ALARB: An Arabic Legal Argument Reasoning Benchmark

Harethah Abu Shairah[1], Somayah AlHarbi[2], Abdulaziz AlHussein[2], Sameer Alsabea[1], Omar Shaqaqi[1], Hebah AlShamlan[2], Omar Knio[1], George Turkiyyah[1]

[1]King Abdullah University of Science and Technology (KAUST), [2]THIQAH

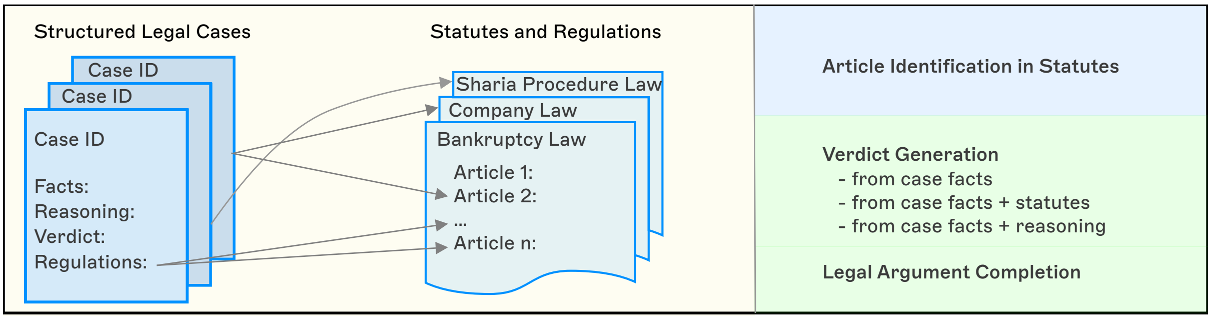


Figure 1: ALARB includes a dataset of structured legal cases. Each case lists the facts presented by the plaintiff and defendant, and an explicit step-by-step chain of the argument reasoning of the court leading to a verdict. Cases are linked to individual articles of applicable statutes and regulations. A set of legal reasoning tasks leverages the data. ALARB is available **here**.

Abstract

We introduce ALARB, a dataset and suite of tasks designed to evaluate the reasoning capabilities of large language models (LLMs) within the Arabic legal domain. While existing Arabic benchmarks cover some knowledge-intensive tasks such as retrieval and understanding, substantial datasets focusing specifically on multistep reasoning for Arabic LLMs, especially in open-ended contexts, are lacking. The dataset comprises over 13K commercial court cases from Saudi Arabia, with each case including the facts presented, the reasoning of the court, the verdict, as well as the cited clauses extracted from the regulatory documents. We define a set of challenging tasks leveraging this dataset and reflecting the complexity of real-world legal reasoning, including verdict prediction, completion of reasoning chains in multistep legal arguments, and identification of relevant regulations based on case facts. We benchmark a representative selection of current open and closed Arabic LLMs on these tasks and demonstrate the dataset's utility for instruction tuning. Notably, we show that instruction-tuning a modest 12B parameter model using ALARB significantly enhances its performance in verdict prediction and Arabic verdict generation, reaching a level comparable to that of GPT-4o.

1 Introduction

The Arabic capabilities of LLMs have been rapidly improving, and many recent models, both closed and open, now demonstrate remarkable fluency and linguistic quality in their generated outputs. This enhanced performance facilitates the development of practical support systems in various knowledge-intensive domains. It also underscores the importance of developing targeted, native Arabic benchmarks to thoroughly evaluate these models in scenarios requiring complex, multistep reasoning.

In English, a variety of benchmarks exist for evaluating the capabilities of emerging LLMs. Several influential benchmarks, such as (Wang et al., 2018; Hendrycks et al., 2021a), have significantly shaped the development of earlier models. As these benchmarks quickly become saturated by rapidly improving models—GPT-4.1, for instance, achieves more than 90% accuracy on MMLU—new benchmarks continue to emerge, offering fresh evaluation challenges (Zhong et al., 2024; Phan et al., 2025; Guha et al., 2023). Notably, tasks requiring multistep reasoning have become an essential focus in recent benchmarks, reflecting the capabilities of current-generation LLMs to plan and execute sequences of reasoning steps prior to generating their outputs.

In contract, there is comparatively a dearth of

Proceedings of The Third Arabic Natural Language Processing Conference, pages 389–406
November 8-9, 2025 ©2025 Association for Computational Linguistics

benchmarks to evaluate the emerging generative abilities of Arabic LLMs, and many existing evaluation and benchmarking resources are in fact translated from English. While in some domains, translations from English or other languages may be quite reasonable, there are others in which LLMs are expected to reason in contexts where social and cultural norms are relevant factors and where translated datasets may suffer from unintended omissions or systematic bias. In order to address this gap, benchmarks that include reasoning tasks in native Arabic contexts are needed.

The Arabic legal domain provides an ideal setting for benchmarking Arabic LLMs, particularly in open-ended scenarios representative of real-world complexity. Legal reasoning involves structured argumentation and contextual sensitivity, and requires flexible inference to handle uncertainties and plausible interpretations that do not exist in mathematical reasoning and inference tasks in closed systems. Additionally, legal tasks often involve linguistic complexity, nuanced text interpretation, and adherence to formal conventions, further testing Arabic comprehension and generation skills. Finally, Arabic remains notably absent from influential multilingual legal datasets (Niklaus et al., 2024), underscoring the importance of developing specialized Arabic legal datasets.

Towards this end, we introduce ALARB, a dataset specifically designed to support the multistep reasoning tasks needed for following legal arguments and predicting verdicts. The dataset is derived from original Arabic judicial sources of cases that appeared in front of commercial courts in Saudi Arabia in recent years.

Our contributions can be summarized as follows:

- We present a 13K+ structured legal cases dataset to support legal argument reasoning, along with their governing statutes.
- We introduce a set of tasks involving this dataset, including identifying applicable articles from case facts and variants of verdict generation.
- We evaluate the performance of the leading open Arabic models on these tasks, and show that the dataset can be used to finetune a 12B model to result in performance that rivals that of GPT-4o.

2 Related Work

2.1 Arabic LLM benchmarks

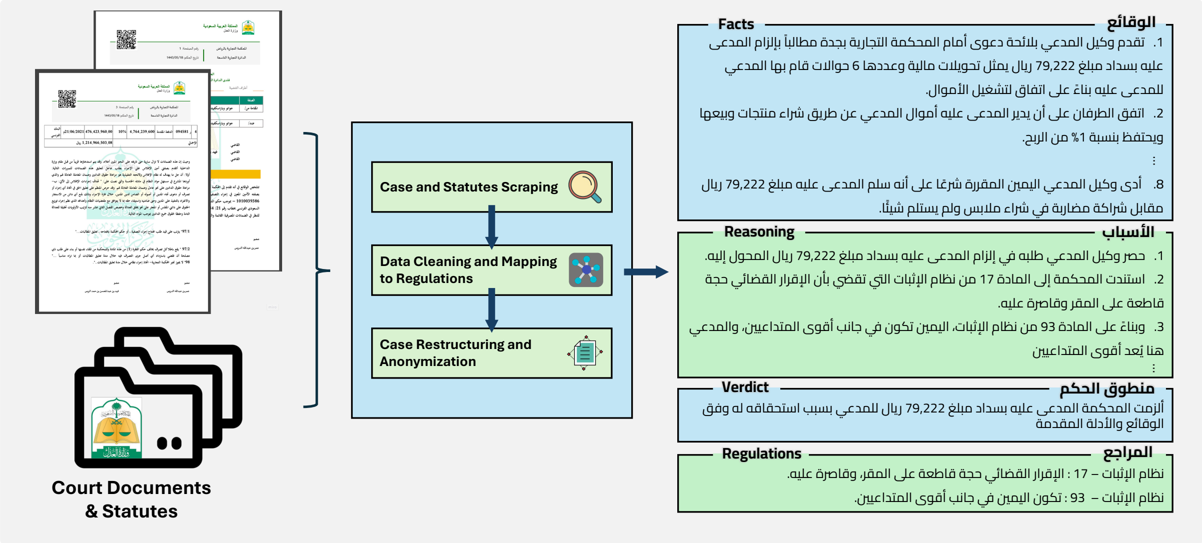
Early benchmarks of Arabic language models largely focused on linguistic-level text classification tasks (Antoun et al., 2020; Abdul-Mageed et al., 2021) consistent with the limited capabilities of models at the time. Despite interest in evaluating deeper linguistic proficiency (Kwon et al., 2023; Sibaee et al., 2025), recent benchmarks have shifted towards more knowledge intensive and reasoning tasks to accompany the rising capabilities of current generation Arabic LLMs. In this category of Arabic LLMs, we include both Arabic-centric models (Sengupta et al., 2023; Huang et al., 2024)—models whose training data is mostly focussed on Arabic and English, as well as the multilingual models such as (Team, 2025; OpenAI, 2024b; Yang et al., 2025) that include Arabic among dozens of supported languages.

Among popular benchmarks for Arabic LLMs, we mention AlGhafa (Almazrouei et al., 2023) and ArabicMMLU (Koto et al., 2024) that have curated multiple choice questions (MCQs) spanning a variety of general knowledge questions. The performance of Arabic models on these and other benchmarks are tracked in public leaderboards including the Open Arabic LLM Leaderboard (El Filali et al., 2024) and BALSAM (King Salman Global Academy for Arabic Language, 2024). There has also been interest in benchmarking Arabic LLM models for cultural alignment (Qian et al., 2024; Mousi et al., 2025).

There is however a need for the evaluation of emerging Arabic LLMs on more challenging tasks that require the generation of conclusions and explanations in open-ended and specialized domains. A task in the domain of poetry understanding and explanation is described in (Alghallabi et al., 2025).

2.2 Legal reasoning benchmarks and tasks

The legal domain has seen tremendous interest in the use of LLMs in tasks related to legal research and writing tools targeting professionals and the public, motivating the need for benchmarking in this domain. Early benchmarks (Chalkidis et al., 2022; Hendrycks et al., 2021b) focussed on classification and recognition tasks in judgement prediction, clause identification, and related tasks. More recent efforts (Guha et al., 2023; Fei et al., 2024; Nigam et al., 2024, 2025) have substantially expanded the evaluation tasks to include a


Court Documents & Statutes
Case and Statutes Scraping
Data Cleaning and Mapping to Regulations
Case Restructuring and Anonymization
Facts
الوقائع
Reasoning
الأسباب
Verdict
منطوق الحكم
Regulations
المراجع

1. تقدم وكيل المدعي بلائحة دعوى أمام المحكمة التجارية بجدة مطالباً بإلزام المدعى عليه بسداد مبلغ 79,222 ريال يمثل تحويلات مالية وعددها 6 حوالات قام بها المدعي للمدعى عليه بناءً على اتفاق لتشغيل الأموال.
2. اتفق الطرفان على أن يدير المدعى عليه أموال المدعي عن طريق شراء منتجات وبيعها ويحتفظ بنسبة 1% من الربح.

⋮

8. أدى وكيل المدعي اليمين المقررة شرعًا على أنه سلم المدعى عليه مبلغ 79,222 ريال مقابل شراكة مضاربة في شراء ملابس ولم يستلم شيئًا.

1. حصر وكيل المدعي طلبه في إلزام المدعى عليه بسداد مبلغ 79,222 ريال المحول إليه.
2. استندت المحكمة إلى المادة 17 من نظام الإثبات التي تقضي بأن الإقرار القضائي حجة قاطعة على المقر وقاصرة عليه.
3. وبناءً على المادة 93 من نظام الإثبات، اليمين تكون في جانب أقوى المتداعيين، والمدعي هنا يُعد أقوى المتداعيين

⋮

ألزمت المحكمة المدعى عليه بسداد مبلغ 79,222 ريال للمدعي بسبب استحقاقه له وفق الوقائع والأدلة المقدمة

نظام الإثبات – 17 : الإقرار القضائي حجة قاطعة على المقر، وقاصرة عليه.
نظام الإثبات – 93 : تكون اليمين في جانب أقوى المتداعيين.

Figure 2: **Data Preparation Workflow.**

broader range of legal reasoning tasks, specifically designed to test logical reasoning, judgment prediction, and question-answering abilities of models. In Arabic, a benchmark inspired by LegalBench appeared in (Hijazi et al., 2024).

However, these benchmarks have not addressed tasks that require understanding or generating chains of legal arguments in support of a decision, making it questionable how much legal reasoning of models is being evaluated. In fact, legal LLMs are still prone to hallucinations (Magesh et al., 2025) that are partly attributed to the models' inability to reason correctly through the text to arrive at the proper conclusion. Reasoning-focused datasets and tasks are needed to support reliable RAG systems, explainability, and trustworthiness of LLMs in legal domains. (Zheng et al., 2025; Chlapanis et al., 2024) are efforts in this direction.

3 Dataset

The ALARB dataset contains legal cases from commercial courts in Saudi Arabia with their applicable statutes. In this section we describe the process of curating this data and its results.

3.1 Data Curation

Figure 2 depicts the data preparation workflow.

Case and Statutes Scraping. Court case descriptions are scraped from the KSA Ministry of Justice (MoJ) website. Each case description includes the facts of the case (arguments presented by the plaintiff and defendant to the court) and the reasoning of the court. Each is usually a few paragraphs long. The description also includes a verdict that is short and authoritative in tone. Eight statues, along with their implementing regulations, were identified as the governing documents for these cases and were also scraped. Each of these governing documents is organized into articles representing specific provisions (مادة).

Data Cleaning and Mapping to Regulations. This involved identifying the statutes and regulation documents, as well as the specific articles from them, that are referenced in each case. These articles are not listed separately in the case descriptions but appear in-line in the text describing the reasoning of the court. In addition, these articles and their statutes are referred to differently in different cases, with inconsistencies in the naming conventions for the same legal document and in the way article numbers appear in the descriptions. This is essentially a named-entity recognition (NER) task and we used an LLM for it. Our experiments showed that modern LLMs can generally understand the context needed to identify the statute names and article numbers referred to in the text. For additional robustness however, this process was repeated twice using different prompts, and the union of the two different outputs was used to minimize the risk of missing any relevant articles and regulations.

Case Restructuring and Anonymization. This involves arranging the facts of a case into a list of individual items, each representing a single fact and generally written in a sentence or two in the text. Similarly, the reasoning was structured as a list of individual steps, each representing a sin-

Articles	Referenced	Document
30	2	لائحة المعلومات والوثائق
129	4665	نظام الإثبات
329	82	نظام الإفلاس و لوائحه التنفيذية
371	84	نظام التنفيذ و لوائحه التنفيذية
281	714	نظام الشركات و لوائحه التنفيذية
356	9652	نظام المحاكم التجارية و لوائحه التنفيذية
55	264	نظام المحاماة و لوائحه التنفيذية
876	3824	نظام المرافعات الشرعية و لوائحه التنفيذية

Table 1: **Statistics of Referenced Legal Statutes.**

Field	Words			Steps		
	Min	Max	Avg	Min	Max	Avg
Facts	31	398	181	3	11	8
Reasoning	18	296	129	1	11	6
Regulations	0	977	186	0	15	3
Verdict	5	26	13	N/A		

Table 2: **Dataset Summary Statistics.**

gle thought in the reasoning process. The scraped textual descriptions of the facts and the reasoning also often contained identifiable information about plaintiffs and defendants, which needed to be removed. Prompts were designed to restructure both the facts and reasoning sections into clear steps and to remove irrelevant or sensitive information, and this step was done with an LLM. The quality of the outputs was verified manually on random samples.

Appendix A shows an example of the generated representation structured as: a list of individual facts, a sequence of reasoning steps, a court verdict, and keys to full text descriptions of cited articles.

3.2 Dataset statistics

Table 1 summarizes the data of legal documents included in the dataset. Each entry shows the number of articles contained in the corresponding statute. On average, each article in the statutes has about 47 words. Also shown in the table are the number of times articles from the statute are referenced. In many of the cases, multiple articles from the same statute are referenced.

Figure 3 shows the composition of the 13,344 legal cases of the dataset. The top left histogram shows their word count distribution, including all text from the list of facts, steps of the reasoning, the verdict, and the referenced articles. There were a few outliers but we had generally chosen cases that are not too lengthy, resulting in the peak of the distribution being around 500 words. The three other histograms show the distribution of the sizes of the case fact lists, reasoning step lists, and the number of articles explicitly referenced from the statutes. We note that most cases involve about half a dozen discrete reasoning steps and use only a few articles in arriving at the verdict. Table 2 shows additional details of these distributions, with the min, max and average number of words and discrete steps. Table 3 shows the verdict distribution of the court rulings, which includes a substantial portion of cases that were deemed not within the court's jurisdiction, with the motivating rationale articulated in the reasoning.

For Plaintiff	For Defendant	Court Dismissal
62%	5%	33%

Table 3: **Case Verdict Breakdown.**

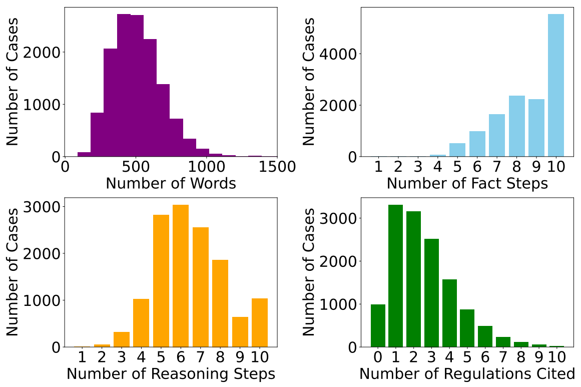


Figure 3: **Distributions of Words and Steps.**

4 Benchmark Tasks

ALRAB introduces two main categories of tasks aimed at evaluating a model's capacity for legal reasoning.

4.1 Verdict Prediction Tasks

The first category focuses on verdict generation in different task setups designed to evaluate the models' capacity for legal reasoning with varying amounts of given contextual information. These tasks specifically test how well the model can analyze case details and generate a verdict grounded in the relevant laws and regulations. In each setup, the model is provided with selected information from the case and is expected to produce a legally sound verdict.

Task 1: From Facts Only. In this task, the models are provided with only the factual details of each case. They are expected to analyze these facts to generate a reasoning chain and a verdict solely based on their understanding of the case.

Task 2: From Facts and Relevant Articles. In this task, the models receive both the case facts and the specific legal articles that were referenced in the court's reasoning. The objective is to assess the model's ability to interpret and apply the relevant articles to the facts of a case and produce a reasoned verdict accordingly.

Task 3: From Facts and Court's Reasoning. In the setup, the models are given the case facts along with the court's official reasoning. Based on this combined input, they are tasked with predicting the final verdict. The objective is to evaluate how well they can understand legal arguments in the context of the facts and reach a verdict.

Task 4: Argument Completion. Tasks 2 and 3 above are two extremes in the spectrum of legal argument reasoning: one provides none of the reasoning of the court and the other provides it all. This task is an intermediate one that provides the models with the first few steps of the reasoning and asks them to complete it and reach a verdict. The task is parameterized by the number of omitted reasoning steps and obviously becomes more difficult as this number increases.

4.2 Article Identification Tasks

The second category of tasks is designed to evaluate the models' ability to identify and recognize the appropriate relevant articles in statutes based solely on their understanding of the case facts. To this end, we initially attempted to create a retrieval-based approach where, given only the case facts, the model would retrieve the relevant articles from the entire set of statutes and regulations available. We embedded all available regulations using text-embedding-large-3 (OpenAI, 2024a) and employed cosine similarity to retrieve the most relevant articles based on embedded case facts. However, the results were extremely poor , which led us to simplify our approach and generate two multiple-choice question tasks instead.

In these MCQ questions, the models are given the complete list of facts from a legal case and asked to choose the most applicable article from a list of four choices: one being an article actually cited in the court's reasoning and three other distractors. The distractors are constructed in two different ways described below, allowing the MCQs to have two levels of difficulty.

Task 1: Articles from the Same Statute In this task, the model is presented with three distractors randomly selected from the same statute as the correct answer. This configuration tests the model's ability to distinguish between somewhat related articles within the same statute. Many articles in the same regulatory document use the same exact words and phrases and require that models understand the full context of an article.

Task 2: Semantically Related Articles In this more challenging task, we employ semantic similarity via embeddings to retrieve articles closely related to the correct article. We utilized the `text-embedding-large-3` model (OpenAI, 2024a) for generating embeddings and calculated cosine similarity scores across the entire regulation corpus. The three most semantically similar articles serve as distractors. These may originate from different legal regulations rather than being confined to a single regulatory document. This creates a more sophisticated evaluation that tests the model's deeper understanding of regulatory nuances, semantic relationships, and subtle differences across various legal texts. A sample MCQ is shown in Figure 11.

5 Results

For all tasks, we conducted evaluations across a diverse set of models, varying in size, language capability (Arabic-centric and multilingual), and accessibility (open-source and proprietary). The list of models included in our evaluation is provided in Table 4. The benchmarks were performed on a subset of **1,329** legal cases.

5.1 Verdict Prediction Tasks Results

For the first category of tasks—*verdict prediction*—the models were provided with detailed prompts outlining both the expected output and the format of the response. In the two setups where the court's reasoning was not included as part of the input, the models were explicitly instructed to perform reasoning before generating a verdict.

To evaluate the predicted verdicts, we used `GPT-4o` as an LLM-as-a-judge (Zheng et al., 2023; Gu et al., 2024). The model was provided with both the predicted and actual verdicts and tasked with assessing their alignment. Reliable automatic evaluation of generated verdicts is not a simple task.

Model	Facts Only			Facts & Reasoning			Facts & Regulations		
	Correct	Partial	Incorrect	Correct	Partial	Incorrect	Correct	Partial	Incorrect
`AceGPT-v2-32B-Chat`	28.9	34.7	35.8	41	55.1	3.9	25.1	27.8	38.3
`AceGPT-v2-8B-Chat`	33.4	33.4	33.1	58.4	38.8	2.7	28.9	30.3	37.3
`ALLaM-7B-Instruct-preview`	14.1	42.7	43.1	39	56.4	4.5	17.2	44.3	38.4
`aya-expanse-32B`	32.9	33	33.9	**70.6**	26.7	2.7	36.3	32	31.5
`aya-expanse-8B`	25.6	38.8	35.6	61.9	34	4.1	24.6	40.7	34.6
`Falcon3-7B-Instruct`	8.7	20.2	70.9	28.7	40.1	31.1	8.7	18.1	73.1
`Gemma-3-12B-it`	15.8	51.8	32.4	51	46.2	2.8	29.6	40.8	29.6
`Gemma-3-4B-it`	13.3	46.2	40.3	46.9	39.2	13.9	24.5	38.5	36.9
`GPT-4o`	**38.7**	31.4	**29.9**	65.7	31.6	2.7	**46**	28.8	**25**
`GPT-o4-mini`	22.9	46	30.9	61.3	36.7	**2**	27.6	43.8	28.5
`Qwen3-14B`	31.5	36.5	31.9	64.5	31.5	4.1	44.6	28.7	26.7
`Qwen3-8B`	27.1	36.4	36.5	58.3	36.2	5.3	32.2	34.2	33.5

Table 4: **Verdict Prediction results:** LLMs Evaluation for Verdict Prediction Across Three Tasks.

Verdicts in commercial cases are not binary and generally require the calculation of fines, which must be done accurately. The judging prompt is shown in in Appendix B. It generates one of three evaluations:

- **CORRECT**: The predicted verdict fully matches the actual court verdict.
- **INCORRECT**: The predicted verdict does not align with the actual court verdict. It may award incorrect amounts, not recognize jurisdiction, or add unnecessary details.
- **PARTIALLY CORRECT**: The prediction demonstrates partial alignment but fails to fully match the court's decision, mostly in minor style and expression.

In the **facts-only** task, `GPT-4o` achieved the highest percentage of correct verdicts, while `Gemma 3-12B` achieved the highest rate of partially correct predictions.

In the **facts and court reasoning** task, `aya-expanse-32B` outperformed all models, followed by `GPT-4o` in the percentage of correct verdicts. Despite being provided with both the case facts and the court's reasoning, and only required to interpret the reasoning to reach a verdict, fewer than half of the models achieved more than **60**% accuracy. This outcome highlights the inherent complexity of correctly interpreting the dense Arabic legal language of the courts.

In the **facts and regulations** task, `GPT-4o` again led in performance, achieving a **46**% correct verdict rate, followed closely by `Qwen3-14B` at **44.6**%. Both models also recorded the lowest percentage of incorrect verdicts, suggesting that they successfully reasoned and applied relevant regulations in approximately **75**% of cases.

Interestingly, several models, including both versions of `AceGPT-v2`, `aya-expanse-8B`, and `Falcon-7B`, performed worse when provided with the relevant regulations compared to when they received only the facts. This suggests that the presence of large amounts of legal text in the context may have introduced confusion in models with less robust reasoning capabilities.

Both versions of `Qwen3` were evaluated with *thinking mode* enabled, allowing us to evaluate the effects of additional test-time reasoning. Under this configuration, the models demonstrated strong reasoning capabilities. `Qwen3-14B` achieves results that closely approach those of `GPT-4o`, and both `Qwen3` models consistently outperform `o4-mini` across most evaluation cases. Specifically, `Qwen3-14B` surpasses `o4-mini` in the percentage of correctly predicted verdicts across all three tasks. In the *Facts and Regulations* task, `Qwen3-14B` achieves a significantly higher rate of fully correct verdicts—**44.6**% compared to `o4-mini`'s **27.6**%—indicating nearly double the accuracy. Even the smaller `Qwen3-8B` model outperforms `o4-mini` in this task in terms of fully correct predictions.

Results for the **argument completion task** with given partial reasoning are discussed in Section 6.2, along with the performance of a fine-tuned model.

5.2 Article Identification Task Results

For the regulation identification task, we evaluated a subset of models on 1,159 MCQs for each of the two tasks. In the task where all answer choices were drawn from the same regulatory document, all models demonstrated strong performance, with accuracy exceeding **80**%. `GPT-4o` achieved the highest accuracy in this setup at 90.42%, followed

Model	Article Identification Accuracy	
	Same Regulation	Semantically Retrieved
AceGPT-v2-8B-Chat	81.79	52.72
Gemma-3-12B-it	82.63	67.47
Qwen3-14B	82.20	71.30
Qwen3-8B	84.60	67.90
GPT-o4-mini	90.07	73.59
GPT-4.1	86.71	**77.30**
GPT-4o	**90.42**	76.79

Table 5: **Article Identification Results.**

by GPT-4.1 at 86.71%. However, the task became significantly more challenging when semantically similar articles —retrieved using embedding-based similarity— were used as distractors. In this more difficult scenario, overall accuracy declined substantially, with GPT-4.1 achieving the highest score at 77.30%.

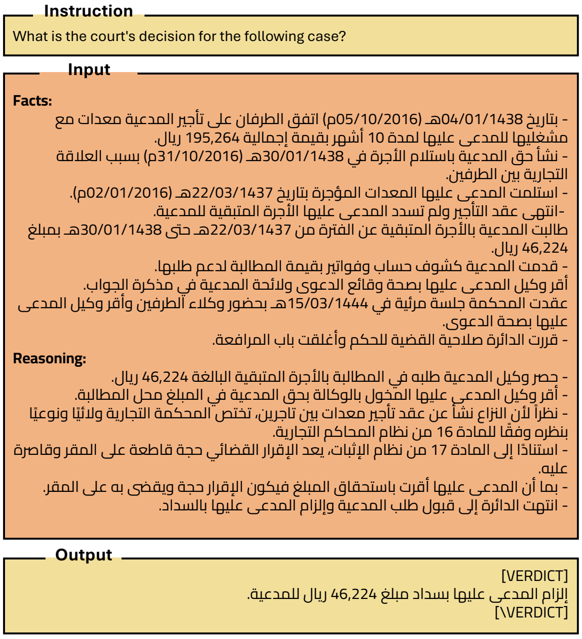
Overall, models with strong reasoning capabilities consistently performed well across both task categories, demonstrating their robustness in legal understanding, verdict prediction, and regulatory interpretation.

6 Additional Experiments

We explore the utilization of our dataset in three focused scenarios: Supervised Fine-tuning (SFT), completion of part of the court's reasoning to predict the verdict, and comparing English versus Arabic reasoning capabilities.

6.1 Supervised Fine-tuning

A primary application of our dataset is supervised fine-tuning of language models for legal reasoning. To investigate this, we constructed an instruction-tuning dataset derived from the existing cases for SFT and assessed whether fine-tuned models could leverage this dataset to enhance performance on predefined verdict prediction tasks. We initially defined three instruction-based tasks: **1)** Given legal case facts and applicable regulations, the model generates the reasoning and predicts the verdict. **2)** Given legal case facts, applicable regulations, and the court's reasoning, the model predicts the verdict. **3)** Given case facts, applicable regulations, and the final verdict, the model infers the court's reasoning. For task variability, we created multiple instructions per task (details available in Appendix C). Subsequently, we converted the training portion of our dataset into training samples for instruction-tuning, as illustrated in Figure 4. We fine-tuned Google's Gemma-3-12B-it using these instructions and evaluated its performance on our benchmark tasks to measure improvements from fine-tuning.



Instruction

What is the court's decision for the following case?

Input

Facts:

- بتاريخ 04/01/1438هـ (05/10/2016م) اتفق الطرفان على تأجير المدعية معدات مع مشغليها للمدعى عليها لمدة 10 أشهر بقيمة إجمالية 195,264 ريال.
- نشأ حق المدعية باستلام الأجرة في 30/01/1438هـ (31/10/2016م) بسبب العلاقة التجارية بين الطرفين.
- استلمت المدعى عليها المعدات المؤجرة بتاريخ 22/03/1437هـ (02/01/2016م).
-انتهى عقد التأجير ولم تسدد المدعى عليها الأجرة المتبقية للمدعية.
طالبت المدعية بالأجرة المتبقية عن الفترة من 22/03/1437هـ حتى 30/01/1438هـ بمبلغ 46,224 ريال.
- قدمت المدعية كشوف حساب وفواتير بقيمة المطالبة لدعم طلبها.
أقر وكيل المدعى عليها بصحة وقائع الدعوى ولائحة المدعية في مذكرة الجواب.
عقدت المحكمة جلسة مرئية في 15/03/1444هـ بحضور وكلاء الطرفين وأقر وكيل المدعى عليها بصحة الدعوى.
- قررت الدائرة صلاحية القضية للحكم وأغلقت باب المرافعة.

Reasoning:

- حصر وكيل المدعية طلبه في المطالبة بالأجرة المتبقية البالغة 46,224 ريال.
- أقر وكيل المدعى عليها المخول بالوكالة بحق المدعية في المبلغ محل المطالبة.
- نظراً لأن النزاع نشأ عن عقد تأجير معدات بين تاجرين، تختص المحكمة التجارية ولائيًا ونوعيًا بنظره وفقًا للمادة 16 من نظام المحاكم التجارية.
- استنادًا إلى المادة 17 من نظام الإثبات، يعد الإقرار القضائي حجة قاطعة على المقر وقاصرة عليه.
- بما أن المدعى عليها أقرت باستحقاق المبلغ فيكون الإقرار حجة ويقضى به على المقر.
- انتهت الدائرة إلى قبول طلب المدعية وإلزام المدعى عليها بالسداد.

Output

[VERDICT]
إلزام المدعى عليها بسداد مبلغ 46,224 ريال للمدعية.
[\VERDICT]

Figure 4: **SFT Training:** Example from the verdict prediction task.

The model went through full parameter fine-tuning on $12,012$ instruction-output pairs for 4 epochs, with an initial learning rate of $5e^{-6}$ with cosine scheduling, a per device batch size of 2 on 3 A100 GPUs, and 2 gradient accumulation steps.

Table 6 summarizes the performance of the fine-tuned model on our 1,329 case test set across the three verdict prediction tasks, highlighting performance gains and drops. The model demonstrates significant improvements across all three tasks, bringing it up on par with the best models in Table 4. The biggest improvements are seen in the "Facts" only task, where the model has to work the hardest to reach the correct verdict. These results highlight the effectiveness of these legal cases as a dataset that can be used for instruction tuning for legal reasoning.

6.2 Partial Reasoning

Table 4 shows a consistent pattern: models consistently exhibit lower rates of incorrect verdict predictions when explicitly provided with court reasoning, compared to when they must infer reasoning independently. To further investigate this behavior, we ran the reasoning completion task testing how the models perform when provided with only a subset of the reasoning steps. Starting with

Model	**Facts**			**Facts & Reasoning**			**Facts & Regulations**		
	Correct	Partial	Incorrect	Correct	Partial	Incorrect	Correct	Partial	Incorrect
`Gemma-3-12B-SFT`	37.3 (+21.5)	38.6 (-13.2)	24.1 (-8.3)	65.9 (+14.9)	31 (-15.2)	3.1 (+0.3)	45.3 (+15.7)	35.7 (-5.1)	19 (-10.6)

Table 6: **Fine-tuning Impact**: Gemma-3-12B-SFT's performance on verdict prediction compared to base model.

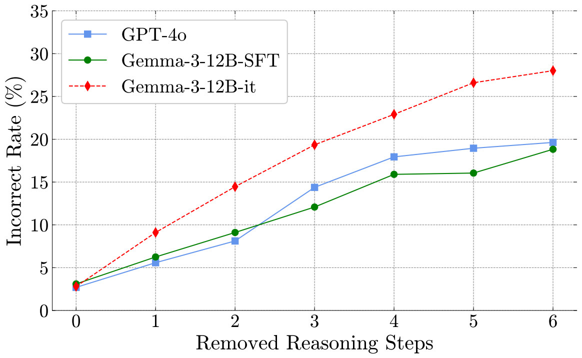


Figure 5: **Error rate with partial reasoning provided.**

the verdict prediction task involving case facts, applicable regulations, and all n reasoning steps, we progressively removed the final $k \in \{0, 1, \ldots, 6\}$ reasoning steps and measured model performance at each stage.

Figure 5 illustrates the increase in error rates as fewer reasoning steps are provided. As anticipated, all models deteriorate in performance when reasoning steps are removed. However, the SFT model demonstrates superior capability at using partial reasoning to reach correct verdicts, surpassing GPT-4o when three or more steps are omitted.

6.3 Reasoning In English

State-of-the-art LLMs are typically trained on extensive multilingual corpora, enabling them to converse and reason across various languages; however, English remains dominant within these datasets. Given that our dataset comprises legal cases exclusively in Arabic, all previously reported results were obtained by explicitly prompting the models to reason and provide verdict predictions in Arabic. We further investigate whether changing the reasoning language from Arabic to English influences model performance. For this experiment, we randomly sampled 100 cases from our test set and used `GPT-4o` to translate only the verdicts into English, avoiding translation of entire cases due to observed quality degradation in translating legal texts. Using these partially translated cases, we explicitly prompted the models to reason and produce verdict predictions in English for the "Facts & Regulations" task.

Model	**Facts & Regulations**		
	Correct	Partial	Incorrect
`Gemma-3-12B-it`	39 (+9.4)	32 (-8.8)	29 (-0.6)
`GPT-4o`	45 (-1)	27 (-1.8)	28 (+3)

Table 7: **Reasoning In English:** English reasoning improves Gemma3's performance, but is not significant for `GPT-4o`.

Table 7 presents the changes in performance for `GPT-4o` and `Gemma-3-12B` when reasoning in English. `GPT-4o` shows minimal variation, with minor performance drops likely attributable to the reduced size of the test sample. On the other hand, `Gemma-3-12B` exhibits substantial improvement when reasoning in English, significantly increasing its rate of fully correct predictions. This suggests that, despite its multilingual training, `Gemma-3-12B` benefits greatly from reasoning in English, likely due to stronger linguistic alignment or familiarity. These findings seem to imply that using English reasoning, even for Arabic legal cases, may offer performance advantages for certain multilingual models, as they may be relying on an English-centric representation space for their internal reasoning (Etxaniz et al., 2024; Schut et al., 2025). Further research is needed to reach broader conclusion, however.

7 Conclusions

We introduced ALARB, a novel Arabic dataset specifically designed to benchmark legal reasoning capabilities in Arabic LLMs. The dataset features multiple variants of verdict prediction tasks, assessing models' abilities to comprehend legal linguistic nuances, accurately apply regulations to given cases, and produce legally sound reasoning chains. Our experiments demonstrate that reasoning-oriented models generally perform better on these tasks; however, significant opportunities for improvement remain. Additionally, we validated ALARB's effectiveness by fine-tuning a 12B-parameter model, resulting in substantial performance gains. For future work, we intend to leverage ALARB in the Reinforcement Learning (RL) post-training of Arabic reasoning models.

Limitations

While this study contributes to evaluating and improving Arabic LLMs, several limitations must be acknowledged and addressed in future work.

First, the dataset is limited to a particular area of the law, obtained from a single country, and is relatively limited in size. Additional diversity is needed to broaden its capabilities. Texts from some areas besides commercial law are publicly available and may be used. Ministries of Justice in many countries of the Arab world have digitized their documents and these represent valuable resources for expanding and enriching the dataset with different styles of reasoning.

Evaluation of the LLM-as-a-judge in verdict prediction tasks merits deeper scrutiny. Instead of the ternary classification we used, a finer scale evaluation may be possible, perhaps separating the substance of the verdict from its expression and form.

When showcasing the effectiveness of the dataset for model finetuning, we used a mid-sized model (Gemma-3-12B-it), primarily for convenience. Larger models need to be investigated to further evaluate its utility.

The reasoning capabilities of the existing Arabic LLMs warrant deeper examination. Our observations of reasoning traces from open models performing test-time inference are that models often pursue incorrect reasoning paths before self-correcting based on additional information, particularly evident when answering multiple-choice questions or applying an article to a case. More thorough analysis is needed to better understand these reasoning dynamics.

Finally, an intriguing question remains regarding the underlying reasons behind the models' improved performance when prompted to reason in English, and how general this behavior is.

Ethics Statement

Legal matters are inherently sensitive and require careful handling. We have anonymized the generated dataset to remove all identifying information about plaintiffs, defendants, as well as the judges that ruled on the cases included. All contributors to this work are properly recognized, either as co-authors or in the Acknowledgments section.

References

Muhammad Abdul-Mageed, Shady Elbassuoni, Jad Doughman, AbdelRahim Elmadany, El Moatez Billah Nagoudi, Yorgo Zoughby, Ahmad Shaher, Iskander Gaba, Ahmed Helal, and Mohammed El-Razzaz. 2021. DiaLex: A benchmark for evaluating multidialectal Arabic word embeddings. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 11–20, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Wafa Alghallabi, Ritesh Thawkar, Sara Ghaboura, Ketan More, Omkar Thawakar, Hisham Cholakkal, Salman Khan, and Rao Muhammad Anwer. 2025. Fann or Flop: A multigenre, multiera benchmark for arabic poetry understanding in LLMs. *Preprint*, arXiv:2505.18152.

Ebtesam Almazrouei, Ruxandra Cojocaru, Michele Baldo, Quentin Malartic, Hamza Alobeidli, Daniele Mazzotta, Guilherme Penedo, Giulia Campesan, Mugariya Farooq, Maitha Alhammadi, Julien Launay, and Badreddine Noune. 2023. AlGhafa evaluation benchmark for Arabic language models. In *Proceedings of ArabicNLP 2023*, pages 244–275, Singapore (Hybrid). Association for Computational Linguistics.

Norah Alshahrani, Saied Alshahrani, Esma Wali, and Jeanna Matthews. 2024. Arabic synonym BERT-based adversarial examples for text classification. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 137–147, St. Julian's, Malta. Association for Computational Linguistics.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Martin Katz, and Nikolaos Aletras. 2022. LexGLUE: A benchmark dataset for legal language understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330. Association for Computational Linguistics.

Odysseas S. Chlapanis, Dimitrios Galanis, and Ion Androutsopoulos. 2024. LAR-ECHR: A new legal argument reasoning task and dataset for cases of the European court of human rights. In *Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 267–279, Miami, FL, USA. Association for Computational Linguistics.

Ali El Filali, Hamza Alobeidli, Clémentine Fourrier, Basma El Amel Boussaha, Ruxandra Cojocaru, Nathan Habib, and Hakim Hacid. 2024. Open arabic llm leaderboard. `https://huggingface.co/`

spaces/OALL/Open-Arabic-LLM-Leaderboard. Accessed 4 July 2025.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lopez de Lacalle, and Mikel Artetxe. 2024. Do multilingual language models think better in English? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 550–564, Mexico City, Mexico. Association for Computational Linguistics.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Alan Huang, Songyang Zhang, Kai Chen, Zhixin Yin, Zongwen Shen, Jidong Ge, and Vincent Ng. 2024. LawBench: Benchmarking legal knowledge of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7933–7962, Miami, Florida, USA. Association for Computational Linguistics.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2024. A survey on LLM-as-a-judge. *Preprint*, arXiv:2411.15594.

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, and 21 others. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Preprint*, arXiv:2308.11462.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

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

Faris Hijazi, Somayah Alharbi, Abdulaziz AlHussein, Harethah Shairah, Reem Alzahrani, Hebah Alshamlan, George Turkiyyah, and Omar Knio. 2024. ArabLegalEval: A multitask benchmark for assessing Arabic legal knowledge in large language models. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 225–249, Bangkok, Thailand. Association for Computational Linguistics.

Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Song Dingjie, Zhihong Chen, Mosen Alharthi, Bang An, Juncai He, Ziche Liu, Junying Chen, Jianquan Li, Benyou Wang, Lian Zhang, Ruoyu Sun, Xiang Wan, Haizhou Li, and Jinchao Xu. 2024. AceGPT, localizing large language models in Arabic. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8139–8163, Mexico City, Mexico. Association for Computational Linguistics.

King Salman Global Academy for Arabic Language. 2024. BALSAM index: Benchmark of arabic language ai systems and models. https://benchmarks.ksaa.gov.sa/b/balsam. Accessed 4 July 2025.

Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman 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. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5622–5640, Bangkok, Thailand. Association for Computational Linguistics.

Sang Kwon, Gagan Bhatia, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. Beyond English: Evaluating LLMs for Arabic grammatical error correction. In *Proceedings of ArabicNLP 2023*, pages 101–119, Singapore (Hybrid). Association for Computational Linguistics.

Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning, and Daniel E. Ho. 2025. Hallucination-free? assessing the reliability of leading ai legal research tools. *Journal of Empirical Legal Studies*, 22(2):216–242.

Basel Mousi, Nadir Durrani, Fatema Ahmad, Md. Arid Hasan, Maram Hasanain, Tameem Kabbani, Fahim Dalvi, Shammur Absar Chowdhury, and Firoj Alam. 2025. AraDiCE: Benchmarks for dialectal and cultural capabilities in LLMs. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4186–4218, Abu Dhabi, UAE. Association for Computational Linguistics.

Shubham Kumar Nigam, Aniket Deroy, Subhankar Maity, and Arnab Bhattacharya. 2024. Rethinking legal judgement prediction in a realistic scenario in the era of large language models. In *Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 61–80, Miami, FL, USA. Association for Computational Linguistics.

Shubham Kumar Nigam, Tanmay Dubey, Govind Sharma, Noel Shallum, Kripabandhu Ghosh, and Arnab Bhattacharya. 2025. LegalSeg: Unlocking the structure of Indian legal judgments through rhetorical role classification. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1129–1144, Albuquerque, New Mexico. Association for Computational Linguistics.

Joel Niklaus, Veton Matoshi, Matthias Stürmer, Ilias Chalkidis, and Daniel Ho. 2024. MultiLegalPile: A

689GB multilingual legal corpus. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15077–15094, Bangkok, Thailand. Association for Computational Linguistics.

Joel Niklaus, Lucia Zheng, Arya D. McCarthy, Christopher Hahn, Brian M Rosen, Peter Henderson, Daniel E. Ho, Garrett Honke, Percy Liang, and Christopher D Manning. 2025. LawInstruct: A resource for studying language model adaptation to the legal domain. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 127–152, Albuquerque, New Mexico. Association for Computational Linguistics.

OpenAI. 2024a. Embeddings. `https://platform.openai.com/docs/guides/embeddings`. Accessed: December 2024.

OpenAI. 2024b. GPT-4o (Omni): A Multimodal AI Model. Model announcement. Supports text, vision, audio, video; faster & cheaper than GPT□4 Turbo.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, Michael Choi, Anish Agrawal, Arnav Chopra, Adam Khoja, Ryan Kim, Richard Ren, Jason Hausenloy, Oliver Zhang, Mantas Mazeika, and 1000+ others. 2025. Humanity's last exam. *Preprint*, arXiv:2501.14249.

Zhaozhi Qian, Faroq Altam, Muhammad Alqurishi, and Riad Souissi. 2024. Cameleval: Advancing culturally aligned arabic language models and benchmarks. *Preprint*, arXiv:2409.12623.

Gerard Salton, Anawat Wong, and Chung-Shu Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Lisa Schut, Yarin Gal, and Sebastian Farquhar. 2025. Do multilingual LLMs think in english? In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*.

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, and 3 others. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *CoRR*, abs/2308.16149.

Serry Sibaee, Omer Nacar, Adel Ammar, Yasser Al-Habashi, Abdulrahman Al-Batati, and Wadii Boulila. 2025. From guidelines to practice: A new paradigm for arabic language model evaluation. *Preprint*, arXiv:2506.01920.

Gemma Team. 2025. Gemma 3. *arXiv preprint*. ArXiv:2503.19786.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. GLUE: A multi□task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *arXiv preprint*. ArXiv:2505.09388.

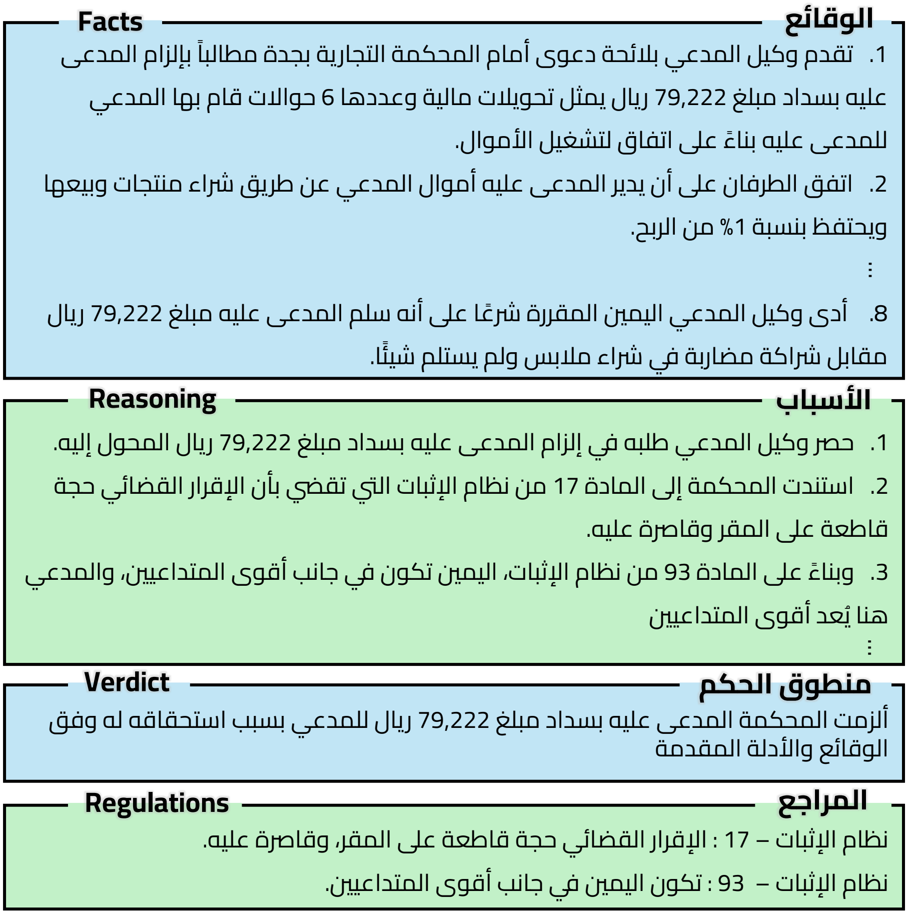
Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Lucia Zheng, Neel Guha, Javokhir Arifov, Sarah Zhang, Michal Skreta, Christopher D. Manning, Peter Henderson, and Daniel E. Ho. 2025. A reasoning-focused legal retrieval benchmark. In *Proceedings of the 2025 Symposium on Computer Science and Law*, CSLAW '25, pages 169–193, New York, NY, USA. Association for Computing Machinery.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2024. AGIEval: A human□centric benchmark for evaluating foundation models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2299–2314, Mexico City, Mexico. Association for Computational Linguistics.

A Sample Case from the Dataset

Figure 6 shows an example of the resulting structured representation of cases. To support reasoning tasks, each legal case is structured into: (1) a list of individual facts and arguments presented to the court; (2) a sequence of steps articulating the reasoning of the court; (3) the final verdict reflecting the court's opinion; and (4) the individual articles form the statutes explicitly cited in the case. The cases reference a core set of eight statutes and regulatory documents. Shown in the figure are the (standardized) keys to full text descriptions of statute articles. For convenience, these descriptions have been inserted in the output so every case has the complete reasoning context.



Facts الوقائع

1. تقدم وكيل المدعي بلائحة دعوى أمام المحكمة التجارية بجدة مطالباً بإلزام المدعى عليه بسداد مبلغ 79,222 ريال يمثل تحويلات مالية وعددها 6 حوالات قام بها المدعي للمدعى عليه بناءً على اتفاق لتشغيل الأموال.
2. اتفق الطرفان على أن يدير المدعى عليه أموال المدعي عن طريق شراء منتجات وبيعها ويحتفظ بنسبة 1% من الربح.

⋮

8. أدى وكيل المدعي اليمين المقررة شرعًا على أنه سلم المدعى عليه مبلغ 79,222 ريال مقابل شراكة مضاربة في شراء ملابس ولم يستلم شيئًا.

Reasoning الأسباب

1. حصر وكيل المدعي طلبه في إلزام المدعى عليه بسداد مبلغ 79,222 ريال المحول إليه.
2. استندت المحكمة إلى المادة 17 من نظام الإثبات التي تقضي بأن الإقرار القضائي حجة قاطعة على المقر وقاصرة عليه.
3. وبناءً على المادة 93 من نظام الإثبات، اليمين تكون في جانب أقوى المتداعيين، والمدعي هنا يُعد أقوى المتداعيين

⋮

Verdict منطوق الحكم

ألزمت المحكمة المدعى عليه بسداد مبلغ 79,222 ريال للمدعي بسبب استحقاقه له وفق الوقائع والأدلة المقدمة

Regulations المراجع

نظام الإثبات – 17 : الإقرار القضائي حجة قاطعة على المقر، وقاصرة عليه.

نظام الإثبات – 93 : تكون اليمين في جانب أقوى المتداعيين.

Figure 6: **Cases Example:** Sample legal case after restructuring.

B Prompts for Inference and Evaluation

B.1 LLM as a Judge

You are a legal assistant. You will be given a judge's verdict from a legal case in Saudi Arabia, and a prediction of the verdict from another legal assistant.
Your task is to evaluate how well the prediction matches the judge's verdict.
The evaluation should be based on the content of the verdicts and how well they align with each other.
A prediction is correct if it is similar to the judge's verdict and captures the essence of the decision. It does not have to be identical, but it should reflect the same outcome and reasoning.
It's acceptable for the prediction to be shorter or more concise than the judge's verdict, or the other way around, as long as the core message is the same. Ignore any noise or irrelevant tokens in the verdicts.
Before you output your evaluation, think about how well the prediction matches the judge's verdict.

Output one of the following for the evaluation:
- `"CORRECT"` if the prediction matches the judge's verdict.
- `"INCORRECT"` if the prediction does not match the judge's verdict.
- `"PARTIALLY CORRECT"` if the prediction is partially correct but does not fully match the judge's verdict.

Follow this format:

```
[THINK]
"Your reasoning here"
[EVALUATION]
"Evaluation here (CORRECT, INCORRECT, or PARTIALLY CORRECT)"
```

Judge's verdict:
`{judge_verdict}`

Predicted verdict:
`{predicted_verdict}`

Begin!

Figure 7: **LLM as a Judge Prompt**: The prompt we use for automatic evaluation of verdicts, provide the predicted and court verdicts to the LLM and ask to think before giving an evaluation.

B.2 Verdict Prediction

```
You are a legal assistant specialized in Saudi Arabian law. Your task is to predict the verdict of a legal
case from Saudi Arabia.
The cases involve trade and finance and commercial laws.

You will be given a set of facts from the case, and you MUST provide BOTH:
1. A reasoning section analyzing the facts
2. A verdict prediction section stating what you think the court will decide

The verdict should be based only on the facts provided without personal opinions or biases.
Think carefully about the facts and how they relate to the laws in Saudi Arabia.
Your verdict and reasoning should be strictly in {language}
The verdict should be short and direct.

Follow the format below:
[REASONING]
”Your reasoning and analysis here”
[\REASONING]

[VERDICT]
”Your verdict here”
[\VERDICT]

Do not output anything else outside these two sections.

Here are the facts of the case:
{case_facts}

Begin!
```

Figure 8: **Prompt for Verdict Prediction from Case Facts.**

You are a legal assistant. Your task is to predict the verdict of a legal case from Saudi Arabia.
The cases involve trade and finance and commercial laws.
You will be given a set of facts from the case, and the reasoning of court on these facts.
You should provide a verdict based on the facts and the reasoning of the court.
The verdict is a sentence that summarizes the outcome of the case showing what do you think the court will decide.
The verdict should be based on the facts and reasoning provided and should not include any personal opinions or biases.
Your verdict should be strictly in {language}.
Your output should only be a direct and short verdict, do not output anything else.
Make sure to label the start and end of the verdict properly.

Follow the format below:
[VERDICT]
”Your verdict here”
[\VERDICT]

Do not output anything else.

Here are the facts of the case:
{case_facts}

Here is the reasoning of the court:
{case_reasoning}

Begin

Figure 9: **Prompt for Verdict Prediction from Case Facts and Reasoning of the Court.**

```
You are a legal assistant. Your task is to predict the verdict of a legal case from Saudi Arabia.
The cases involve trade and finance and commercial laws.
You will be given a set of facts from the case, and the laws and regulations applicable to this case, and you MUST provide BOTH:
1. A reasoning section analyzing the facts
2. A verdict prediction section stating what you think the court will decide

You should provide a verdict based on the facts and the given laws.
The verdict is a sentence that summarizes the outcome of the case showing what do you think the court will decide.
The verdict should be based on the facts and laws provided and should not include any personal opinions or biases.
Your verdict and reasoning should be strictly in language.
Think about the case facts and how they relate to the given laws.

Follow the format below:
[REASONING]
”Your reasoning and analysis here”
[\REASONING]

[VERDICT]
”Your verdict here”
[\VERDICT]

Do not output anything else.

Here are the facts of the case:
{case_facts}

Here are the laws related to this case:
{case_laws}

Begin!
```

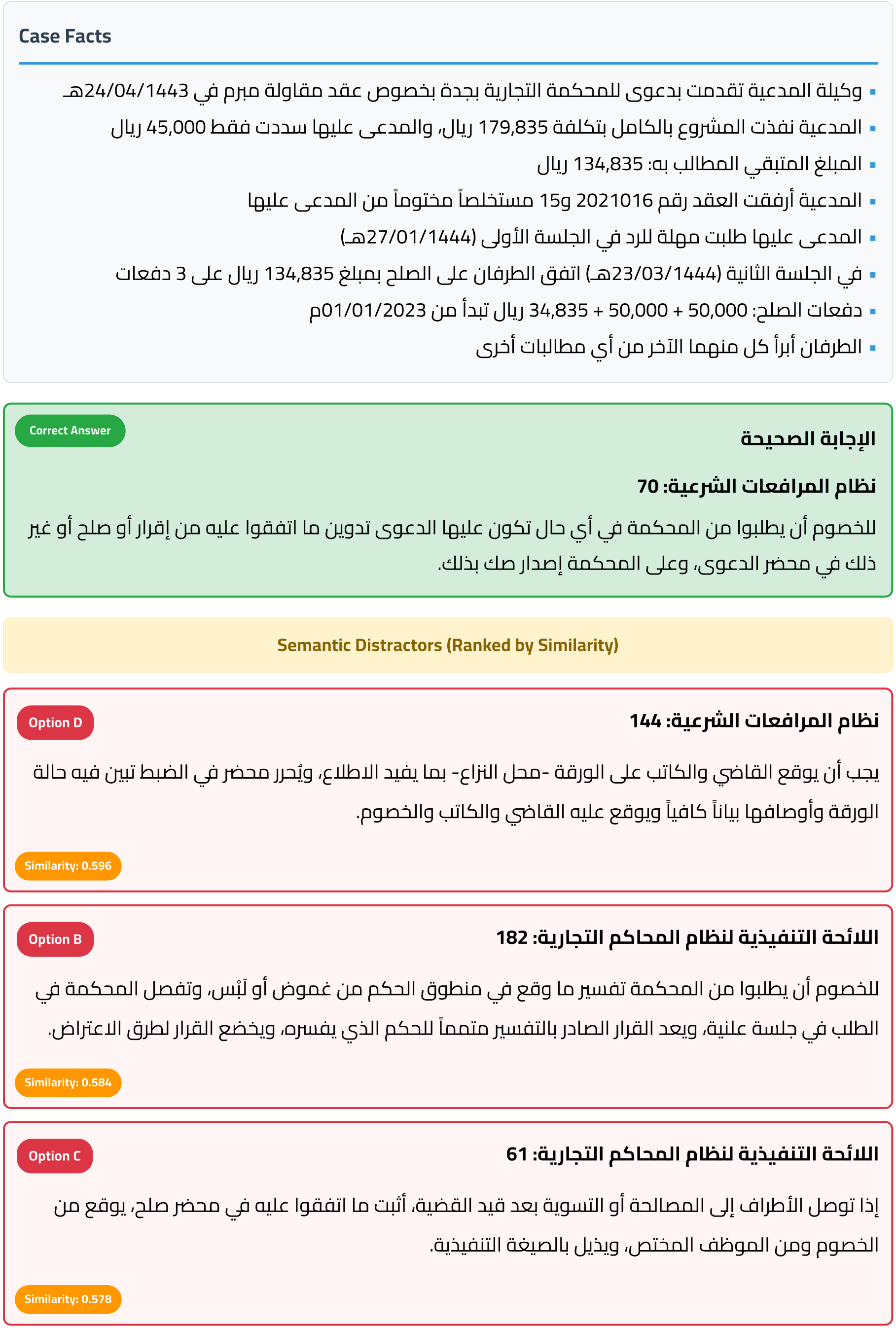
Figure 10: **Prompt for Verdict Prediction from Case Facts and Applicable Regulations.**

C SFT Instructions

Task	Instruction
Verdict Prediction	What is the court’s decision for the following case? Given the information, how should the court rule, and why? Based on the facts and reasoning, what is the final verdict of the court? Analyze the case details and provide the court’s verdict. Given the facts and reasoning, what is the court’s decision?
Reasoning & Verdict Prediction	Given the following facts and laws, provide the verdict. Read the facts and applicable laws below, then summarize the court’s decision. Given the case details, generate a summary of the reasoning and the final verdict. Analyze the following facts and laws, then provide your reasoning and the verdict. What is the court’s decision for the following case? Include reasoning. After reviewing the facts and applicable laws, explain the court’s reasoning process and final decision.
Verdict Justification (Reasoning)	Given the facts, laws, and final verdict, explain the legal reasoning of the court step by step. Analyze the case details and provide a detailed explanation of the court’s reasoning leading to the verdict. Explain the court’s reasoning process based on the provided facts, laws, and final verdict. Given the case facts and laws, summarize the court’s reasoning and how it led to the final verdict.

Table 8: **Categories of Instructions for SFT.**

D Sample MCQ from Article Identification Task



Case Facts

- وكيلة المدعية تقدمت بدعوى للمحكمة التجارية بجدة بخصوص عقد مقاولة مبرم في 24/04/1443هـ
- المدعية نفذت المشروع بالكامل بتكلفة 179,835 ريال، والمدعى عليها سددت فقط 45,000 ريال
- المبلغ المتبقي المطالب به: 134,835 ريال
- المدعية أرفقت العقد رقم 2021016 و15 مستخلصاً مختوماً من المدعى عليها
- المدعى عليها طلبت مهلة للرد في الجلسة الأولى (27/01/1444هـ)
- في الجلسة الثانية (23/03/1444هـ) اتفق الطرفان على الصلح بمبلغ 134,835 ريال على 3 دفعات
- دفعات الصلح: 50,000 + 50,000 + 34,835 ريال تبدأ من 01/01/2023م
- الطرفان أبرأ كل منهما الآخر من أي مطالبات أخرى

Correct Answer

الإجابة الصحيحة

نظام المرافعات الشرعية: 70

للخصوم أن يطلبوا من المحكمة في أي حال تكون عليها الدعوى تدوين ما اتفقوا عليه من إقرار أو صلح أو غير ذلك في محضر الدعوى، وعلى المحكمة إصدار صك بذلك.

Semantic Distractors (Ranked by Similarity)

Option D

نظام المرافعات الشرعية: 144

يجب أن يوقع القاضي والكاتب على الورقة -محل النزاع- بما يفيد الاطلاع، ويُحرر محضر في الضبط تبين فيه حالة الورقة وأوصافها بياناً كافياً ويوقع عليه القاضي والكاتب والخصوم.

Similarity: 0.596

Option B

اللائحة التنفيذية لنظام المحاكم التجارية: 182

للخصوم أن يطلبوا من المحكمة تفسير ما وقع في منطوق الحكم من غموض أو لَبْس، وتفصل المحكمة في الطلب في جلسة علنية، ويعد القرار الصادر بالتفسير متمماً للحكم الذي يفسره، ويخضع القرار لطرق الاعتراض.

Similarity: 0.584

Option C

اللائحة التنفيذية لنظام المحاكم التجارية: 61

إذا توصل الأطراف إلى المصالحة أو التسوية بعد قيد القضية، أثبت ما اتفقوا عليه في محضر صلح، يوقع من الخصوم ومن الموظف المختص، ويذيل بالصيغة التنفيذية.

Similarity: 0.578

Figure 11: Sample MCQ showing semantically similar distractors