting provisions from all publicly available Belgian codes of law via the official government website, forming an exhaustive foundation of 23,759 articles across 35 legal codes, thereby encapsulating a wide range of legal subjects. To enhance the corpus, we then incorporate additional laws, decrees, and ordinances that are frequently cited as supportive references but absent from the initial collection, adding 4,183 articles from 34 legal acts. This brings the final evidence corpus to 27,942 articles. We proceed by assigning a unique identifier to each article and employ regular expressions to match the plain text legal references linked to the questions with their corresponding article in our corpus. Next, we cleanse the articles of recurrent noisy textual elements, such as nested brackets, superscripts, or footnotes that may be present due to revisions or repeals by new legislation. Besides the article's content, we also collate the complete legislative path leading up to that article, which starts from the statute's name and progresses through the name of the book, act, chapter, section, and subsection where the article resides. This auxiliary information provides valuable contextual insight on the article's subject matter. Lastly, we partition the articles into their constituent paragraphs, which serve as the basic units for rationale extraction.

Generating paragraph-level rationales. Only 10.4% of the collected questions come with paragraph-level references, i.e., with mentions to specific article paragraphs as the relevant information to the question. To extend this level of interpretability across all questions in LLeQA, we leverage a large-scale language model to synthetically generate these paragraph-level relevance signals for the remaining samples. This approach serves as an affordable and efficient alternative to hiring costly legal experts for annotation. Utilizing the closed-source gpt-3.5-turbo-0613 model via the OpenAI API, we present the model with all paragraphs from relevant legal articles for each question-answer pair, and instruct it to identify those contributing to the answer. The model responds with a comma-separated list of identifiers corresponding to the deemed relevant paragraphs, or "None" if it discerns no contribution. After parsing the model's responses, we incorporate these synthetically generated paragraph-level rationales into the dataset.

Finalizing the dataset. We assign the questions with expert-annotated paragraph-level references to the test set and randomly partition the remaining samples into training and development sets, yielding a 10/90/10 split, respectively.

Dataset Analysis

Comparative overview. Table 1 presents a comparative review of existing LQA datasets, including ours, across several key factors, such as the length of textual elements, number of samples, type of answer, legal domain covered, source of annotations, and language used. LLeQA distinguishes itself by being the only dataset targeting any statutory law subjects, with long-form answers derived from legal experts. Its questions are succinct, intending to mirror a real-world scenario where laypeople may struggle to elaborate on their legal concerns. In contrast, the answers are more detailed than in other datasets, as they often compensate for the lack of provided information by exploring all potential scenarios contingent on an individual's circumstances, such as age, employment situation, or marital status.

Question diversity. In Table 3, we provide a breakdown of the major question subjects in LLeQA. Housing and healthcare represent the two largest topics, accounting for almost half of all questions together. Family, work, and immigration follow, collectively constituting over a third of the dataset, while money, privacy, and justice questions are less prevalent. We then examine the type of information requested in the questions based on their interrogative words, as shown

Word	(%)	Example Question
Can	33.9	Can I continue to work if I am retired?
How	21.2	How can I end my student lease ?
What	14.8	What is the role of the guardian of a minor?
Must	8.6	Must I say I am pregnant in an interview?
Who	3.9	Who has to pay the funeral expenses?
Which	3.8	Which costs are covered for work accidents?
When	1.6	When do I have to hand in my resignation?
Where	0.5	Where can I get my criminal record extract?
Why	0.1	Why shall I declare the birth of my child?
Misc.	11.6	Do my assets become joint after marriage?

Table 2: LLeQA questions by interrogative word. All examples in the paper are translated from French for illustration.

Topic	Train	Dev	Test	(%)
Housing	382	54	83	27.8
Healthcare	286	40	67	21.0
Family	217	22	16	13.7
Work	167	26	9	10.8
Immigration	156	22	3	9.7
Money	120	14	7	7.5
Privacy	80	14	10	5.6
Justice	64	9	0	3.9
Total	1472	201	195	-

Table 3: Topic distribution of questions in LLeQA.

in Table 2. "Can", "how", "what", and "must" are the most frequently used question words, indicating that people primarily seek information on legal permissions, procedures, definitions, and obligations. These distributions reflect the variety of legal concerns that citizens face, thereby providing valuable guidance for user-centric LQA systems.

Question evidence. In LLeQA, approximately 80% of the questions associate with fewer than five articles from the knowledge corpus, with the median number of relevant articles per question being two. These articles have a median length of 84 words, yet 1,515 articles exceed 500 words, with the longest reaching several thousand words. When combining all relevant articles for each question, we find an average evidence length of 1,857 words per question, positioning LLeQA at the upper range among published datasets regarding evidence length. Interestingly, a mere 8% of the articles within our corpus are referenced as relevant to at least one question within the dataset. Moreover, nearly half of these referenced articles originate from five statutes only, implying the critical role a select few laws play in answering the most frequently posed legal inquiries.

Assessment of annotation quality. To assess the quality of the synthetically generated paragraph-level rationales, we evaluate the performance of gpt-3.5-turbo-0613 against the ground truth annotations from the test set. Although far from expert quality, we find that the model demonstrates decent annotation performance, achieving a F1 score of 47.5%. By comparison, a naive baseline that randomly selects the relevant paragraphs achieves 15.3% F1, whereas one that always marks the first paragraph as the relevant one scores 27.2% F1. After experimenting with alternative LLMs, we observe that, as of the time of writing, gpt-3.5-turbo-0613 achieves the best overall performance within a limited cost budget. Despite the apparent margin of error, we believe that these imperfect synthetic annotations may still be beneficial for in-context learning purposes as ground truth labels bear less significance in such settings (Min et al. 2022).

Method

In this section, we detail the "retrieve-then-read" framework we use for interpretable long-form legal question answering, illustrated in Figure 1. First, a *retriever* selects a small subset of statutory articles, some of which being relevant to the question. Then, a *generator* conditions its answer on the subset of articles returned by the retriever. We describe these two components in more detail below.

Retriever

The role of our retrieval component is to pinpoint all statutory articles relevant to a question and present them at the forefront of the returned results. More formally, the retriever can be expressed as a function $R : (q, \mathcal{C}) \mapsto \mathcal{F}$ that takes as input a question q along with a knowledge corpus of legal provisions $\mathcal{C} = \{p_1, p_2, \cdots, p_N\}$, and returns a notably smaller filtered set $\mathcal{F} \subset \mathcal{C}$ of the supposedly relevant provisions, ordered by decreasing relevance.

We employ the widely adopted bi-encoder architecture (Bromley et al. 1993) as the foundation of our retriever, which maps a question q and a legal provision p into dense vector representations and computes a relevance score $s : (q, p) \mapsto \mathbb{R}_+$ between the two by the similarity of their embeddings $\boldsymbol{h}_q, \boldsymbol{h}_p \in \mathbb{R}^d$, i.e.,

$$s(q, p) = \mathrm{sim}\left(\boldsymbol{h}_q, \boldsymbol{h}_p\right), \tag{1}$$

where $\mathrm{sim} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a similarity function, such as the dot product or cosine similarity. These embeddings are typically derived from a pooling operation on the output representations of a pretrained autoencoding language model, such as BERT (Devlin et al. 2019), so that

$$\begin{aligned} \boldsymbol{h}_q &= \mathrm{pool}\left(E(q; \theta_1)\right), \text{and} \\ \boldsymbol{h}_p &= \mathrm{pool}\left(E(p; \theta_2)\right), \end{aligned} \tag{2}$$

where the model $E(\cdot; \theta_i) : \mathcal{W}^n \to \mathbb{R}^{n \times d}$ with weights θ_i transforms an input text sequence of n terms from vocabulary $\mathcal{W}$ to d-dimensional real-valued word vectors. The pooling function $\mathrm{pool} : \mathbb{R}^{n \times d} \to \mathbb{R}^d$ implements a mean or max operation on the output word embeddings to distill a global representation for the text passage. Beyond the conventional *two-tower* bi-encoder (Karpukhin et al. 2020; Yang et al. 2020), which employs two independent encoder models to map queries and articles separately into distinct embedding spaces, the *siamese* variant (Reimers and Gurevych 2019; Xiong et al. 2021) uses a unique encoder, i.e., $\theta_1 = \theta_2$, to encode the question and article in a shared dense vector space. We use the latter variant in this work.

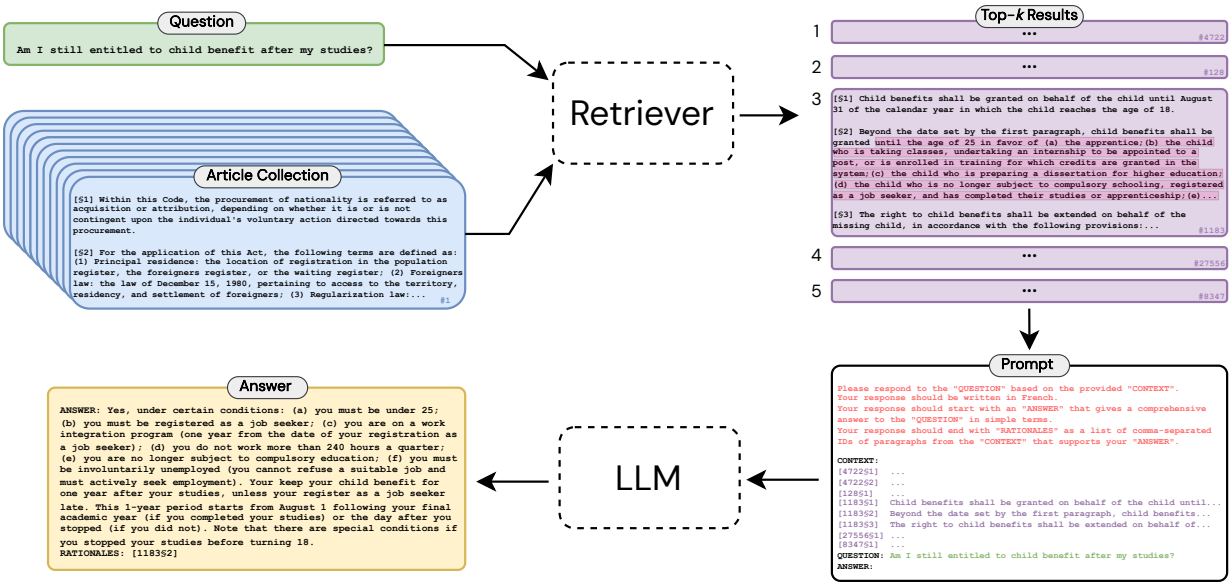

Figure 1: An illustration of the "retrieve-then-read" pipeline for interpretable long-form legal question answering.

The training objective for our dense retriever is to learn a high-quality low-dimensional embedding space for questions and legal provisions such that relevant question-provision pairs appear closer to each other than irrelevant ones. Assume our training data $\mathcal{D} = \{\langle q_i, p_i^+ \rangle\}_{i=1}^N$ contains N instances, each comprising a question q_i linked to a relevant provision p_i^+. For each question q_i, we sample a set of irrelevant provisions $\mathcal{P}_i^-$, thereby constituting a training set $\mathcal{T} = \{\langle qi, p_i^+, \mathcal{P}_i^- \rangle\}_{i=1}^N$. Subsequently, we use the instances in $\mathcal{T}$ to contrastively optimize the negative log-likelihood of the relevant provision against the non-relevant ones, i.e.,

$$\mathcal{L}_\theta \left(q_i, p_i^+, \mathcal{P}_i^-\right) = -\log \frac{e^{s(q_i, p_i^+)/\tau}}{\sum_{p \in \mathcal{P}_i^- \cup \{p_i^+\}} e^{s(q_i,p)/\tau}}, \quad (3)$$

where $\tau > 0$ is a temperature parameter that we set to 0.01. To select irrelevant provisions, we employ two distinct sampling strategies: *random sampling* using in-batch negatives (Chen et al. 2017b; Henderson et al. 2017), which considers provisions paired with the other questions within the same mini-batch; and *hard negative sampling* using BM25 (Robertson et al. 1994), which includes the top provisions returned by BM25 that bear no relevance to the question.

Generator

Our generator aims at formulating a comprehensive answer to a short legal question, leaning on corroborative data. Formally, the task can be cast as a conditional text generation problem, where the model requires conditioning its response on a context string that incorporates the question and several supporting statutory articles. We turn to autoregressive large language models (LLMs) based on the Transformer's decoder block as the backbone architecture for our generator. We then delve into two learning scenarios, specifically in-context learning (Radford et al. 2019) and parameter-efficient finetuning (Lester et al. 2021; Liu et al. 2022).

In-context learning. To assess the innate performance of our generator without additional training, we start by examining three prevalent in-context learning strategies, namely *zero-shot*, *one-shot*, and *few-shot* learning. Under the zero-shot learning paradigm, the generator receives a natural language instruction and seeks to answer the question directly. The context c provided to the model is formulated as

$$c = \left[d, \mathcal{P}_t^+, q_t\right], \quad (4)$$

where d stands for the task description, q_t is the test question, and $\mathcal{P}_t^+$ represents the top-k most pertinent legal provisions to q_t as identified by the retriever (k is set to 5). In one-shot learning, the generator benefits from an additional demonstration to guide its understanding of the task, whereas in a few-shot setting, the model accommodates as many demonstrations as can be included within its context window. Under these scenarios, the context string becomes

$$c = \left[d, \left[\mathcal{P}_j^+, q_j, a_j\right]_{j=1}^n, \mathcal{P}_t^+, q_t\right], \quad (5)$$

where a_j is the gold long-form answer to the training question q_j, and n denotes the number of demonstrations. We dynamically select demonstrations in the training pool through a similarity-based retrieval method based on the question.

Parameter-efficient finetuning. Due to the gigantic size of contemporary LLMs, performing full finetuning of all model parameters would be prohibitively expensive. Instead, we employ parameter-efficient finetuning to train our generator, which significantly trims both the training duration and computational cost. Specifically, we apply the QLoRA technique (Dettmers et al. 2023), which undertakes a preliminary quantization of the pretrained model to 4-bit before freezing all its parameters for training. A small set of learnable low-rank adapter weights (Hu et al. 2022) are then injected into each linear layer of the Transformer and tuned by

backpropagating gradients through the quantized weights. Formally, given a training sample $(q_i, \mathcal{P}_i^+, a_i)$ where $a_i = (y_1, \cdots, y_T)$ represents the target output sequence, we optimize the adapter's parameters ϕ using a standard language modeling objective that maximizes the negative log likelihood of generating the target answer a_i conditioned on the input context, encompassing the source question q_i and the set of relevant legal provisions $\mathcal{P}_i^+$, i.e.,

$$\begin{aligned}\mathcal{L}_\phi\left(q_i, \mathcal{P}_i^+, a_i\right) &= -\log p_\phi\left(a_i \mid \mathcal{P}_i^+, q_i\right) \\ &= -\log \prod_{t=1}^{T} p_\phi\left(y_t \mid \mathcal{P}_i^+, q_i, y_{<t}\right).\end{aligned} \tag{6}$$

Context window extension. LLMs typically come with a predefined context window limit, beyond which perplexity steeply rises due to the weak extrapolation properties of positional encoding. This limitation poses significant challenges for applications requiring the processing of extensive inputs, like ours. Recent efforts have aimed to extend the context window sizes of pretrained LLMs employing rotary position embedding (Su et al. 2021, RoPE), such as LLaMA (Touvron et al. 2023), by interpolating positional encoding (Chen et al. 2023b). Guided by promising findings from the open-source community, we perform dynamic NTK-aware scaling,[2] which retains the exact position values within the original context window and progressively down-scales the input position indices using a nonlinear interpolation from neural tangent kernel theory (Jacot et al. 2018). Preliminary results suggest this approach substantially mitigates perplexity degradation for sequences exceeding the maximum window size, without necessitating additional finetuning.

Rationales extraction. Given the serious implications of flawed legal guidance, ensuring interpretability in generated answers is crucial. This enables users to cross-verify responses through reliable sources while understanding the underlying reasoning, thereby enhancing the trustworthiness of LQA systems. To this end, we impose additional constraints on the model to furnish proper justification for its answers. While prior work on rationale generation has predominantly focused on creating free-form natural language explanations (Latcinnik and Berant 2020; Narang et al. 2020), abstractive models have shown a propensity for fabricating convincing yet misleading justifications, inadvertently supporting inaccurate predictions (Camburu et al. 2020; Wiegreffe et al. 2021). Besides, adapting this strategy for applications involving multiple extensive evidence documents proves challenging. Consequently, we adopt an extractive rationale generation strategy (Lakhotia et al. 2021), prompting the model to generate evidence paragraph markers rather than raw explanations. This technique ensures the production of unaltered rationales that are easily interpretable.

Experiments

Experimental Setup

Models. Our dense retrieval model leverages CamemBERT (Martin et al. 2020), a leading autoencoding model for the French language. To furnish a well-rounded comparative analysis, we incorporate two robust retrieval baselines: BM25 (Robertson et al. 1994), a widely utilized bag-of-words retrieval function; and mE5 (Wang et al. 2022a), currently the top-performing multilingual dense model on the MTEB benchmark (Muennighoff et al. 2023). With regard to our generator, we experiment with several instruction-tuned open-source models derived from LLaMA (Touvron et al. 2023) to harness the benefits of dynamic NTK-aware scaled RoPE. Specifically, we consider four models that are notably high-ranking on the MT-Bench leaderboard (Zheng et al. 2023): Vicuna-1.3 (Chiang et al. 2023), WizardLM-1.0 (Xu et al. 2023a), TÜLU (Wang et al. 2023), and Guanaco (Dettmers et al. 2023). Due to limited computational resources, we restrict our study to their 7B variant and curtail the extended context window size to 8192 tokens for in-context learning and 4096 tokens for finetuning.

Implementation. We start by fully finetuning CamemBERT on LLeQA with a batch size of 32 and a maximum sequence length of 384 tokens for 20 epochs (i.e., 5.8k steps), using AdamW (Loshchilov and Hutter 2017) with $\beta_1 = 0.9$, weight decay of 0.01, and learning rate warm up along the first 60 steps to a maximum value of 2e-5, after which linear decay is applied. We use 16-bit automatic mixed precision to accelerate training, which takes about 1.7 hours. We then optimize our baseline LLMs through 4-bit QLoRA finetuning with an effective batch size of 8 for 10 epochs (i.e., 1.1K steps) using paged AdamW optimizer with default momentum parameters and constant learning rate schedule of 2e-4. We employ NormalFloat4 with double quantization for the base models and add LoRA adapters on all linear layers by setting $r = 16$, $\alpha = 32$ while utilizing float16 as computation datatype. Training takes around 7.5 hours to complete. We generate from the LLMs using nucleus sampling (Holtzman et al. 2020) with $p = 0.95$ and a temperature of 0.1.

Hardware & Libraries. Computations are performed on a single 32GB NVIDIA V100 GPU hosted on a server with a dual 20-core Intel Xeon E5-2698 v4 CPU and 512GB of RAM, operating under Ubuntu 16.04.

Automatic Evaluation

Evaluation metrics. To evaluate our framework's effectiveness, we assess three core aspects: *retrieval performance*, *generation quality*, and *rationales accuracy*. Firstly, it is essential that the retriever returns as many pertinent provisions as possible within the first top-k results, given the generator's limited context window size. This requirement implies a primary interest in recall at small cutoffs. Additionally, we report the mean reciprocal rank, as it offers valuable insights into the position of the first relevant result. Gauging the quality of long-form answers presents more intricate challenges. Automatic metrics, such as ROUGE (Lin 2004), proved to be inadequate due to the intrinsically open-ended nature of long-form responses, which allows for an array of possible formulations that retain semantic similarity (Wang et al. 2022b; Xu et al. 2023b). Besides, these lexical overlap metrics fail to assess essential aspects such as factual correctness and query relevance. Ultimately, a thorough assessment of

[2] https://reddit.com/r/LocalLLaMA/comments/14mrgpr/

	Model	Size	R@5	R@10	MRR@10
Baselines					
1	BM25	-	17.4	22.8	22.0
2	$mE5_{BASE}$	278M	15.4	21.7	25.8
3	$mE5_{LARGE}$	560M	16.5	26.7	28.3
4	CamemBERT	111M	**48.6**	**60.6**	**60.0**

Table 4: Retrieval scores of our dense retriever benchmarked against other strong retrieval baselines on LLeQA dev set.

	Model	0-shot	1-shot	2-shot	Fine-tuned
	Question Answering (METEOR)				
1	Vicuna	11.6	16.2	15.3	19.7
2	WizardLM	12.3	15.5	16.6	20.4
3	TÜLU	2.9	4.6	8.5	12.7
4	Guanaco	11.2	11.2	11.3	20.1
	Rationales Extraction (F1)				
1	Vicuna	0.4	0.6	0.2	0.0
2	WizardLM	0.0	0.0	0.0	2.0
3	TÜLU	0.1	0.0	0.0	3.5
4	Guanaco	1.3	0.4	0.0	0.0

Table 5: Scores of our baseline LLMs on LLeQA test set.

such systems requires human evaluation, although the latter introduces its own set of challenges (Krishna et al. 2021), in addition to being expensive and noisy. Owing to resource constraints in our study, we opt for an automated evaluation metrics, fully cognizant of its limitations, and earmark human evaluation for future work. In particular, we report METEOR (Banerjee and Lavie 2005), which demonstrated superior correlation with human judgment in the context of long text generation (Sharma et al. 2017; Chen et al. 2022). Lastly, we evaluate the model's capacity for rationale extraction by measuring the F1 score over the set of predicted paragraph markers as compared to the ground truth markers.

Results. As depicted in Table 4, our dense retriever, finetuned on a mere 1.5k in-domain examples, significantly outperforms robust retrieval baselines, underlining the essential role of domain adaptation in enhancing performance. However, the results leave substantial room for improvement; on average, less than half of the relevant articles appear within the top five returned results. This shortfall represents a major bottleneck for our generator, which faces the challenge of answering questions based on partially irrelevant provisions. The impact of this limitation is palpable in the poor performance of rationale extraction, which approaches near-zero effectiveness for most LLMs, though part of this deficiency can be attributed to the models' tendency to hallucinate. In terms of answer quality, WizardLM and Vicuna display a high degree of overlap with the ground truth responses, indicating accurate engagement with the subject matter. Providing a demonstration appears to improve generation quality as compared to a zero-shot setup, except for Guanaco which shows strong zero-shot results. However, performance does not seem to vary significantly when more demonstrations are provided. Finally, the results suggest that finetuning on our task-specific dataset consistently enhances performance.

Qualitative Analysis

To discern the strengths and shortcomings of our generators, we conduct a detailed manual analysis of 10 randomly selected samples from the test set. We find that TÜLU exhibits a propensity for producing concise answers, a tendency likely due to its extensive finetuning on instruction datasets averaging a relatively short completion length of 98.7 words, which also explains its low METEOR score as LLeQA answers are markedly longer. We further observe that Guanaco and WizardLM are prone to repetitiveness in their responses, occasionally echoing identical phrases. While presence and frequency penalties could mitigate this issue, they may prove ineffective in instances addressing specialized topics where the same terminology is intrinsically repeated. Regarding response quality, WizardLM and Vicuna stand out, far exceeding Guanaco, which tends to produce nonsensical or linguistically convoluted sentences. In contrast, WizardLM and Vicuna's outputs are articulated in impeccable French, displaying a persuasive flair that could potentially mislead an unsuspecting reader. Nevertheless, a deeper probe unveils striking hallucinations (Ji et al. 2023). Despite seemingly addressing the question, many facts, dates, sources, and conditions appear to be fabricated, as if the models leveraged the provided context less for factual accuracy and more as a foundation upon which to weave a convincing fictitious answer.

Conclusion

In this work, we introduce LLeQA, an expert-annotated dataset tailored to facilitate the development of models aimed at generating comprehensive answers to legal questions, while supplying interpretable justifications. We experiment with the "retrieve-then-read" pipeline on LLeQA and explore various state-of-the-art large language models as readers, that we adapt to the task using several learning strategies. We find that this framework tends to produce syntactically correct answers pertinent to the question's subject matter but occasionally fabricate facts. Overall, we believe LLeQA can serve as a robust foundation for advancements in interpretable, long-form legal question answering, thereby contributing to the democratization of legal access.

Limitations and future work. Despite our efforts to make LQA systems more factually grounded in supporting legal provisions, the framework we propose remains vulnerable to hallucinations in both the constructed answers and associated rationales. Additionally, consistent with prior studies (Krishna, Roy, and Iyyer 2021; Xu et al. 2023b), we observe that conventional automatic metrics may not accurately mirror answer quality, leading to potential misinterpretations. These challenges present great avenues for future work.

Ethical considerations. The premature deployment of LQA systems poses a tangible risk to laypersons, who may uncritically rely on the furnished guidance and inadvertently exacerbate their circumstances. To ensure the responsible development of legal aid technologies, we are committed to limiting the use of our dataset strictly to research purposes.

Acknowledgments

This research is partially supported by the Sector Plan Digital Legal Studies of the Dutch Ministry of Education, Culture, and Science. In addition, this research was made possible, in part, using the Data Science Research Infrastructure (DSRI) hosted at Maastricht University.

References

Balmer, N. J.; Buck, A.; Patel, A.; Denvir, C.; and Pleasence, P. 2010. Knowledge, capability and the experience of rights problems. *London: PLEnet*.

Banerjee, S.; and Lavie, A. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and Summarization*, 65–72.

Bastings, J.; Aziz, W.; and Titov, I. 2019. Interpretable Neural Predictions with Differentiable Binary Variables. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2963–2977.

Baumel, T.; Eyal, M.; and Elhadad, M. 2018. Query Focused Abstractive Summarization: Incorporating Query Relevance, Multi-Document Coverage, and Summary Length Constraints into seq2seq Models. *CoRR*, abs/1801.07704.

Berant, J.; Chou, A.; Frostig, R.; and Liang, P. 2013. Semantic Parsing on Freebase from Question-Answer Pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1533–1544.

Bhattacharya, P.; Hiware, K.; Rajgaria, S.; Pochhi, N.; Ghosh, K.; and Ghosh, S. 2019. A Comparative Study of Summarization Algorithms Applied to Legal Case Judgments. *Advances in Information Retrieval*, 11437: 413–428.

Bromley, J.; Guyon, I.; LeCun, Y.; Säckinger, E.; and Shah, R. 1993. Signature Verification Using a Siamese Time Delay Neural Network. *Advances in Neural Information Processing Systems*, 6: 737–744.

Camburu, O.; Shillingford, B.; Minervini, P.; Lukasiewicz, T.; and Blunsom, P. 2020. Make Up Your Mind! Adversarial Generation of Inconsistent Natural Language Explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4157–4165.

Chalkidis, I.; Androutsopoulos, I.; and Aletras, N. 2019. Neural Legal Judgment Prediction in English. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, 4317–4323.

Chalkidis, I.; Fergadiotis, M.; Tsarapatsanis, D.; Aletras, N.; Androutsopoulos, I.; and Malakasiotis, P. 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*, 226–241.

Chen, A.; Yao, F.; Zhao, X.; Zhang, Y.; Sun, C.; Liu, Y.; and Shen, W. 2023a. EQUALS: A Real-world Dataset for Legal Question Answering via Reading Chinese Laws. In *Proceedings of the 19th International Conference on Artificial Intelligence and Law*, 71–80.

Chen, D.; Fisch, A.; Weston, J.; and Bordes, A. 2017a. Reading Wikipedia to Answer Open-Domain Questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 1870–1879.

Chen, S.; Wong, S.; Chen, L.; and Tian, Y. 2023b. Extending Context Window of Large Language Models via Positional Interpolation. *CoRR*, abs/2306.15595.

Chen, T.; Sun, Y.; Shi, Y.; and Hong, L. 2017b. On Sampling Strategies for Neural Network-based Collaborative Filtering. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 767–776.

Chen, Y.; Song, Z.; Wu, X.; Wang, D.; Xu, J.; Chen, J.; Zhou, H.; and Li, L. 2022. MTG: A Benchmark Suite for Multilingual Text Generation. In *Findings of the Association for Computational Linguistics: NAACL 2022*, 2508–2527.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; Stoica, I.; and Xing, E. P. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality. https://lmsys.org/blog/2023-03-30-vicuna/. Accessed: 2023-12.

Currie, A. 2009. The legal problems of everyday life. In *Access to justice*, 1–41. Emerald Group Publishing Limited.

Denvir, C. 2016. Online and in the know? Public legal education, young people and the Internet. *Computers & Education*, 92-93: 204–220.

Dettmers, T.; Pagnoni, A.; Holtzman, A.; and Zettlemoyer, L. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. *CoRR*, abs/2305.14314.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 4171–4186.

Duan, X.; Wang, B.; Wang, Z.; Ma, W.; Cui, Y.; Wu, D.; Wang, S.; Liu, T.; Huo, T.; Hu, Z.; Wang, H.; and Liu, Z. 2019. CJRC: A Reliable Human-Annotated Benchmark Dataset for Chinese Judicial Reading Comprehension. In *Proceedings of the 18th China National Conference on Computational Linguistics*, 439–451.

Farrow, T. C.; Currie, A.; Aylwin, N.; Jacobs, L.; Northrup, D.; and Moore, L. 2016. Everyday legal problems and the cost of justice in Canada: Overview report. *Osgoode Legal Studies Research Paper*, 12(57).

Garavano, G. C.; Kaag, S.; Schwartz, P.; Jilani, H.; Alvarez, A.; Ardyanto, D.; Goldston, J.; de Greiff, P.; Hossain, S.; Kennou, K.; Maru, V.; Maynard-Gibson, A.; Molokomme, A.; Pell, O.; Pais, M. S.; Rodriguez, M. F.; van Wieren, J.; and Osho, B. 2019. *Justice for All: The Report of the Task Force on Justice*. Center on International Cooperation.

Hagan, M.; and Li, Y. 2020. Legal Help Search Audit: Are Search Engines Effective Brokers of Legal Information? *Available at SSRN 3623333*.

Henderson, M. L.; Al-Rfou, R.; Strope, B.; Sung, Y.; Lukács, L.; Guo, R.; Kumar, S.; Miklos, B.; and Kurzweil, R. 2017. Efficient Natural Language Response Suggestion for Smart Reply. *CoRR*, abs/1705.00652.

Holtzman, A.; Buys, J.; Du, L.; Forbes, M.; and Choi, Y. 2020. The Curious Case of Neural Text Degeneration. In *Proceedings of the 8th International Conference on Learning Representations*.

Holzenberger, N.; Blair-Stanek, A.; and Durme, B. V. 2020. A Dataset for Statutory Reasoning in Tax Law Entailment and Question Answering. In *Proceedings of the Natural Legal Language Processing Workshop 2020*, 31–38.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *Proceedings of the 10th International Conference on Learning Representations*.

Ishigaki, T.; Huang, H.; Takamura, H.; Chen, H.; and Okumura, M. 2020. Neural Query-Biased Abstractive Summarization Using Copying Mechanism. *Advances in Information Retrieval*, 12036: 174–181.

Jacot, A.; Hongler, C.; and Gabriel, F. 2018. Neural Tangent Kernel: Convergence and Generalization in Neural Networks. *Advances in Neural Information Processing Systems*, 31: 8580–8589.

Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.; Madotto, A.; and Fung, P. 2023. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12): 248:1–248:38.

Karpukhin, V.; Oguz, B.; Min, S.; Lewis, P. S. H.; Wu, L.; Edunov, S.; Chen, D.; and Yih, W. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 6769–6781.

Krishna, K.; Roy, A.; and Iyyer, M. 2021. Hurdles to Progress in Long-form Question Answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*, 4940–4957.

Kumar, S.; and Talukdar, P. P. 2020. NILE : Natural Language Inference with Faithful Natural Language Explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8730–8742.

Kwok, C. C. T.; Etzioni, O.; and Weld, D. S. 2001. Scaling question answering to the web. *ACM Transactions on Information Systems*, 19(3): 242–262.

Lakhotia, K.; Paranjape, B.; Ghoshal, A.; Yih, S.; Mehdad, Y.; and Iyer, S. 2021. FiD-Ex: Improving Sequence-to-Sequence Models for Extractive Rationale Generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 3712–3727.

Latcinnik, V.; and Berant, J. 2020. Explaining Question Answering Models through Text Generation. *CoRR*, abs/2004.05569.

Lei, T.; Barzilay, R.; and Jaakkola, T. S. 2016. Rationalizing Neural Predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 107–117.

Lester, B.; Al-Rfou, R.; and Constant, N. 2021. The Power of Scale for Parameter-Efficient Prompt Tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 3045–3059.

Lin, C. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, 74–81.

Litvak, M.; and Vanetik, N. 2017. Query-based summarization using MDL principle. In *Proceedings of the Workshop on Summarization and Summary Evaluation Across Source Types and Genres*, 22–31.

Liu, X.; Ji, K.; Fu, Y.; Tam, W.; Du, Z.; Yang, Z.; and Tang, J. 2022. P-Tuning: Prompt Tuning Can Be Comparable to Fine-tuning Across Scales and Tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 61–68.

Loshchilov, I.; and Hutter, F. 2017. Decoupled Weight Decay Regularization. In *Proceedings of the 7th International Conference on Learning Representations*.

Louis, A.; and Spanakis, G. 2022. A Statutory Article Retrieval Dataset in French. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 6789–6803.

Mansouri, B.; and Campos, R. 2023. FALQU: Finding Answers to Legal Questions. In *Proceedings of the 1st International Workshop on Legal Information Retrieval*, 22–24.

Martin, L.; Müller, B.; Suárez, P. J. O.; Dupont, Y.; Romary, L.; de la Clergerie, É.; Seddah, D.; and Sagot, B. 2020. CamemBERT: a Tasty French Language Model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7203–7219.

Min, S.; Lyu, X.; Holtzman, A.; Artetxe, M.; Lewis, M.; Hajishirzi, H.; and Zettlemoyer, L. 2022. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 11048–11064.

Muennighoff, N.; Tazi, N.; Magne, L.; and Reimers, N. 2023. MTEB: Massive Text Embedding Benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2006–2029.

Narang, S.; Raffel, C.; Lee, K.; Roberts, A.; Fiedel, N.; and Malkan, K. 2020. WT5?! Training Text-to-Text Models to Explain their Predictions. *CoRR*, abs/2004.14546.

Nema, P.; Khapra, M. M.; Laha, A.; and Ravindran, B. 2017. Diversity driven attention model for query-based abstractive summarization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 1063–1072.

Otterbacher, J.; Erkan, G.; and Radev, D. R. 2009. Biased LexRank: Passage retrieval using random walks with question-based priors. *Information Processing and Management*, 45(1): 42–54.

Ponce, A.; Chamness Long, S.; Andersen, E.; Gutierrez Patino, C.; Harman, M.; A Morales, J.; Piccone, T.; Rodriguez Cajamarca, N.; Stephan, A.; Gonzalez, K.; VanRiper, J.; Evangelides, A.; Martin, R.; Khosla, P.; Bock, L.; Campbell, E.; Gray, E.; Gryskiewicz, A.; Ibrahim, A.; Solis, L.; Hearn-Desautels, G.; and Tinucci, F. 2019. *Global Insights on Access to Justice 2019: Findings from the World Justice Project General Population Poll in 101 Countries*. World Justice Project.

Rabelo, J.; Goebel, R.; Kim, M.; Kano, Y.; Yoshioka, M.; and Satoh, K. 2022. Overview and Discussion of the Competition on Legal Information Extraction/Entailment (COLIEE) 2021. *Review of Socionetwork Strategies*, 16: 111–133.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Rajani, N. F.; McCann, B.; Xiong, C.; and Socher, R. 2019. Explain Yourself! Leveraging Language Models for Commonsense Reasoning. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, 4932–4942.

Ravichander, A.; Black, A. W.; Wilson, S.; Norton, T. B.; and Sadeh, N. M. 2019. Question Answering for Privacy Policies: Combining Computational and Legal Perspectives. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 4946–4957.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 3980–3990.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144.

Roberts, A.; Raffel, C.; and Shazeer, N. 2020. How Much Knowledge Can You Pack Into the Parameters of a Language Model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 5418–5426.

Robertson, S. E.; Walker, S.; Jones, S.; Hancock-Beaulieu, M.; and Gatford, M. 1994. Okapi at TREC-3. In *Proceedings of the 3rd Text REtrieval Conference*, 109–126.

Sharma, S.; Asri, L. E.; Schulz, H.; and Zumer, J. 2017. Relevance of Unsupervised Metrics in Task-Oriented Dialogue for Evaluating Natural Language Generation. *CoRR*, abs/1706.09799.

Shukla, A.; Bhattacharya, P.; Poddar, S.; Mukherjee, R.; Ghosh, K.; Goyal, P.; and Ghosh, S. 2022. Legal Case Document Summarization: Extractive and Abstractive Methods and their Evaluation. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, 1048–1064.

Su, J.; Lu, Y.; Pan, S.; Wen, B.; and Liu, Y. 2021. RoFormer: Enhanced Transformer with Rotary Position Embedding. *CoRR*, abs/2104.09864.

Tombros, A.; and Sanderson, M. 1998. Advantages of Query Biased Summaries in Information Retrieval. In *Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2–10.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023. LLaMA: Open and Efficient Foundation Language Models. *CoRR*, abs/2302.13971.

Trautmann, D.; Petrova, A.; and Schilder, F. 2022. Legal Prompt Engineering for Multilingual Legal Judgement Prediction. *CoRR*, abs/2212.02199.

Voorhees, E. M. 1999. The TREC-8 Question Answering Track Report. In *Proceedings of The 8th Text REtrieval Conference*, volume 500–246.

Wang, L.; Yang, N.; Huang, X.; Jiao, B.; Yang, L.; Jiang, D.; Majumder, R.; and Wei, F. 2022a. Text Embeddings by Weakly-Supervised Contrastive Pre-training. *CoRR*, abs/2212.03533.

Wang, S.; Xu, F.; Thompson, L.; Choi, E.; and Iyyer, M. 2022b. Modeling Exemplification in Long-form Question Answering via Retrieval. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics*, 2079–2092.

Wang, Y.; Ivison, H.; Dasigi, P.; Hessel, J.; Khot, T.; Chandu, K. R.; Wadden, D.; MacMillan, K.; Smith, N. A.; Beltagy, I.; and Hajishirzi, H. 2023. How Far Can Camels Go? Exploring the State of Instruction Tuning on Open Resources. *CoRR*, abs/2306.04751.

Wiegreffe, S.; Marasovic, A.; and Smith, N. A. 2021. Measuring Association Between Labels and Free-Text Rationales. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 10266–10284.

Xiong, L.; Xiong, C.; Li, Y.; Tang, K.; Liu, J.; Bennett, P. N.; Ahmed, J.; and Overwijk, A. 2021. Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval. In *Proceedings of the 9th International Conference on Learning Representations*.

Xu, C.; Sun, Q.; Zheng, K.; Geng, X.; Zhao, P.; Feng, J.; Tao, C.; and Jiang, D. 2023a. WizardLM: Empowering Large Language Models to Follow Complex Instructions. *CoRR*, abs/2304.12244.

Xu, F.; Song, Y.; Iyyer, M.; and Choi, E. 2023b. A Critical Evaluation of Evaluations for Long-form Question Answering. In *Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics*, 3225–3245.

Yang, Y.; Cer, D.; Ahmad, A.; Guo, M.; Law, J.; Constant, N.; Ábrego, G. H.; Yuan, S.; Tar, C.; Sung, Y.; Strope, B.; and Kurzweil, R. 2020. Multilingual Universal Sentence Encoder for Semantic Retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 87–94.

Yu, M.; Yin, W.; Hasan, K. S.; dos Santos, C. N.; Xiang, B.; and Zhou, B. 2017. Improved Neural Relation Detection for Knowledge Base Question Answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 571–581.

Zheng, L.; Chiang, W.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E. P.; Zhang, H.; Gonzalez, J. E.; and Stoica, I. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *CoRR*, abs/2306.05685.

Zhong, H.; Xiao, C.; Tu, C.; Zhang, T.; Liu, Z.; and Sun, M. 2020. JEC-QA: A Legal-Domain Question Answering Dataset. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, 9701–9708.