ArabLegalEval: A Multitask Benchmark for Assessing Arabic Legal Knowledge in Large Language Models

Faris Hijazi[1], **Somayah AlHarbi**[1], **Abdulaziz AlHussein**[1], **Harethah Abu Shairah**[2],
Reem AlZahrani[2], **Hebah AlShamlan**[1], **Omar Knio**[2], **George Turkiyyah**[2]
[1]THIQAH, [2]KAUST

Abstract

The rapid advancements in Large Language Models (LLMs) have led to significant improvements in various natural language processing tasks. However, the evaluation of LLMs' legal knowledge, particularly in non-English languages such as Arabic, remains under-explored. To address this gap, we introduce `ArabLegalEval`, a multitask benchmark dataset for assessing the Arabic legal knowledge of LLMs. Inspired by the `MMLU` and `LegalBench` datasets, `ArabLegalEval` consists of multiple tasks sourced from Saudi legal documents and synthesized questions. In this work, we aim to analyze the capabilities required to solve legal problems in Arabic and benchmark the performance of state-of-the-art LLMs. We explore the impact of in-context learning and investigate various evaluation methods. Additionally, we explore workflows for generating questions with automatic validation to enhance the dataset's quality. We benchmark multilingual and Arabic-centric LLMs, such as `GPT-4` and `Jais`, respectively. We also share our methodology for creating the dataset and validation, which can be generalized to other domains. We hope to accelerate AI research in the Arabic Legal domain by releasing the ArabLegalEval dataset and code: https://github.com/Thiqah/ArabLegalEval

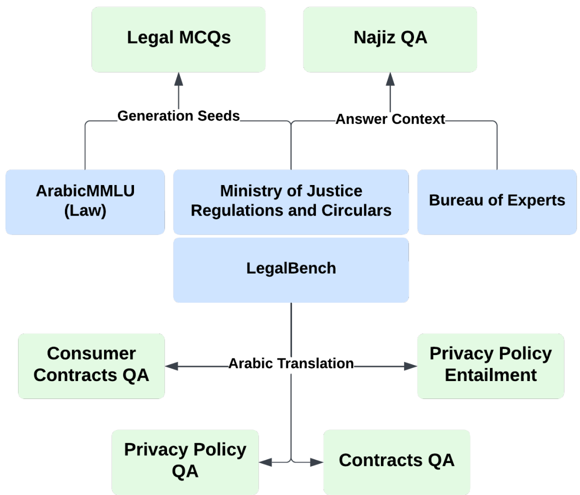

Figure 1: Tasks included in ArabLegalEval and their source documents.

1 Introduction

The development of LLMs has revolutionized various fields by enhancing natural language understanding and generation capabilities. However, the applicability and performance of these models in specialized domains, such as legal contexts, specially in low- and medium-resource languages such as Arabic, remain active research areas. In this paper, we report on ongoing work for evaluating the proficiency of large language models in understanding and processing Arabic legal texts. Given the complexity and richness of legal language, especially in Arabic, it is crucial to develop benchmarks that accurately assess the models' capabilities in this domain in order to guide model development. One of the key motivations for this work and benchmark is to find out the current state of LLMs in the Arabic Legal domain. Thus, we benchmark a wide range of LLMs from proprietary multilingual LLMs, such as `GPT-4` (OpenAI, 2024a), to open-source Arabic-centric LLMs, such as `Jais` (Sengupta et al., 2023).

There are perhaps two broad categories of evaluation criteria that are useful for assessing the performance of legal LLMs. The first category is the ability of a model to use specific regulations, facts, and data that are relevant to a particular conversation. This can be achieved by having the LLM memorize specific facts, presumably by finetuning it on particular legal corpora, or more conveniently by using a Retrieval-Augmented Generation (RAG) system to retrieve information relevant to the context at hand. The second broad category of assessment criteria is related to the model's ability to exhibit logical reasoning, understand relationships between entities and events, and apply these skills to answer

Proceedings of The Second Arabic Natural Language Processing Conference, pages 225–249
August 16, 2024

questions.

In this initial release of ArabLegalEval, we include tasks to assess the legal reasoning capabilities in Arabic as well as tasks to measure the ability to recall and use legal knowledge embedded in the finetuned models. Much like ArabicMMLU (Koto et al., 2024) was designed to test the general reasoning capabilities of Arabic LLMs but includes a mix of tasks with some requiring previous knowledge, the ArabLegalEval benchmark seeks to develop Arabic legal tasks and questions derived from original Arabic legal sources, in consultation with legal professionals, to test the legal reasoning capabilities of Arabic LLMs. The benchmark also includes some high quality translations, verified by legal experts, of tasks from English legal benchmarks (LegalBench) (Guha et al., 2023).

The benchmark includes Arabic legal Multiple Choice Questions (MCQs), Question & Answer (QA) pairs where the relevant Saudi regulations are included in the question, as well as Arabic translations of tasks from LegalBench. The benchmark does not assess the ability of models to retrieve specific facts and laws relevant to a particular context from external knowledge bases. While important for the successful deployment of a legal LLM, these retrieval abilities will be assessed in future tasks.

The primary contributions of this paper are: the development of a novel methodology for generating a legal QA dataset that can be adapted to other domains, such as finance; and ArabLegalEval, an initial dataset specifically designed for assessing Arabic legal knowledge in large language models.

2 Related work

2.1 Arabic and multilingual reasoning capabilities in LLMs

The success of LLMs, such as GPT-4 and GPT-4o (OpenAI, 2024a,b), Claude-3 (Anthropic, 2024), Command R and Command R Plus (Cohere, 2024a,b), and Llama3 (LLaMA, 2024), in exhibiting general purpose reasoning abilities - when queried in English - have naturally led to the development of models trained on Arabic content as well as benchmarks to evaluate the quality of reasoning in Arabic. Jais (Sengupta et al., 2023), for example, trained on native Arabic and quality translations, has shown better performance than other open-weight models on an Arabic version of the MMLU multiple-choice school exam questions benchmark (Koto et al., 2024). Arabic content and questions are also part of many broadly used multilingual benchmarks for assessing general reasoning. For example, MLQA (Lewis et al., 2020) assesses reading comprehension and question answering capabilities in languages including Arabic.

While substantial experimentation is ongoing for evaluating reasoning and knowledge-based tasks in multilingual models, there are some model scale effects that appear to be emerging. On one hand, smaller models and models trained with limited data do not generally perform very well. Even GPT-3.5 (Ouyang et al., 2022) performs worse when responding to queries in Arabic and Non-English languages as compared to English on various tasks (Koto et al., 2024; Jin et al., 2023). On the other hand, the performance generally improves steadily with increasing model scale (Shi et al., 2023). GPT-4, for example, exhibits remarkable performance on all sections of ArabicMMLU surpassing all other models including Arabic-centric ones, even when few-shot prompts are used. It appears that multilingual reasoning is an emergent ability of large language models. Beyond a certain (task-dependent) scale, LLMs have evidently strong multilingual reasoning abilities.

2.2 LLMs and evaluation benchmarks in legal domains

Given the importance of natural language to law, the advancing capabilities of large language models have been very quickly recognized and used for performing legals tasks. Current efforts are ongoing to explore whether current LLMs can be used as legal assistants for producing background research, drafting initial documents, summarizing contracts, answering questions about reports, and handling related tasks (Nay et al., 2023; Perlman, 2022; Goth, 2023). The announcement that GPT-4 has "passed the bar exam" (Katz et al., 2024) has shown potential that intelligent legal advisors are not too far off. In the area of tax law for example, LLMs, particularly when combined with prompting enhancements and the correct legal texts, can perform at high levels of accuracy (Nay et al., 2023). Systems for Arabic court rulings and QA related to Legal Palestinian cooperative maters were presented in (Ammar et al., 2024; Maree et al., 2024).

However, closer inspection has revealed that certification exams or narrow domains are not always representative of the actual use-cases in law practice or for LLMs (Fei et al., 2023). As a result, a number of efforts have been ongoing to develop models, agents, and benchmarks that use tasks and general conceptual frameworks similar to those used in the legal profession. `LegalBench` and a Chinese counterpart, `LawBench` (Fei et al., 2023), have sought to collect and categorize a broad set of tasks that appear in law practice, from simple tasks requiring rule-recall or issue recognition to sophisticated tasks that require interpretation, drawing conclusions, and multi-hop reasoning. As many of the LegalBench tasks, particularly those requiring interpretation, appear to involve reasoning skills that are not particular to US law, it seems reasonable to use translated versions of these tasks for testing LLMs in different languages.

3 Data Sources

3.1 Raw Arabic data sources

Obtaining comprehensive Arabic legal data is challenging. After consulting Saudi legal experts, we identified key sources that we have started to incorporate (Figure 1). These include scanned documents from the Ministry of Justice (MoJ, وزارة العدل) and the Board of Experts (BoE, هيئة الخبراء بمجلس الوزراء). MoJ documents cover 67 regulations (5720 subjects) and 388 circulars, while BoE documents include 448 rules (14134 subjects). See Appendix B for samples.

Both sets of documents contain regulations and statutes, they differ primarily in the topics they cover. MoJ documents specialize in topics issued by the Ministry, providing a comprehensive database of judicial regulations and legislation essential for legal research and practice.

Note that all data sources used are open and publicly available and scraped from official websites with no confidential information. These were the raw data sources used to generated the benchmark

3.1.1 Preparation Steps

For both BoE and MoJ documents, we followed a systematic data preparation process to ensure that the data is rich and easy to work with. The data was scraped from the web to capture all the regulations while preserving all metadata to allow for careful filtering later.

An example of a processed, structured, MoJ data document is shown in Figure 10.

3.1.2 Frequently asked Questions (FAQs)

In addition to these raw sources, we also rely on human-written FAQs that are publicly available. The questions and answers in this data are generally available in the BoE and MoJ documents, and we use them to build an open-ended question answering task in the benchmark, which we call *NajizQA*. A sample of this data can be found in Figure 11.

3.2 LegalBench

In addition to native Arabic sources, we rely on the translation of English legal documents. `Legalbench` is a benchmark for legal reasoning in English LLMs (Guha et al., 2023). We selected four datasets from it and translated them from English to Arabic. These were specifically chosen because they have fewer localization requirements and are more universal. This makes them ideal for assessing the ability of LLMs to understand and interpret legal clauses and contracts in Arabic.

Consumer Contracts QA: this dataset consists of 400 yes/no questions about the rights and obligations outlined in online terms of service agreements.

Contracts QA: this dataset consists contract clauses and questions about these contracts. It has 88 examples, with 80 examples for testing and 8 examples for training.

Privacy Policy QA: the dataset consists of questions and corresponding clauses from a privacy policy. It consists of a total of 10,931 instances, with 8 examples for training and 10,923 for testing.

Privacy Policy Entailment: this dataset has 4385 examples in training and testing, each example is a privacy policy clause and a description. The goal is to determine if the description for the clause is correct or not.

3.3 ArabicMMLU

`ArabicMMLU` (Koto et al., 2024), an Arabic knowledge evaluation benchmark constructed from human-written school exams from Arabic-speaking countries, served as one of the inspirations for this work. With a subset of its samples focused on the legal domain, `ArabicMMLU` provided a valuable starting point for us to generate our MCQs.

4 Benckmark Tasks

In this section, we describe the three broad task categories in the benchmark, including 10,000+ MCQs from the native Arabic MoJ and BoE documents, a set of QA from these documents, and a quality translation of a subset of the English

LegalBench benchmark related to consumer contracts and privacy policies. See Figure 1. We believe that the mixture of questions from native Arabic documents and translated questions gives us a somewhat diverse set of tasks and allows the benchmark to test a broader set of capabilities. In addition, this allows us to test the observation that, with increasing model scale, multilingual LLMs can display reasoning abilities and semantic judgment in Arabic as well as in they do in English (Shi et al., 2023). The quantitative details and human performance baseline are presented in Appendix J.

4.1 MCQs

One standard method of benchmarking reasoning and memorization capabilities in neural networks are MCQs, such as MMLU (Hendrycks et al., 2021). It is easy to verify the correctness of the answer using exact matching or regular expressions making MCQs ideal for automatic evaluation.

Given the availability of a large collection of raw legal documents from the MoJ and BoE, we aim to generate synthetic MCQs from them, using them the documents as context. Generating MCQs poses two main challenges: formulating questions and generating options (correct answers and plausible distractors).

We approached this using a robust LLM and experimented with three methods: 1. *QA to MCQ*, 2. *Chain of Thought (CoT)*, and 3. *Retrieval-based in-context learning*. See Figure 2.

In all cases, we prompt the model to synthesize questions in the same format and style as MMLU.

4.1.1 QA to MCQ

Here we use a two-step prompt. Given a legal document, the model is prompted to generate both a question and its answer, then a follow-up prompt to convert the question into a Multiple Choice Question (MCQ) by rephrasing the answer and generating appropriate distractors,

$$f(c_i) \rightarrow q_i, a_i; \quad g(q_i, a_i) \rightarrow D_i$$

where f is a language model that is given some legal context c_i, and then prompted to generate a QA pair (q_i, a_i), and g is another instance of the model that is given the QA pair and is tasked to generate a set of distractors D_i for the question.

4.1.2 QA to MCQ with CoT

CoT is a relatively recent method of prompting LLMs where instead of directly producing the answer, the model is given space to reason and "think out loud" before answering, essentially providing itself with more context. This simple technique has led to relatively huge gains in performance, in a variety of tasks (Wei et al., 2023). We want to utilize this idea to generate reasoning-based plausible distractors. This approach can be formulated as:

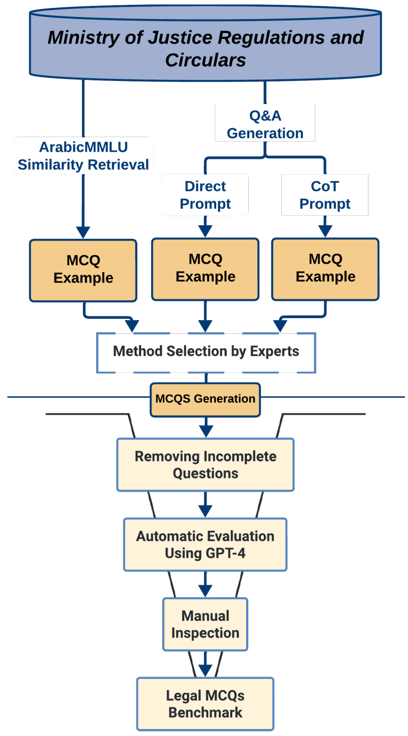

Figure 2: MCQs Generation and Filtering

$$f(c_i) \rightarrow q_i a_i; \quad g_{\text{CoT}}(q_i, a_i) \rightarrow D_i$$

where f is a language model prompted with some context c from the scraped MoJ documents to produce a question q and its answer a, these are then fed to another instance of the model g with CoT prompt to produce a set of plausible distractors D for the question.

4.1.3 Direct MCQs generation with in-context examples

Since the inspirations of this work are popular knowledge evaluation benchmarks such as MMLU and ArabicMMLU, our goal is to generate questions in a similar format. We begin by taking a subset of ArabicMMLU questions that have 'Law' as their subject tag (about 300 examples). For each document in the MoJ dataset, we perform a semantic similarity search (Risch et al., 2021) to retrieve the top k examples from these questions. These examples are then added to the prompt to guide the model in generating questions of a similar style.

$$f(c_i, E_k) \rightarrow q_i, a_i, D_i$$

The MoJ context is augmented with a set of `ArabicMMLU` examples E_k that are fed to a model f to generate a question q, where k is the number of retrieved examples. These examples provide context for the model for in-context learning so that the generated answer has the same style and format as the `ArabicMMLU` examples. From this, the model generated a question a and all of the distractors D in one go.

4.1.4 MCQ filtering and curation

Each of the above techniques was tested and then had the results manually reviewed and inspected by the legal experts according to the metrics in Appendix 3.1

It was concluded that the best method for generating MCQs was *in-context examples* with $k = 3$. Based on this, we decided to use this technique to generate all the MCQs.

After generating the MCQs dataset (approximately 12k samples), we did some automatic filtering using `GPT-4` (Chiang and yi Lee, 2023), where it was prompted to check if each sample satisfied our criteria (see appendix subsection 3.1). Any sample that failed to satisfy all of the above were removed, which left us with 10,583 MCQs for our benchmark. At the end, we extracted a random subset of the dataset for a final manual inspection.

4.1.5 Models for Generation

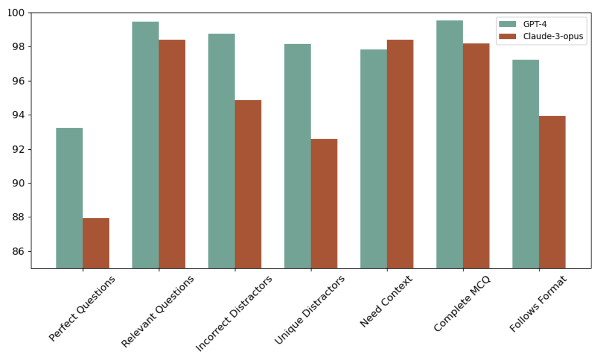

Figure 3: GPT-4 vs Claude-3-opus MCQs Generation

It has been observed that models tend to perform better on synthetic data generated by themselves as apposed to another model (Huang et al., 2024). To mitigate this unfair advantage, we split our documents and alternate between two state of the art models: `Claude-3-opus` and `GPT-4`. We make sure a model isn't evaluated on questions generated by itself.

4.2 QA

This dataset includes filtered legal QAs from a publicly available human written set of FAQs. The questions include only those referencing specific statues and regulations and articles in the MoJ and BoE Arabic documents, and are therefore particularly valuable for evaluating Arabic legal LLMs. See figure 4. A sample of the data shown in the Appendix B.

Embedding techniques were used in the semantic similarity matching phase (Figure 4). Specifically, we utilized `text-embedding-3-small` (OpenAI, 2024c) to generate embeddings for both the questions and contextual information. Subsequently, `cosine similarity` was employed to identify relevant texts corresponding to each question. This methodological framework was selected to evaluate the models' reasoning capabilities with and without context.

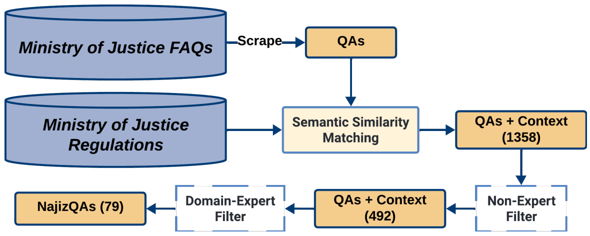

Figure 4: NajizQA curation pipeline

4.3 Arabic translation of LegalBench

4.3.1 Translation Strategy

We evaluated three different machine translation models, Azure Translation Services (Microsoft, n.d.), Google API Translation (Google, n.d.), and Opus MT (Tiedemann and Thottingal, 2020) along with `GPT-4` to determine the best translator for a publicly available dataset with legal context. The dataset used in these experiments is the **United Nations Parallel Corpus** (Ziemski et al., 2016), which consists of UN documents in the six official UN languages. We focused solely on the English and Arabic datasets, which comprise 20 million rows. However, to expedite the experimentation process, we utilized a subset of 14,000 examples from this dataset. This subset was used to evaluate the performance of the selected models without any preprocessing. Rouge metrics were used in evaluating the translation quality.

By comparing the results obtained from the different models, we aimed to identify the one that consistently produced the best translation from English to Arabic in the legal domain. The results of our experiments on this subset of data, without any prepossessing, are shown in Table 3 in Appendix D. Overall, the Opus model outperformed the other models across all the metrics. The Opus model achieved a ROUGE-1 score of 0.52, a ROUGE-2 score of 0.3, and a ROUGE-L score of 0.51. In addition to this, we had a sample of the translations manually reviewed by legal experts.

These findings established a basis for selecting Opus as the translator for our main task: translating selected datasets from `LegalBench` from English to Arabic.

4.3.2 Evaluation of Translated Output

To ensure the highest translation quality, we employed the three models from our initial experiments to translate the "Consumer Contract QA" dataset. We conducted a manual inspection of the overall results and, to further enhance the quality assessment, engaged both `GPT-4` and legal experts as evaluators for the translated content.

GPT-4 Evaluation We tasked `GPT-4` with evaluating a sample of the translated text alongside the original text, asking it to rate the translation quality on a scale from 1 to 5. Interestingly, `GPT-4` consistently rated the translations as either 4 or 5 out of 5, indicating a high level of perceived quality, see Appendix D for `GPT-4` evaluation prompt and result example.

Human Evaluation We selected five examples from the translated texts, each consisting of a contract text and a corresponding question. To evaluate the performance of the three translation models — Azure Translation Services, Google API Services, and Opus MT — we asked legal experts to assess each contract and question separately. The experts provided scores for each model's based on the translation quality and how well the legal context was preserved in the translation. The overall scores for all three models were relatively close, but Opus MT consistently achieved the highest score among them:

- Google API: average score of 3.3 out of 5
- Azure: average score of 3.6 out of 5
- Opus MT: Average score of 4 out of 5

In addition to the overall scores, we asked the experts to rank the best model for each example. Opus MT was chosen as the best model 60% of the time, while Azure Translation Services was selected 40% of the time. Interestingly, Google API Translation was never selected as the best model for any of the examples.

These results suggest that while all three models performed reasonably well, Opus MT demonstrated superior translation quality for the given legal texts and questions, as determined by the expert evaluations. See Figure 20 in Appendix D for a translated example.

5 Evaluation and results

5.1 MCQ Evaluation

In this section, we evaluate the performance of language models on our synthetic MCQs dataset using tailored prompts, where the instructions in the prompt are provided in English for each model.

5.1.1 Experiment Setup

We aim to improve the models' capabilities by modifying the given prompt. Different parts of the prompt can be optimized according to a given metric, and in this evaluation, we started with optimizing the instruction and few-shot examples to determine which method is more effective. Unfortunately, instruction optimization yielded no significant performance gain. On the other hand, few-shot optimization boosted the performance of many models. Hence, we decided to focus on few-shot optimization. `ArabicMMLU` is a benchmark to assess the capabilities of models, similar to `MMLU` benchmark, but with localized data in Arabic. A subset relevant to the legal domain of `ArabicMMLU` was sampled resulting in a total size of 524 questions after filtering questions that require context. Out of those 524 samples, 314 and 210 are law and political science, respectively. We optimize the prompts on this subset to use it for evaluating our generated MCQs.

DSPy (Khattab et al., 2024) is a Language Model (LM) programming framework to optimize LM prompts and weights automatically by recompiling the entire pipeline to optimize it on a specific task. We relied on this framework for prompt optimization. Initially, all of the models were given a zero-shot prompt with an English answer instruction and the input-output format. Then, this zero-shot prompt is optimized for each model to achieve a higher performance by augmenting it with either plain few-shot examples or few-shot with reasoning demonstrations using CoT. Teacher and student models were used to create few-shot examples with

CoT demonstrations, where the teacher is either a clone of the student or another model. Figure 5 demonstrates DSPy's bootstrapped few-shot optimization.

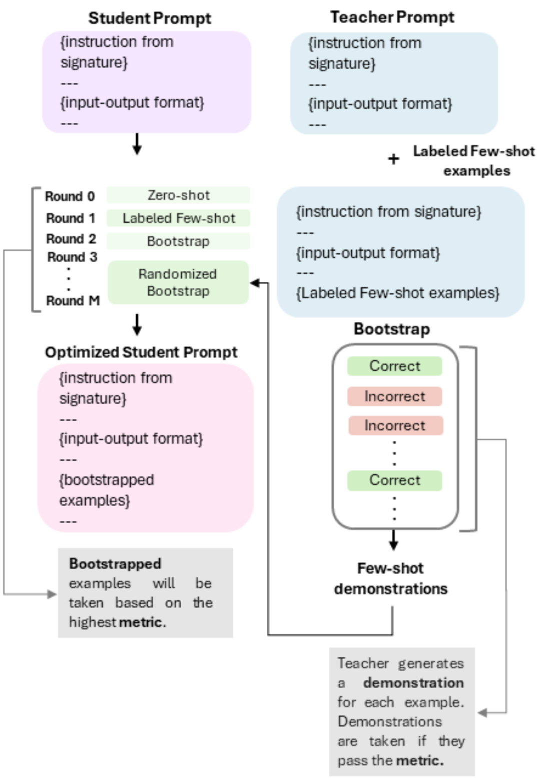

Figure 5: DSPy's prompt optimizer process.

5.1.2 MCQ Results Analysis

The MCQ dataset contains 10,583 questions, with 5,544 generated by GPT-4 and 5,039 by Claude-3-opus. We evaluated several LMs on this dataset, including Command R, Command R Plus, and Llama3 (8B and 70B). However, GPT-4o was tested only on Claude-3-opus subset to mitigate potential bias towards its own generated questions (Panickssery et al., 2024). Our evaluation metric assesses LM performance by comparing the selected answer with the correct one for each question. For further ArabicMMLU detailed results with optimized prompts, refer to appendix G.

Table 1 summarizes the performance of LMs on our generated MCQs using prompts optimized with DSPy on ArabicMMLU's legal subsets. Interestingly, many of the optimized few-shot prompts shared identical examples, suggesting that certain examples play a more significant role in improving LMs' performance than others. In addition, few-shot examples coupled with CoT reasoning boosted the capabilities of the models. For further testing, we employed GPT-4 as a teacher model for smaller LMs in both plain and CoT few-shot prompting. Surprisingly, these smaller LMs demonstrated greater performance in few-shot CoT when the teacher was a clone of themselves, rather than the more advanced GPT-4. This unexpected result suggests that LMs may have a better grasp of their own reasoning.

We observed that the choice of language plays a crucial role in the reasoning abilities of smaller models. In many cases, these LMs generated answers without providing accompanying reasoning. Figure 6 shows the differences in reasoning language for Command R Plus, revealing a degradation of the model's reasoning capability when the language choice is Arabic.

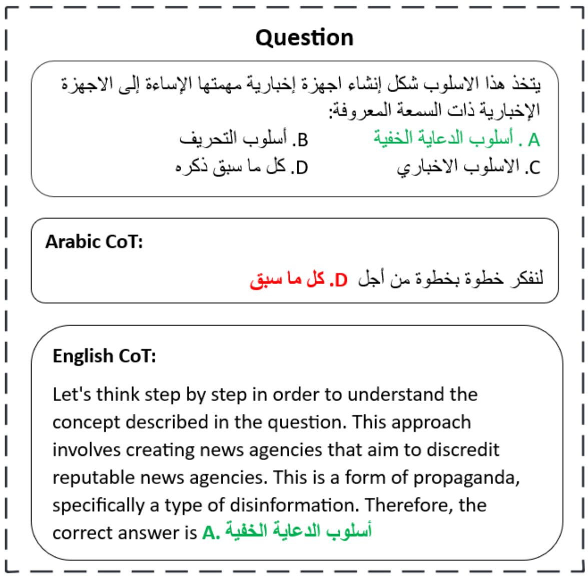

Figure 6: Command R Plus CoT Reasoning. The English translation is *'This style takes the form of creating news networks whose purpose is to harm reputable news networks: A. The style of hidden propaganda B. The style of manipulation C. The journalistic style D. All of the above'*. The correct answer, A, is highlighted in green.

Table 1 shows that GPT-4o demonstrated superior performance across all prompting methods, achieving 79.10% accuracy with few-shot prompting. Llama3 (8B) achieved a 5% increase using few-shot prompting with CoT reasoning. Similarly, the other LMs, except Command R, obtained better results when the prompt included their own CoT reasoning. Our findings reveal potential for improving LMs' question-answering capabilities through task-specific optimized prompts.

Table 1: Experimental results on Generated MCQs

	Original	Few-shot	Few-shot (GPT-4 Teacher)	CoT Few-shot	CoT Few-shot (GPT-4 Teacher)
GPT-4o	76.80%	**79.10%**	-	76.50%	-
Command R	68.10%	71.40%	72.30%	71.00%	70.09%
Command R+	68.10%	71.40%	72.30%	73.10%	69.86%
Llama3 (8B)	68.20%	71.20%	72.30%	73.60%	70.45%
Llama3 (70B)	68.17%	71.47%	72.34%	71.93%	70.53%

5.2 Open ended QA Evaluation

5.2.1 Experiment Setup

In open-ended QA—contrary to closed-set QA where the possible answers are fixed—the answer can be expressed in natural language, which is harder to evaluate since it relies on semantics and meaning. We observe that English prompts for the exam-taker LLM and the judge-LLM outperform Arabic prompts, so we use those English prompts for the instructions.

Traditional evaluation of QA models uses metrics like Exact Match (EM), F1-score, and top-n-accuracy, focusing on lexical matching. These metrics often miss semantically similar answers that use different words.

We use use an *LLM-as-a-judge* (Huang et al., 2024), (Kim et al., 2023), (Zheng et al., 2023) to rate the answer similarity on a scale from 0 to 5 given the generated answer and the reference ground truth answer. In our case, `GPT-4` is the judge, and we refer to the output score as the *answer similarity metric*, see figure 23 for details. We notice that the judge LLMs give a lower score to answers that are in a different language than that of the reference answer even when the content is correct. To mitigate this, we prompt the models to output answers in the same language as the reference answer (which is Arabic in our dataset).

We run the evaluation for each LLM with two cases: one where it is given the question and the needed context, and one where it is given only the question. We also add a "golden model" that always answers the perfect ground truth answer, just so that we can compare against the upper bound of what the judge-LLM is going to score.

5.2.2 QA Results

We run our experiments on the 79 NajizQA pairs that have been filtered and verified by legal experts.

From the original 1358 QA pairs, we chose the filtered 79 QA + context sets. We run them with and without context and show result in Figure 7.

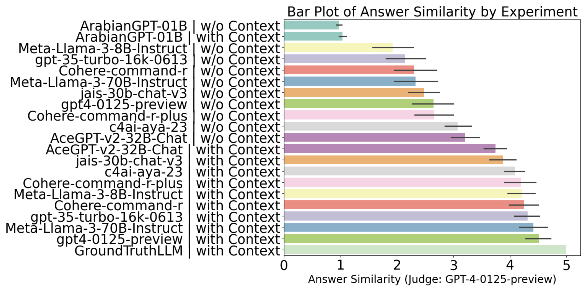

Figure 7: Filtered NajizQA (79) QAs with scores based on answer similarity by GPT-4 as judge

We notice that all models perform significantly better when the context is provided, and on closer inspection (Figure 11), we can see that the answer can easily be extracted from the context, making the context a giveaway of the answer which only measures the model's ability to recall and extract information and not legal reasoning.

5.3 Arabic LegalBench Evaluation

5.3.1 Experiments Setup

In total, we carried out 96 unique experiments with different prompting techniques to assess the performance of the models in legal reasoning. All prompts were in English, as this yielded better model performance. Appendix I provides prompt examples for each technique used.

For the *Contract QA* and *Consumer Contract QA* datasets, we utilized the entire testing data in our experiments. However, for the larger datasets, we selected a representative sample. For each technique and dataset, we created tailored prompts. The training examples in one-shot and few-shot learning were fixed across all models for each specific dataset. We then benchmarked the performance of all models using the following four approaches: zero-shot learning with simple prompts where models are asked straightforward questions without extensive instructions; zero-shot learning with detailed instructions in the prompt; one-shot learning; and few shot learning.

5.3.2 Results Analysis

A comprehensive overview of model performance across all tasks and learning techniques is presented in Table 6 in Appendix H. This table provides a summary of F1 scores for each model and task combination, offering a comparative analysis of the various approaches evaluated in this study.

The **consumer contract QA** task can assess the models' ability to answer questions based on a long context, in this case consumer contracts. `GPT-4` with one-shot learning, and `Llama3 (70B)`, with a zero-shot basic prompt, achieved the highest F1 score of 90%. This suggests that both models can extract relevant information from consumer contracts and provide answers, even when training examples are not available or limited. We observed that most of the models performed well on the Contract QA task, which assessed the models capabilities to answer questions related to contracts. `Command R Plus`, using few-shot learning, achieved the highest F1 score of 99%. This high score indicates that the model can accurately understand and respond to questions about contracts when provided with a small number of training examples. However, this task proved to be the least challenging among the four tasks. On the other hand,the privacy policy entailment task proved to be the most challenging for the LLMs across all techniques, highlighting the complexity of this task. `Command R Plus`, using one-shot learning, achieved the best F1 score of 66%. This result suggests that while the models struggle with this task, `Command R Plus` is more capable of understanding of privacy policies when given a single training example.

In the final task, **Privacy Policy QA**, which evaluated the models' ability to answer questions based on privacy policies, `GPT-4` with one-shot learning achieved the highest performance. This result demonstrates `GPT-4`'s strong capability in extracting relevant information from privacy policies and providing accurate answers when given a single training example.

Overall, one-shot learning achieved the best results for most of the models across the various tasks. This finding suggests that providing a single example can significantly improve the models' performance in understanding and responding to questions related to legal documents such as consumer contracts and privacy policies.

6 Limitations

The `ArabLegalEval` benchmark currently relies heavily on Saudi Arabian legal documents, with some tasks translated from universal benchmarks. Including documents from more Arabic-speaking countries would improve geographic representation. Our study did not evaluate all models, which limits generalizability; future work should include a broader range of models. Limited access to legal experts affected validation depth; involving more experts would improve quality control. The dataset lacks granular categorization, such as task-specific prior knowledge, document origin, and AI-generated content labels. Adding more granular metadata and task categories would aid nuanced model training and evaluation.

7 Conclusions and Future Work

We are developing an Arabic benchmark to evaluate LLMs' legal reasoning, using Saudi regulations and translated LegalBench problems. Future plans include adding more KSA regulatory documents, court cases, and judicial decisions to enhance this benchmark and promote advancements in Arabic legal AI.

Acknowledgments

We would like to express our sincere gratitude to Yusra Alonaizan for her contributions in envisioning this project, and dedicated team of six members—Lama Alsaif, Alanoud Alangari, Alaa Bazaid, and Faisal Alessa for gathering and curating the data sources. Additionally, we extend our thanks to the legal team at THIQAH for their valuable time in reviewing the labeled data, we are grateful to Norah AlHussein, an external legal expert, for her meticulous double-review of the labeled data and the translated LegalBench dataset, as well as for her clarifications and support in explaining the legal data.

References

Ahmed Abdelali, Hamdy Mubarak, Shammur Absar Chowdhury, Maram Hasanain, Basel Mousi, Sabri Boughorbel, Samir Abdaljalil, Yassine El Kheir, Daniel Izham, Fahim Dalvi, Majd Hawasly, Nizi Nazar, Youssef Elshahawy, Ahmed Ali, Nadir Durrani, Natasa Milic-Frayling, and Firoj Alam. 2024. LAraBench: Benchmarking Arabic AI with Large Language Models. In *Proceedings of the 18th Conference of the European Chapter of the Association* *for Computational Linguistics: Volume 1, Long Papers*, Malta. Association for Computational Linguistics.

Emad A. Alghamdi, Jezia Zakraoui, and Fares A. Abanmy. 2023. Domain adaptation for arabic machine translation: The case of financial texts. *Preprint*, arXiv:2309.12863.

Reem Alghamdi, Zhenwen Liang, and Xiangliang Zhang. 2022. ArMATH: a dataset for solving Arabic math word problems. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 351–362, Marseille, France. European Language Resources Association.

Adel Ammar, Anis Koubaa, Bilel Benjdira, Omer Nacar, and Serry Sibaee. 2024. Prediction of arabic legal rulings using large language models. *Electronics*, 13(4):764.

Anthropic. 2024. Claude 3 haiku: Our fastest model yet. `https://www.anthropic.com/news/claude-3-haiku`. Accessed: 2024-06-26.

Cheng-Han Chiang and Hung yi Lee. 2023. Can large language models be an alternative to human evaluations? *Preprint*, arXiv:2305.01937.

Cohere. 2024a. Command r. `https://docs.cohere.com/docs/command-r`. Accessed: 2024-06-26.

Cohere. 2024b. Command r+. `https://docs.cohere.com/docs/command-r-plus`. Accessed: 2024-06-26.

Fahim Dalvi, Maram Hasanain, Sabri Boughorbel, Basel Mousi, Samir Abdaljalil, Nizi Nazar, Ahmed Abdelali, Shammur Absar Chowdhury, Hamdy Mubarak, Ahmed Ali, Majd Hawasly, Nadir Durrani, and Firoj Alam. 2024. LLMeBench: A flexible framework for accelerating llms benchmarking.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Songyang Zhang, Kai Chen, Zongwen Shen, and Jidong Ge. 2023. LawBench: Benchmarking Legal Knowledge of Large Language Models. *arXiv*.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2023. A framework for few-shot language model evaluation.

Google. n.d. Google cloud translation api.

Gregory Goth. 2023. Why are lawyers afraid of AI? *Commun. ACM, https://doi.org/10.1145/3631935*, 67(1):14–16.

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

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Hui Huang, Yingqi Qu, Jing Liu, Muyun Yang, and Tiejun Zhao. 2024. An empirical study of llm-as-a-judge for llm evaluation: Fine-tuned judge models are task-specific classifiers.

Yiqiao Jin, Mohit Chandra, Gaurav Verma, Yibo Hu, Munmun De Choudhury, and Srijan Kumar. 2023. Better to Ask in English: Cross-Lingual Evaluation of Large Language Models for Healthcare Queries. *arXiv*.

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(20230254).

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan A, Saiful Haq, Ashutosh Sharma, Thomas Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2024. Dspy: Compiling declarative language model calls into state-of-the-art pipelines. In *12th International Conference on Learning Representations, ICLR 2024, Messe Wien Exhibition and Congress Center, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2023. Prometheus: Inducing fine-grained evaluation capability in language models.

Noam Kolt. 2022. Predicting consumer contracts. *Berkeley Tech. LJ*, 37:71.

Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Boda Sadallah, Aisha Alraeesi, Khalid Almubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, Nizar Habash, Preslav Nakov, and Timothy Baldwin. 2024. ArabicMMLU: Assessing massive multitask language understanding in arabic. *Preprint*, arXiv:2402.12840.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics.

Meta LLaMA. 2024. Introducing meta llama 3: The most capable openly available llm to date. `https://ai.meta.com/blog/meta-llama-3/`. Accessed: 2024-06-26.

Mohammed Maree, Rabee Al-Qasem, and Banan Tantour. 2024. Transforming legal text interactions: leveraging natural language processing and large language models for legal support in palestinian cooperatives. *International Journal of Information Technology*, 16(1):551–558.

Microsoft. n.d. Azure translation services.

Subhabrata Mukherjee, Xiaodong Liu, Guoqing Zheng, Saghar Hosseini, Hao Cheng, Greg Yang, Christopher Meek, Ahmed Hassan Awadallah, and Jianfeng Gao. 2021. Few-shot learning evaluation in natural language understanding. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

John J. Nay, David Karamardian, Sarah B. Lawsky, Wenting Tao, Meghana Bhat, Raghav Jain, Aaron Travis Lee, Jonathan H. Choi, and Jungo Kasai. 2023. Large language models as tax attorneys: A case study in legal capabilities emergence. *arXiv, https://arxiv.org/abs/2306.07075*.

OpenAI. 2024a. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

OpenAI. 2024b. Hello gpt-4o. `https://openai.com/index/hello-gpt-4o/`. Accessed: 2024-06-26.

OpenAI. 2024c. New embedding models and api updates. `https://openai.com/index/new-embedding-models-and-api-updates/`. Accessed: 2024-07-02.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. Llm evaluators recognize and favor their own generations. *Preprint*, arXiv:2404.13076.

Andrew Perlman. 2022. The implications of chatgpt for legal services and society. *Suffolk University Law School Research Paper*, (22-14). Available at SSRN: http://dx.doi.org/10.2139/ssrn.4294197.

Julian Risch, Timo Möller, Julian Gutsch, and Malte Pietsch. 2021. Semantic answer similarity for evaluating question answering models. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 149–157, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, Lalit Pradhan, Zain Muhammad Mujahid, Massa Baali, Alham Fikri Aji, Zhengzhong Liu, Andy Hock, Andrew Feldman, Jonathan Lee, Andrew Jackson, Preslav Nakov, Timothy Baldwin, and Eric Xing. 2023. Jais and Jais-chat: Arabic-Centric Foundation and Instruction-Tuned Open Generative Large Language Models. *arXiv, https://arxiv.org/abs/2308.16149*.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language Models are Multilingual Chain-of-Thought Reasoners. *ICLR*.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

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

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534, Portorož, Slovenia. European Language Resources Association (ELRA).

A LMs descriptions

Table 2: Models used in evaluation

Model	Language	# parameters
GPT 4o	Multilingual	Undisclosed
GPT 4 (0125-preview)	Multilingual	Undisclosed
GPT 3.5 turbo 16k (0613)	Multilingual	Undisclosed
meta-llama/Meta-Llama-3-8B-Instruct	Multilingual	8B
meta-llama/Meta-Llama-3-70B-Instruct	Multilingual	70B
CohereForAI/c4ai-command-r-v01	Multilingual	35B
CohereForAI/c4ai-command-r-plus	Multilingual	104B
CohereForAI/aya-101	Multilingual	13B
Claude 3 opus (20240229)	Multilingual	137B

B Examples of Data Sources


```
JSON                                                    Copy  Caption  ...
{
    "Regulation_Name": "نظام الأحوال الشخصية",
    "Details": {
        "تاريخ الإصدار": "1443/08/06 هـ الموافق : 09/03/2022 مـ",
        "تاريخ النشر": "1443/08/15 هـ الموافق : 18/03/2022 مـ",
        "الحالة":"ساري",
        "أدوات إصدار النظام":"مرسوم ملكي رقم (م/73) وتاريخ 1443/8/6هـ قرار مجلس الوزراء رقم (429) وتاريخ 1443/8/5هـ"

    },
    "Subjects": {
        "المادة الأولى.": "الخطبة هي طلب الزواج والوعد به",
        "المادة الثانية.": "لكل من الخاطب والمخطوبة العدول عن الخطبة",
        "المادة الثالثة": "جميع ما يقدمه الخاطب أو المخطوبة إلى الآخر خلال فترة الخطبة يعد هدية؛ ما لم يصرح الخاطب بأن ما قدمه يعد مهراً أو يجر عرف على أنه من المهر.",
        "المادة الرابعة": "إذا عدل أي من الخاطب أو المخطوبة عن الخطبة بسبب يعود إليه، فليس له الرجوع في الهدية التي قدمها. وللطرف الآخر أن يسترد منه ما قدمه من هدية إن كانت قائمة وإلا بمثلها، أو قيمتها يوم قبضها، ما لم تكن الهدية مما يستهلك بطبيعتها.\nوفي جميع الأحوال، إذا انتهت الخطبة بالوفاة، أو بسبب لا يد لأحد الطرفين فيه، فلا يسترد شيء من الهدايا.",
        "المادة الخامسة": "إذا عدل أي من الخاطب أو المخطوبة عن إبرام عقد الزواج أو مات قبل العقد، وكان الخاطب قد سلّم إلى مخطوبته قبل العقد مالاً على أنه من المهر، يحق للخاطب أو لورثته الرجوع فيما سلم بعينه إن كان قائماً وإلا بمثله، أو بقيمته يوم القبض.\nإذا كانت المخطوبة اشترت بالمهر أو بعضه لمصلحة الزواج -وفق ما جرى به العرف- وكان العدول من الخاطب بلا سبب من قبلها، أو كان العدول منها بسبب من الخاطب، فلها الخيار بين إعادة المهر أو تسليم ما اشترته بحاله.",
        "المادة السادسة": "الزواج عقد بأركان وشروط، يرتب حقوقاً وواجبات بين الزوجين، غايته الإحصان وإنشاء أسرة مستقرة يرعاها الزوجان بمودة ورحمة.",
        "المادة السابعة": "الخلوة -في سبيل تطبيق هذا النظام- هي انفراد الزوجين في مكان ليس عندهما من يميِّز.",
        }
...
}
```


Figure 9: Example data document from Board of experts (BOE) data source


```
{
    "Regulation_name": "نظام الإثبات",
    "Regulation_Content": "تضمن النظام: أحكام عامة عن الإثبات، الإقرار، استجواب الخصوم، المحررات الرسمية، المحررات العادية، إلزام الخصم بتقديم المحررات، إثبات صحة المحررات، تحقيق الخطوط، دعوى التزوير، الكتابة، الدليل الرقمي، الشهادة، القرائن وحجية الأمر المقضي، العرف، اليمين واليمين الحاسمة والمتممة، المعاينة، الخبرة، أحكام ختامية، النشر والنفاذ.",
    "Details": {
        "تاريخ الإصدار": "٢٦ جمادي الأول ١٤٤٣",
        "التصنيف": "القضاء",
        "نوع التشريع": "نظام",
        "تاريخ النشر": "٤ جمادي الآخر ١٤٤٣",
        "حالة التشريع": "ساري",
        "أداة اصدار التشريع": "المرسوم الملكي رقم م/43",
        "مرفق أداة الاصدار": "اداة الاصدار-47",
        "وثيقة التشريع": "نسخة التشريع-91",
    },
    "Subjects": {
        "الباب الأول أحكام عامة": {
            "status": "Original",
            "المادة الأولى": {
                "description": "تسري أحكام هذا النظام على المعاملات المدنية والتجارية.",
                    "status": "Original"
            },
            "المادة الثانية": {
                "description": "على المدعي أن يثبت ما يدعيه من حق، وللمدعى عليه نفيه.  يجب أن تكون الوقائع المراد إثباتها متعلقة بالدعوى، ومنتجة فيها، وجائز قبولها.  لا يجوز للقاضي أن يحكم بعلمه الشخصي.",
                    "status": "Original"
            },
            "المادة الثالثة": {
                "description": "البينة على من ادعى، واليمين على من أنكر.  البينة لإثبات خلاف الظاهر، واليمين لإبقاء الأصل.  البينة حجة متعدية، والإقرار حجة قاصرة.  الثابت بالبرهان كالثابت بالعيان.",
                    "status": "Original"
                    }
                    }
...
}
```


Figure 10: Example data document from Ministry of Justice (MOJ) data source

NajizQA Examples	
Question	ما هي الحالات التي لا يستحق فيها دفع تكاليف قضائية؟
Answer	إذا حكم بعدم الاختصاص فلا تستحق تكاليف جديدة لإقامة الدعوى أمام المحكمة المختصة ما لم يتغير موضوع الدعوى، إذا قضت المحكمة المختصة بنقض الحكم وإعادة الدعوى إلى المحكمة التي أصدرت العكم المعترض عليه إلى المحكمة أو إلى أي .محكمة أخرى
Context	لاتفرض تكاليف جديدة على الدعوى في أي من :الحالتين الأتيتين ذا حكم بعدم الاختصاص وأقيمت.١ دعوى أمام المحكمة المختصة، ٢.إذا قضت المحكمة المختصة بنقض الحكم وإعادة الدعوى إلى المحكمة التي أصدرت الحكم المعترض عليه أو إلى محكمة أخرى
Question	متى ترد التكاليف القضائية؟
Answer	ترد التكاليف القضائية متى تبين عدم وجوبها على دافعها أو استفادته من إحدى حالات ،الإعفاء ومن ذلك: إذا حكم لمصلحة دافع التكاليف القضائية. طلب رد القاضي أو القضاة وتم قبول طلب الرد، طلب تصحيح الحكم وتفسيره إذا قضي بإجابة الطلب، طلب الاستثناف إذا حكم بنقض الحكم المستأنف كليا فيعفى من قيمة الطلب، إذا نقض الحكم جزئيا فيعقى بقدر ذلك الجزء وطلب التقض إذا قضي فيه بإعادة القضية إلى المحكمة مصدرة الحكم المعترض عليه، إذا ترك المدعي دعواه قبل عقد الجلسة الأولى وقها للإجراءات النظامية. الدعاوى التي .تنتهي بالصلح قبل رفع الجلسة الأولى
Context	تُرد التكاليف القضائية المدفوعة متى تبين عدم وجوبها على دافعها أو استفادته من حالة من حالات الإعفاء ومن ذلك مايأتي: ١إذا حكم لمصلحة دافع التكاليف القضائية. ٢ طلب رد القاضي أو القضاة إذا قبل طلب .الرد طلب تصحيح الحكم أو تفسيره إذا.٣ قضي بإجابة الطلب، ٤ طلب الاستئناف إذا حكم بنقض الحكم المستأنف كلياً، وإذا نقض الحكم جزئياً فيعفى بقدر ذلك الجزء وطلب النقض إذا قضي فيه بإعادة القضية إلى المحكمة مصدرة الحكم المعترض عليه. ٥إذا ترك المدعي عليه دعواه قبل عقد الجلسة الأولى؛ وققاً للإجراءات النظامية الدعاوى التي تنتهي بالصلح قبل.٦ رفع الجلسة الأولى. ٧.الدعاوى المتعلقة بالحقوق الخاصة التي ترفع بالتبعية للقضايا الجزائية إذا انتهت بالصلح على أي حال كانت فيها الدعوى. وتحدد اللائحة الإجراعات والقواعد الخاصة .بذلك
Question	كم تبلغ رسوم إصدار ترخيص مكتب محاماة أجنبي؟
Answer	يبلغ رسم إصدار الترخيص لمكتب المحاماة الأجنبي (٢٠٠٠) ريال، وذلك بناء على ما ورد في المادة (٢\٨٤) من نظام المحاماة
Context	يبلغ رسم إصدار الترخيص لمكتبتكون مدة الترخيص لمزاولة مكتب -١ المعاماة الأجنبي مهنة المحاماة فيٍ المملكة (خمس) سنوات، قابلة للتجديد لمدة أو لمدد أخرى مماثلة وفقاً للشروط المحددة في المادة (الخامسة والأربعين) من هذا النظام ٢- يكون رسم إصدار الترخيص لمكتب المحاماة الأجنبي بمبلغ قدره .(أألفا) ريال، و(ألف) ريال عند تجديدها

Figure 11: Examples from the NajizQA data source

C MCQs Generation

3.1 MCQs Evaluation Criteria

```
You are a law Professor.
  You have a bank of multiple choice questions.
  Your task is to filter out the questions that are not
relevant to the context information provided.
  The questions are in Arabic.
  For the following context and question:
  Question: {question}

  Answer the following:
  - Is the question complete, meaning it has a context,
question, correct answer, and distractors? 0 for NO, 1 for
YES
  - Is the question relevant to the context information? 0
for NO, 1 for YES
  - Are the ditractors for the question all incorrect? 0 for
NO, 1 for YES
  - Is the correct answer the first option? 0 for NO, 1 for
YES
  - Does the question need the provided context to be
answered? 0 for NO, 1 for YES
  - Are all the distractors unique? 0 for NO, 1 for YES

  Provide the answers in the following format, Do not
output anything else:
  complete_mcq: 0 or 1
  question_relevance: 0 or 1
  distractors_correctness: 0 or 1
  correct_answer_first: 0 or 1
  context_needed: 0 or 1
  unique_distractors: 0 or 1
  total_score: between 0 and 6
```

Figure 12: MCQs Evaluation Prompt

3.2 MCQs Example

سياق النص: تجرى محاكمة الحدث أمام المحكمة بحضوره وولي أمره أو من يقوم مقامه، فإن تعذر ذلك فمندوب من الدار، وذلك دون الإخلال بحق الحدث في الاستعانة بمحام وفق الأحكام المقررة نظاماً. وللمحكمة -بناء على طلب مَنْ له مصلحة- أن تسمح بعدم حضور الحدث أو ولي أمره أو من يقوم مقامه للمحاكمة، ويكتفى بحضور من يمثله وتعد المحاكمة حضورية في حقه.

السؤال: في حال تعذر حضور الحدث وولي أمره أو من يقوم مقامه لمحاكمة الحدث، فإنه يحضر:

A. محامي الحدث
B. أحد أصدقاء الحدث
C. أحد أقارب الحدث
D. مندوب من الدار

الإجابة الصحيحة: مندوب من الدار

Figure 13: Example of an MCQ generated with ArabicMMLU context

3.3 Prompts

```
You are a law Professor.
  Your task is to setup {num_questions_per_chunk} questions for
an upcoming quiz.
  The questions should be diverse in nature across the document.
  Restrict the questions to the context information provided.
  A student should be able to answer your questions using only the
context information.
  Avoid asking about dates or specific numbers.
  Ask short questions only
  Quesitons should be solvable with short answers
  The questions should be in Arabic.

  Context information is below.
  ---------------------
  {context_str}
  ---------------------
  Given the context information and not prior knowledge, generate
the relevant questions.
```

Figure 14: Question generation prompt

```
Context information is below.
  ---------------------
  {context_str}
  ---------------------
  Given the context information and not prior knowledge, answer
the query in Arabic.
  The answer should be concise and to the point, make it as short
as possible.
  Avoid answering with lists or numbered points.
  Give short and direct answers only.
  Query: {query_str}
  Answer:
```

Figure 15: Question answering prompt

```
You are a law Professor.
  Your task is to setup {num_questions_per_chunk} question/s for
an upcoming quiz.
  The questions should be diverse in nature across the document.
  Do not ask literal questions from the context information, but
rather ask questions that require understanding of the context.
  The questions should be about logical reasoning, inference, or
any other form of higher-order thinking.
  The questions must be multiple choice questions.
  The distractors should be plausible and tricky.
  The correct answer should always be the first option.
  For each question, provide 3 distractors in addition to the correct
answer.
  The questions should be in Arabic.
  End each example question with ####.

  Your questions should have simmilar style and format to the
following examples:

  {few_shot_examples}

  Context information is below.
  ---------------------
  {context_str}
  ---------------------
  Given the context information and not prior knowledge, generate
relevant questions.
```

Figure 16: Prompt for generating MCQs using in-context examples

You are a law Professor.
You have a bank of questions that are in the form of question-answer pairs.
Your task is to convert the question-answer pairs to multiple choice questions.
The distractors should be plausible and tricky, and similar to the style of the answer.
Make sure all distractors are incorrect by comparing them agains the answer.
If the answer is too long, shotren it to a reasonable length, similar to the length of the distractors, by making the answer direct.
The questions are in Arabic.
For the following question and answer:
Question: {question}
Answer: {answer}
Write a multiple choice question with the correct answer and three distractors in the following format:

<start_thought>
"A space for you to think step by step about 3 incorrect distractors, do it in Arabic"
<end_thought>

"Repeat the quesiton"

1-"Correct Answer"
2-"Distractor 1"
3-"Distractor 2"
4-"Distractor 3"

Figure 17: Converting QAs to MCQs using CoT

D Translation

4.1 ROUGE scores for UN English-Arabic translation

Model	Rouge-1	Rouge-2	Rouge-L
Azure Translation Services	0.446	0.242	0.435
Google Translation API	0.408	0.207	0.399
Opus MT	**0.519**	**0.308**	**0.509**
GPT 4 Turbo	0.327	0.138	0.316

Table 3: Machine Translation Results

4.2 GPT-4 Translation Evaluation Prompt

Prompt= ''' You are an expert translator who can translate legal contracts, legal documents and terms and conditions from English to Arabic. I will give you a legal contract in English and in Arabic and I would like you to give me a score from 1 to 5 to rate how good is the transaltion, where 1 means very bad and 5 means accurate and complete translation. The final answer should be a number between 1 and 5 only.

English_Contract: {en_contract}

Arabic_contract: {ar_contract}

Answer: '''

Figure 18: Prompt used with GPT-4 to evaluate the translation of LegalBench

4.3 GPT-4 Translation Result Evaluation

Prompt= ''' You are an expert translator who can translate legal contracts, legal documents and terms and conditions from English to Arabic. I will give you a legal contract in English and in Arabic and I would like you to give me a score from 1 to 5 to rate how good is the transaltion, where 1 means very bad and 5 means accurate and complete translation. The final answer should be a number between 1 and 5 only.

English_Contract: {en_contract}

Arabic_contract: {ar_contract}

Answer: '''

Figure 19: GPT-4 Evaluation Result

4.4 Sample of the Translated Data

{

"index":"0",

"question": “Will Google help me if I think someone has taken and used content I’ve created without my permission?",

"answer": “Yes",

"contract": “Your content\nSome of our services give you the opportunity to make your content publicly available for example, you might post a product or restaurant review that you wrote, or you might upload a blog post that you created.\nSee the Permission to use your content section for more about your rights in your content, and how your content is used in our services\nSee the Removing your content section to learn why and how we might remove user-generated content from our services\nIf you think that someone is infringing your intellectual property rights, you can send us notice of the infringement and well take appropriate action. For example, we suspend or close the Google Accounts of repeat copyright infringers as described in our Copyright Help Centre.",

"Question_translation":"هل ستساعدني Google إذا اعتقدت أن شخصًا ما قد أخذ المحتوى الذي أنشأته واستخدمه دون إذن مني؟",

"answer_Translation":"أجل",

"contract_Translation":" تمنحك بعض خدماتنا الفرصة لجعل المحتوى الخاص بك متاحًا للجمهور ، على سبيل المثال ، يمكنك نشر منتج أو مراجعة لمطعم كتبته ، أو يمكنك تحميل مشاركة مدونة قمت بإنشائها. راجع قسم إزالة المحتوى الخاص بك لمعرفة لماذا وكيف يمكننا إزالة المحتوى الذي ينشئه المستخدمون من خدماتنا إذا كنت تعتقد أن شخصًا ما ينتهك حقوق الملكية الفكرية الخاصة بك ، فيمكنك إرسال إشعار إلينا بالانتهاك واتخاذ الإجراء المناسب. على سبيل المثال ، نقوم بتعليق أو إغلاق حسابات Google لمنتهكي حقوق الطبع والنشر المتكررين كما هو موضح في مركز مساعدة حقوق الطبع والنشر الخاص بنا."

},

Figure 20: Sample of the Translated Data

E QA

You are an expert Q&A system that is trusted around the world
for your factual accuracy.
Always answer the query using the provided context information, and not prior
knowledge. Ensure your answers are fact-based and accurately reflect the context
provided.
Some rules to follow:
1. Never directly reference the given context in your answer.
2. Avoid statements like 'Based on the context, ...' or 'The context information
...' or anything along those lines.
3. Focus on succinct answers that provide only the facts necessary, do not be
verbose.Your answers should be max two sentences, up to 250 characters.

{context_str}

Given the context information and not prior knowledge, answer the query in Arabic.
Query: {query_str}
Answer:

Figure 21: QA model prompt

Considering the reference answer and the new answer to the following question, on a scale of 0 to 5, where 5 means the same and 0 means not at all similar, how similar in meaning is the new answer to the reference answer? Respond with just a number and no additional text.

Question: "{question}"
Reference answer: "{ground_truth}"
Reference new answer:
{llm_answer}

Figure 22: LM Judge answer similarity prompt

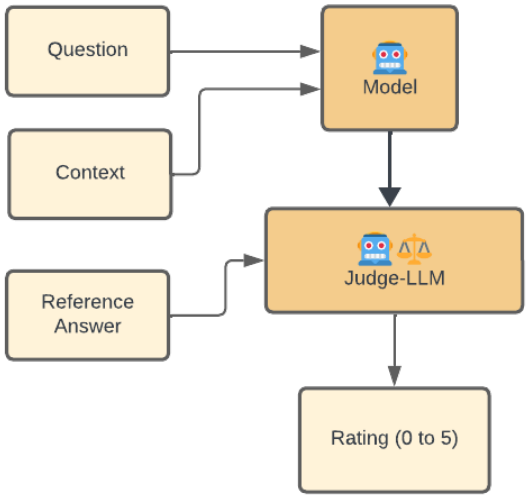

Figure 23: LLM as a judge

F Data statistics

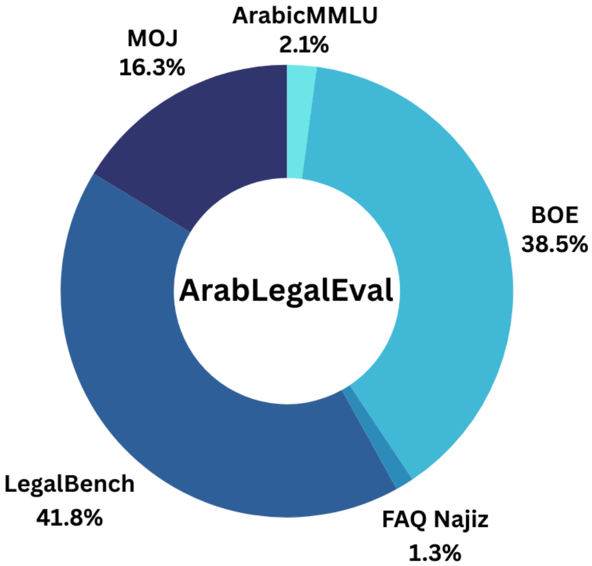

Figure 24: Source documents used to create ArabLegal-Eval and their percentages

Source	Category	Value
LegalBench	Privacy Policy Entailment	4385
	Privacy Policy QA	10931
	Contracts QA	88
	Consumer Contracts QA	400
BOE	Rules Count	448
	Rules Count - Subject	14134
MOJ	Regulations	67
	Regulations - Subjects	5720
	Circular	388
FAQ Najiz	FAQ Najiz	492
ArabicMMLU	ArabicMMLU	800

Table 4: Tasks counts in ArabLegalEval and their source documents.

G MCQs Evaluation

7.1 DSPy Signatures

```
You are given an Arabic Multiple Choice Question, answer it
by choosing the correct option with the letter included.

---

Follow the following format.

Question: Arabic MCQ question
Option: Option

---

Question: {question}
```

Figure 25: DSPy signature for few-shot prompt

```
You are given an Arabic Multiple Choice Question, answer
it by choosing the correct option with the letter included.

---

Follow the following format.

Question: Arabic MCQ question
Reasoning: Let's think step by step in order to ${produce
the option}. We ...
Option: Option

---

Question: {question}
```

Figure 26: DSPy signature for few-shot with CoT prompt

7.2 ArabicMMLU Results with DSPy optimization

Table 5 displays the performance of models on ArabicMMLU's law and political science subsets with optimized prompts. The prompts were optimized with 15 examples as labeled inputs from law and political science subsets, respectively. Therefore, the accuracies were computed using the remaining 299 and 195 samples from the respective law and political science sets.

Table 5: Experimental results on Law and Political Science subsets of ArabicMMLU

		Original	Few-shot	Few-shot (GPT-4 Teacher)	CoT Few-shot	CoT Few-shot (GPT-4 Teacher)
GPT-4	Law	73.60%	74.90%	-	**77.30%**	-
	Political Science	70.80%	**75.90%**	-	70.80%	-
GPT-4o	Law	68.20%	**82.90%**	-	81.90%	-
	Political Science	71.30%	**74.90%**	-	73.80%	-
Command R	Law	56.90%	61.50%	65.20%	69.20%	**69.90%**
	Political Science	65.10%	67.20%	67.70%	**72.80%**	66.70%
Command R+	Law	56.50%	61.20%	65.60%	**72.20%**	68.20%
	Political Science	65.10%	66.70%	67.70%	**72.30%**	66.20%
Llama3 8B	Law	56.90%	61.50%	65.90%	**73.20%**	69.20%
	Political Science	65.10%	67.20%	67.20%	**73.30%**	65.10%
Llama3 70B	Law	56.60%	61.50%	65.90%	**73.90%**	67.60%
	Political Science	65.10%	67.20%	67.70%	**70.30%**	66.70%
Aya101	Law	45.80%	42.80%	19.10%	21.70%	**47.10%**
	Political Science	47.20%	50.80%	51.30%	**51.80%**	26.10%

H Details for Arabic LegalBench Experiments Results

Table 6: Experiments Results on Arabic LegalBench Data

Technique	Model	consumer_contract_qa	contract_qa	privacy_policy_entailment	privacy_policy_qa
few_shot	Cohere-command-r	43%	89%	61%	64%
	Cohere-command-r-plus	35%	**99%**	41%	**68%**
	gpt-35-turbo-16k-2023-03-15-preview	51%	89%	50%	66%
	gpt4-2024-02-15-preview	62%	97%	**62%**	67%
	Meta-Llama-3-70B-Instruct	**89%**	62%	49%	40%
	Meta-Llama-3-8B-Instruct	35%	1%	0%	14%
One_Shot	Cohere-command-r	81%	95%	61%	66%
	Cohere-command-r-plus	89%	94%	**66%**	68%
	gpt-35-turbo-16k-2023-03-15-preview	73%	92%	47%	64%
	gpt4-2024-02-15-preview	**90%**	**96%**	60%	**74%**
	Meta-Llama-3-70B-Instruct	86%	**96%**	50%	60%
	Meta-Llama-3-8B-Instruct	65%	56%	4%	62%
Zero_shot_basic	Cohere-command-r	88%	**94%**	44%	**65%**
	Cohere-command-r-plus	59%	50%	**54%**	43%
	gpt-35-turbo-16k-2023-03-15-preview	82%	50%	36%	30%
	gpt4-2024-02-15-preview	31%	92%	37%	60%
	Meta-Llama-3-70B-Instruct	**90%**	90%	38%	35%
	Meta-Llama-3-8B-Instruct	75%	26%	12%	35%
Zero_shot_detailed	Cohere-command-r	55%	62%	44%	63%
	Cohere-command-r-plus	43%	29%	**52%**	**65%**
	gpt-35-turbo-16k-2023-03-15-preview	77%	59%	35%	62%
	gpt4-2024-02-15-preview	57%	**92%**	43%	55%
	Meta-Llama-3-70B-Instruct	**89%**	59%	44%	62%
	Meta-Llama-3-8B-Instruct	45%	18%	13%	23%

I Araic LegalBench Prompts

Classify if the clause is relevant ذات صلة to answering the question or irrelevant غير ذات صلة. The final answer should be the phrase ذات الصلة or غير ذات صلة only.

Clause: {clause}

Question: {question}

Figure 27: Zero-Shot Basic Technique Example

You are an expert legal professional who can understand legal documents and answer questions related to legal context. Given a clause and a question in Arabic, classify if the clause is relevant ذات صلة to answering the question or irrelevant غير ذات صلة. Provide your Answer in Arabic with only ذات صلة or غير ذات صلة.

Clause: {clause}

Question: {question}

Figure 28: Zero-Shot Detailed Technique Example

Classify if the clause is relevant ذات صلة to answering the question or irrelevant غير ذات صلة. The final answer should be the phrase ذات الصلة or غير ذات صلة only.

Clause: {Clause 1 Example}

Question: {Question1 Example}

Answer: ذات صلة

Clause: {Clause 2 Example}

Question: {Question 2 Example}

Answer: غير ذات صلة

Clause: {clause}

Question: {question}

Answer:

Figure 29: One-Shot Basic Technique Example

Given a text from a contract and a related question, answer the question only with أجل for Yes or لا for No in Arabic, don't add any extra details to your asnwer.

Contract: {contract 1 example}

Question: {Question 1 example}

Answer: أجل

Contract: {Contract 2 example}

Question: {Question 2 example}

Answer: أجل

Contract: {Contract 3 example}

Question: {Question 3 example}

Answer: لا

Contract: {Contract 4 example}

Question: {question 4 example}

Answer: لا

Contract: {clause}

Question: {question}

Figure 30: Zero-Shot Basic Technique Example

J Dataset Quantitative Analysis

10.1 MCQs

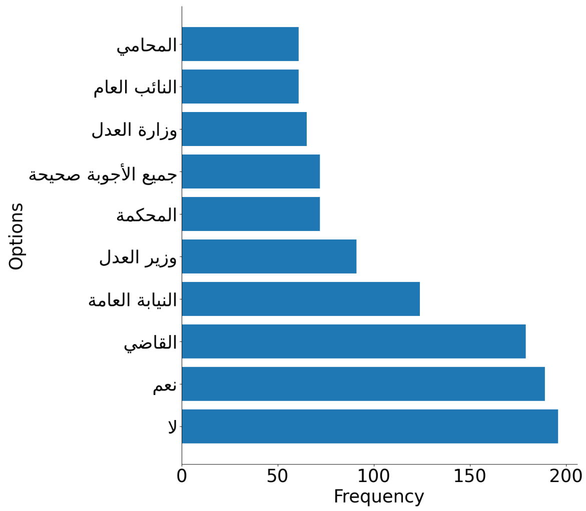

Figure 31: Top 10 most frequent choices

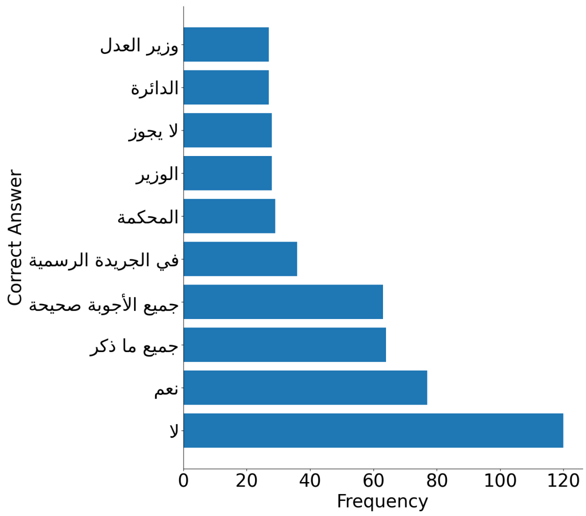

Figure 32: Top 10 most frequent correct answers

	Min Length	Max Length	Avg Length
Context	4	1497	46.0
Question	2	48	12.0

Table 7: Length of context documents and questions (words)

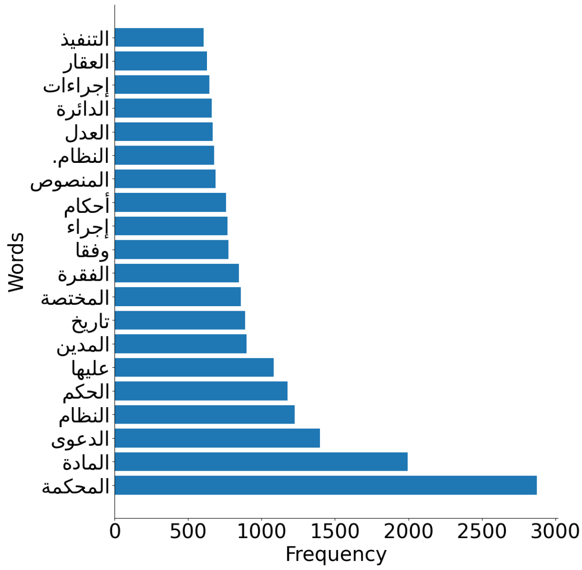

Figure 33: Top 20 frequency of words in the context of the MCQs

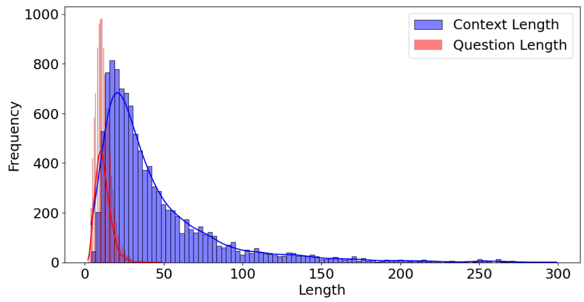

Figure 34: Length histogram of questions and contexts. Outliers longer than 300 words are excluded

10.2 QA

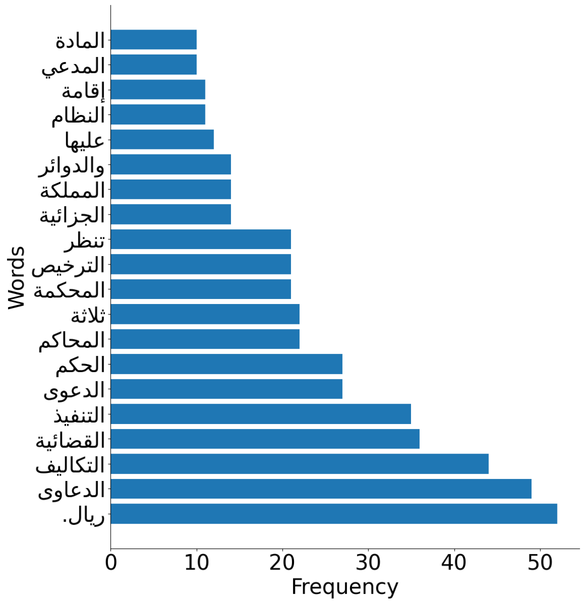

Figure 35: Top 20 frequency of words in the context of the QAs

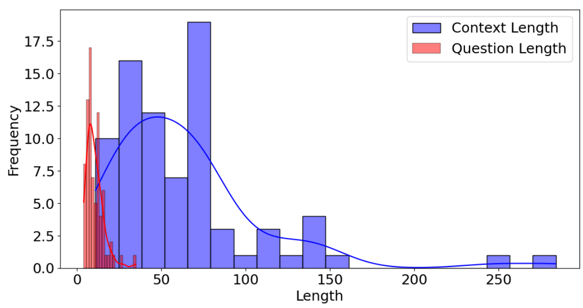

Figure 36: Length histogram of questions and contexts.

	Min Length	Max Length	Avg Length
Context	11	284	64.0
Question	4	35	11.0

Table 8: Length of context documents and questions (words)

10.3 Arabic LegalBench

Dataset	Size	Average Context Length (Words)	Min Context Length (Words)	Max Context Length (Words)	Average Question Length (Words)	Min Question Length (Words)	Max Question Length (Words)	Answer Choices	Answers Distribution
privacy_policy_qa	10508	25	5	88	8	3	23	ذات الصلة / غير ذات صلة	ذات الصلة: 5453, 52% غير ذات صلة: 5055, 48%
consumer_contract_qa	400	424	70	1208	14	4	38	أجل / لا	أجل: 180, 45% لا: 220, 55%
contract_qa	88	34	8	86	6	4	9	أجل/ لا	أجل: 43, 49% لا: 45, 51%
privacy_policy_entailment	4343	85	4	439	15	10	29	صحيح / غير صحيح	صحيح: 3745, 86% غير صحيح: 598, 14%

Figure 37: Summary Statistics for Arabic LegalBench

10.4 Experts Performance on the Tasks

To establish a human performance baseline and provide a meaningful comparison for the LLMs' performance, a sample of the data was presented to legal experts for evaluation. These experts, with their specialized knowledge and experience, were tasked with solving the same problems that were presented to the LLMs. Table 9 summarize the results of the experts assessment.

	MCQs	NajezQA	Arabic LegalBench
F1 Score	-	-	0.73
Accuracy	0.9	0.65	0.8

Table 9: Experts Answers