MizanQA: A Benchmark for Multi-Answer Moroccan Legal QA

Adil Bahaj
Mohammed 6 Polytechnic University

Mounir Ghogho
Mohammed 6 Polytechnic University

Abstract

We present MizanQA, a benchmark for assessing LLMs on Moroccan legal MCQs, many with multiple correct answers. Covering 1,776 expert-verified questions in Modern Standard Arabic enriched with Moroccan idioms, the dataset reflects influences from Maliki jurisprudence, customary law, and French legal traditions. Unlike single-answer settings, MizanQA features variable option counts, creating added difficulty. We evaluate multilingual and Arabic-centric models in zero-shot, native-Arabic prompts, measuring accuracy, a precision-penalized F1-like score, and calibration errors. Results show large performance gaps and miscalibration, particularly under stricter penalties. By scoping this benchmark to parametric knowledge only, we provide a baseline for future retrieval-augmented and rationale-focused setups.

1 Introduction

Large language models (LLMs) have driven major advances in natural language understanding and generation, yet their effectiveness in specialized domains such as legal contexts—especially in low- and medium-resource languages like Arabic—remains an open research challenge. This paper investigates LLMs' ability to comprehend and process Arabic legal corpora within the Moroccan legal system.

Moroccan legal language intensifies the difficulties Arabic already poses for LLMs (Bayan Kmainasi et al., 2025; Daoud et al., 2025). Although written in Modern Standard Arabic, Moroccan law is permeated with local idioms and cultural references. It reflects a blend of Islamic Maliki jurisprudence, customary law, and French/international influences, which introduces "cultural specificities inherent to legal terminology" (Ismail Mellouki, 2021). As a result, statutes often use archaic or region-specific expressions absent from standard Arabic corpora. For NLP systems, this mix of formal syntax and Morocco-specific terminology creates major challenges, making accurate legal QA dependent on handling precise phrasing while recognizing concepts unique to Morocco's legal system.

We introduce **MizanQA**, a benchmark for evaluating LLMs on Moroccan legal question answering. It contains over 1,700 MCQ pairs spanning basic legal knowledge to detailed reasoning in various legal categories. A key feature is the presence of multi-answer questions, which increase task difficulty beyond standard single-answer formats.

In summary, this paper makes the following key contributions:

- A curated Arabic MCQ benchmark[1] for Moroccan law with multi-answer items and variable option counts.
- Clearer evaluation criteria for multi-answer MCQ: strict accuracy, precision-penalised F1-like, and ECE variants (per-option, set-level).
- Zero-shot, native-Arabic evaluation of multilingual and Arabic-centric LLMs, revealing accuracy and calibration gaps.
- A parametric-knowledge baseline (no retrieval), to be complemented by RAG and reasoning tracks in future work.

2 Related Work

The success of multilingual LLMs (e.g., GPT (OpenAI et al., 2024), Gemini (Yang et al., 2024; Team et al., 2023)) has led to native Arabic models such as ALLAM (Bari et al., 2024) and JAIS (Sengupta et al., 2023), yet these still show domain-specific knowledge gaps (Bayan Kmainasi et al., 2025; Daoud et al., 2025). Existing legal benchmarks are mostly English-focused (Fei et al., 2024; Hijazi et al., 2024; Guha et al., 2023; Pipitone and Alami, 2024; Li et al., 2024; Dahl et al., 2024), with

[1] https://huggingface.co/datasets/adlbh/MizanQA-v0

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 5: Industry Track), pages 132–144
March 25-27, 2026

only limited coverage in Chinese (Fei et al., 2024; Li et al., 2024) and Saudi Arabic (Hijazi et al., 2024). To date, just one Arabic legal benchmark exists (Hijazi et al., 2024), largely based on translated content and Saudi law. This work introduces the first Moroccan legal QA dataset, capturing its unique linguistic and cultural complexity. Unlike prior benchmarks with only single-answer MCQs, Moroccan legal exams often require multiple correct answers from variable option sets, motivating new evaluation metrics for this setting.

3 MizanQA Dataset

3.1 General Description

MizanQA is constructed from publicly available Moroccan law MCQ banks and exams. The dataset contains 1,776 questions, option counts range 2–12, and correct-answer counts 1–10 across 9 law categories. Table 1 summarises different statistics of MizanQA. The dataset contains a varying number of options and correct answers, which increases the complexity of the benchmark. Table 2 lists the number of questions per legal topic category. Table 3 gives an example of a question present in MizanQA. Figure 1 shows the distribution of the number of options per question in the dataset. Figure 2 shows the distribution of the number of correct options in the dataset.

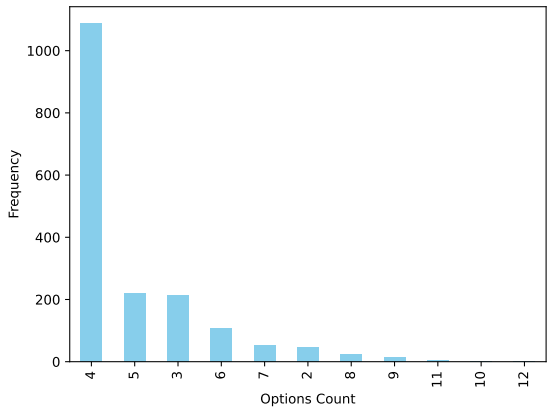

Figure 1: Distribution of the number of options in the dataset.

3.2 Construction Process

The dataset's construction process went through multiple phases, with hybrid manual and automated steps.

- **Step 1: Collection.** We collected a set of publicly available Moroccan-law MCQ sources.
- **Step 2: Temporal curation** A legal expert curated the collected documents to sift out any documents that use outdated legislation.

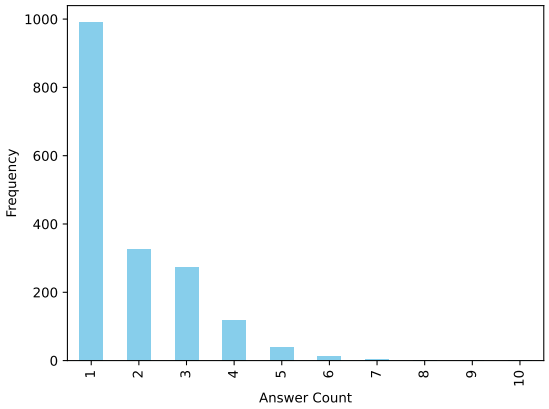

Figure 2: Distribution of the number of correct options in the dataset.

- **Step 3: Organisation.** MCQs were grouped into image batches to enable automated extraction. For structured documents with consistent formatting, this was automated, while irregular documents (e.g., spanning pages or with answers at the end) required manual organization. In these cases, annotators captured screenshots of complete question–option–answer sets, ensuring each page was self-contained before conversion to images.
- **Step 4: Extraction.** The images containing batches of MCQs produced in the previous step are fed to a multimodal LLM (i.e. Gemini-2.0-Flash in our case) to extract MCQs in a standardised format.
- **Step 5: Verification.** The extracted MCQs in the previous step are verified manually. The curators follow a set of verification guidelines (appendix A.5) to ensure that the extracted questions are identical to the original ones.
- **Step 6: Categorisation.** Depending on the original documents, MCQs are categorised manually based on the set of legislation they represent (e.g. Criminal law, constitution, etc). This is followed by normalisation of the categories to remove any redundancy.

4 Benchmarking Study

4.1 Evaluation metrics

Accuracy Measures We found that most MCQs from Moroccan sources have multiple options. An answer is considered correct only if all the right options are chosen. To our knowledge, prior QA benchmarks do not target multi-answer Arabic MCQs with variable option counts. Consequently, we created different performance met-

Statistic	Values
Number of questions	1776
Number of categories	9
Number of options per question	min: 2, max: 12
Number of words per question	min: 1, max: 63
Number of correct options per question	min: 1, max: 10
Number of words per option	min: 1, max: 71

Table 1: General statistics of MizanQA. min and max signify the range of values that a statistic has in the MizanQA.

Category (EN)	Category (AR)	Count
Civil Procedure	المسطرة المدنية	460
Criminal Law	القانون الجنائي	847
Exam	الامتحانات	131
Family Code	مدونة الأسرة	38
Family Law	المادة الأسرية	66
Law of Obligations and Contracts	قانون الالتزامات والعقود	37
The Judicial System of the Kingdom	التنظيم القضائي للمملكة	88
The Justice Sector	قطاع العدل	39
The Moroccan Constitution	الدستور المغربي	70

Table 2: Distribution of topic categories in MizanQA.

	Arabic	English Translation
Question	إذا نسب لباشا أو خليفة أول لعامل، أو رئيس دائرة أو قائد أو لضابط شرطة قضائية غير المشار إليهم سابقا، ارتكابهم لجناية أو جنحة اثناء مزاولة مهامهم، فإن	If it is alleged that a Pasha, a first deputy to a governor, a head of a department, a commander, or a judicial police officer other than those previously mentioned, has committed a felony or misdemeanor while performing their duties, then
Options	'A': 'الرئيس الأول لمحكمة الاستئناف المعروضة عليه القضية من طرف الوكيل العام للملك إذا قرر إجراء بحث فإنه يعين مستشارا مكلفا بالتحقيق بمحكمته', 'B': 'إذا تعلق الأمر بجناية فإن المستشار المكلف بالتحقيق يصدر أمرا بإحالة القضية إلى غرفة الجنايات', 'C': 'إذا تعلق الأمر بجنحة، فإنه يحيل القضية إلى محكمة ابتدائية غير التي يزاول المتهم فيها مهامه', 'D': 'يرجع الاختصاص إلى محكمة النقض إذا كان ضابط الشرطة القضائية مؤهلا لمباشرة وظيفته في مجموع تراب المملكة', 'E': 'يمكن للطرف المدني التدخل لدى هيئة الحكم', 'F': 'جميع الأجوبة صحيحة'	'A': The first president of the Court of Appeal to whom the case is referred by the Public Prosecutor, if he decides to conduct an investigation, shall appoint an advisor in charge of the investigation in his court., 'B': If it is a felony, the investigating advisor issues an order referring the case to the criminal chamber., 'C': If it is a misdemeanor, he refers the case to a court of first instance other than the one in which the accused performs his duties., 'D': Jurisdiction reverts to the Court of Cassation if the judicial police officer is qualified to perform his duties throughout the Kingdom., 'E': The civil party may intervene before the arbitral tribunal., 'F': All the answers are correct
Answer	F	F

Table 3: An example of a Question and its corresponding answer in MizanQA.

rics to evaluate LLMs on this task. Let $\mathcal{Q} = (Q_i, O_i, C_i)_i$ be the set of questions Q_i, their corresponding options O_i and the correct options C_i. Let $\mathbf{P}(Q_i, O_i)$ be a prompt parameterised by question Q_i and its corresponding options O_i and let $S_i = \text{LLM}(\mathbf{P}(Q_i, O_i))$ be the set of options predicted by an LLM to be correct for question Q_i. $S_i = \{(\hat{c}_j, p_j)\}_j$ is composed of tuples $(\hat{c}_j, p_j)$, where $\hat{C} = \{\hat{c}_j\}_j$ is the set of predicted options, $\hat{c}_j \in O_i$ is an option selected by the LLM and $p_j \in [0, 1]$ is the LLM's corresponding confidence that option j is the right option. We define strict accuracy as:

$$\text{ACC} = \frac{1}{|\mathcal{Q}|} \sum_{i}^{|\mathcal{Q}|} \mathbb{1}_{[\hat{C}_i \setminus C_i = C_i \setminus \hat{C}_i = \emptyset]} \tag{1}$$

$\mathbb{1}_{[A]}$ is the indicator function, which equals 1 if A is true and 0 otherwise. ACC rewards only perfectly correct answers. Additionally, to reward partial correctness while penalising incorrect selections, we propose a metric inspired by the F1 metric (Sitarz, 2022):

$$\text{F1-like}_\alpha = \frac{1}{|\mathcal{Q}|} \sum_{i}^{|\mathcal{Q}|} \frac{2P_i R_i}{P_i + R_i} \tag{2}$$

where $R_i = \frac{TP_i}{TP_i + FN_i}$ is equivalent to recall and $P_i = \frac{TP_i}{TP_i + \alpha \cdot FP_i}$ is equivalent to precision, such that $TP_i = |C_i \cap \hat{C}_i|$, $FP_i = |\hat{C}_i \setminus C_i|$ and $FN_i = |C_i \setminus \hat{C}_i|$ are true positives (correct answers selected), false positives (wrong answers selected) and false negatives (missed correct answers), respectively. $\alpha \geq 1$ increases the penalty for wrong choices. We also propose Partial Match Penalized Accuracy (PMPA):

$$\text{PMPA}_\beta = \frac{1}{|\mathcal{Q}|} \sum_{i}^{|\mathcal{Q}|} \max\left(0, \min\left(1, \frac{TP_i - \beta \cdot FP_i}{|C_i|}\right)\right) \tag{3}$$

where $\beta \in [0, 1]$ is a penalty factor for incorrect answers. The F1-like score and the PMPA score have a similar objective, but the PMPA score is more advantageous in cases where the number of correct options varies significantly. This is particularly important since the number of options per question in our dataset varies from 2 to 12.

Confidence calibration measures A model exhibits well-calibrated uncertainty when its predicted probabilities are congruent with observed empirical frequencies; specifically, events assigned a probability p occur with a relative frequency of p in empirical validation. Following (Naeini et al., 2015), we estimate Expected Calibration Error (ECE) by binning predicted confidences of N samples into M equally-spaced bins $B = \{B_m\}_{m=1}^{M}$ w.r.t. the prediction confidence estimated for each sample. The empirical ECE estimator is given by,

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} |\text{conf}(B_m) - \text{acc}(B_m)| \tag{4}$$

We use this measure in two settings: a) the Per-Option Calibration and b) Set-level Calibration.

- **Per-option Calibration Setting**: Let $\mathcal{D}_{\text{opt}} = \{(y_{i,j}, p_{i,j})\}$ such that i is the index of examples and j is the index of options (i.e. jth predicted option of the ith example). Let $y_{i,j} = \mathbb{1}_{[\hat{c}_{i,j} \in C_i]}$.
  - The empirical accuracy in bin B_m is:
  $$\text{acc}(B_m) = \frac{1}{|B_m|} \sum_{(y,p) \in B_m} \mathbb{1}_{[y=1]} \tag{5}$$
  - The average predicted confidence is:
  $$\text{conf}(B_m) = \frac{1}{|B_m|} \sum_{(y,p) \in B_m} p \tag{6}$$
  - Number of examples N: $N = |\mathcal{D}_{\text{opt}}|$
- **Set-level Calibration**: let $\mathcal{D}_{\text{set}} = \{(z_i, q_i)\}_i$ such that $z_i = \mathbb{1}_{[\hat{C}_i = C_i]}$ is an indicator which equals 1 if and only if the predicted set exactly matches the ground truth. Set-level confidence multiplies option confidences, implicitly assuming independence: $q_i = \prod_{(\hat{c}_j, p_j) \in S_i} p_j$. We use it as a conservative proxy for joint correctness without adding model-specific calibration tricks. After binning the pairs (z_i, q^i) the following metrics can be calculated :
  - Empirical accuracy in each bin ($\text{acc}(B_m)$):
  $$\text{acc}(B_m) = \frac{1}{|B_m|} \sum_{(z_i, q_i) \in B_m} z_i \tag{7}$$
  - Average predicted joint confidence ($\text{conf}(B_m)$):
  $$\text{conf}(B_m) = \frac{1}{|B_m|} \sum_{(z_i, q_i) \in B_m} q_i \tag{8}$$
  - Number of examples N: $N = |\mathcal{D}_{\text{set}}|$

Practically, the Per-option Calibration Setting (ECE_{opt}) and the Set-level Calibration error (ECE_{set}) are obtained by replacing their respective expressions of $\text{conf}(B_m)$, $\text{acc}(B_m)$ and N in equation 4.

Why F1-like and PMPA? Multi-answer MCQs require selecting all and only correct options, so naïve accuracy both under-rewards partial knowledge and conflates omission with commission errors. To address this, F1-like$_\alpha$ penalizes extra wrong selections more heavily, while PMPA$_\beta$ normalizes by the true set size, ensuring comparability across variable option counts and numbers of correct answers (as in Moroccan legal MCQs). Alongside ECE at both option and set levels, these metrics capture not only prediction accuracy but also confidence calibration under multi-answer uncertainty.

4.2 Baselines

We evaluated various multilingual and specialised Arabic LLMs on MizanQA. These models have varying levels of complexity (i.e. number of parameters, support for reasoning etc). We evaluated the following models: Allam-2 (7b) (Bari et al., 2024), Gemini-1.5-Flash (Yang et al., 2024; Team et al., 2023), Gemini-2.0-Flash (Yang et al., 2024; Team et al., 2023), Llama-3.3 (70b) (Grattafiori et al., 2024), Llama-4-Maverick (17b) (Team, 2025), and Llama-4-Scout (17b) (Team, 2025).

4.3 Experimental Setting

All models are evaluated zero-shot in native Arabic script with a fixed prompt (English translation in Fig. 3), requiring outputs as option letters with per-option confidence. Responses are parsed, malformed outputs are re-prompted, and failures are marked incorrect. Experiments use temperature = 1, with no tool use or retrieval (details in Appx. B). We deliberately exclude retrieval-augmented settings to isolate models' parametric legal knowledge in Moroccan Arabic. This avoids confounding retrieval with reasoning, ensures comparability to prior legal QA benchmarks, and establishes a baseline for future retrieval-augmented extensions.

4.4 Results

Table 1 summarises the overall results. Gemini-2.0-Flash leads ACC and PMPA, and is best on F1-like(2) (higher penalty on extra selections). Llama-4-Maverick narrowly tops F1-like(1) and exhibits the lowest ECE at both option and set levels, indicating more conservative confidence allocation. Performance declines as penalty strength increases. Results confirm substantial gaps and miscalibration, especially under stricter penalties.

4.4.1 Performance vs. category

Appendix B.1 shows that LLM performance generally improves from Allam-2 (7b) to Gemini-2.0-Flash, with Gemini models outperforming the Llama series. Accuracy is higher in the Law of Obligations and Contracts and the Moroccan Constitution, likely due to alignment with international legal standards, while lower scores in the Family Code and Criminal Law reflect challenges tied to Islamic jurisprudence and human rights frameworks. Calibration errors vary across models and categories, revealing inconsistencies between confidence and predictive accuracy.

4.4.2 Performance vs. Options Count

Appendix B.2 shows that performance declines as the number of options increases: accuracy and selection-sensitive metrics (F1-like, PMPA) drop, while calibration errors rise at both option and set levels. The steepest losses occur in F1-like(2) and PMPA(1), with ACC falling more gradually and

Prompt(EN)

- You have been given a question about Moroccan law.
- Answer the question by choosing the correct option indicator.
- You can choose multiple options that you think are correct.
- Make sure to choose only the correct options or you will be penalized.
-Give your confidence score from 1 to 100 for each option you choose.
- Your output must be in the following format only [("Confidence Score", "Option 1"), ("Confidence Score", "Option 2")...].
Question:
<QUESTION>
Options:
<OPTIONS>
#Answer:

Figure 3: English translation of instructions used to prompt various LLMs to answer MizanQA questions.

Model	PM(1) ↑	PM(0.5)↑	F1(1) ↑	F1(2) ↑
Allam-2 (7b)	26.88	34.04	43.07	39.93
Gemini-1.5-Flash	35.90	44.23	53.30	48.93
Gemini-2.0-Flash	**53.57**	**58.34**	**64.84**	**62.16**
Llama-3.3 (70b)	46.78	50.73	59.21	56.18
Llama-4-Maverick (17b)	49.97	55.53	**64.90**	61.29
Llama-4-Scout (17b)	44.06	49.01	59.51	55.60

(a) F1(α) refers to the F1-Like metric in equation 2. and PM(β) refers to the measure in equation 3.

Model	ACC ↑	ECE_{opt} ↓	ECE_{set} ↓
Allam-2 (7b)	15.32	28.42	51.43
Gemini-1.5-Flash	24.26	34.77	48.52
Gemini-2.0-Flash	**42.11**	28.15	41.16
Llama-3.3 (70b)	33.28	35.27	59.40
Llama-4-Maverick (17b)	36.83	**17.64**	**29.10**
Llama-4-Scout (17b)	31.27	36.99	61.78

(b) ACC refers to equation 1; ECE_{opt} and ECE_{set} refer the options and set variants of equation 4 respectively.

Table 4: Evaluation results of various models on MizanQA.

ECE_{set} growing faster than ECE_{opt} due to compounding uncertainty. While model rankings remain stable, performance gaps widen at high option counts, underscoring choice-set size as a key challenge and the importance of selection-aware metrics and set-level calibration.

5 Conclusion

This paper introduces MizanQA, the first benchmark for evaluating LLMs on Moroccan legal question answering. The dataset comprises 1,776 expert-validated MCQs from authentic legal texts, including many multi-answer items with variable option counts that reflect the linguistic and conceptual complexity of Moroccan law. Initial results indicate baseline competence but persistent gaps—especially as choice sets grow; by scoping this benchmark to parametric knowledge only (no retrieval), we establish a clear foundation for future retrieval-augmented and rationale-focused tracks.

6 Real world Impact

Morocco is home to a population of over 37 million and a vibrant multilingual legal ecosystem, yet many citizens—especially in rural areas, among Amazigh-speaking communities, or in economically disadvantaged settings, face acute barriers when it comes to accessing and understanding legal knowledge. Despite recent reforms and laws promising transparency, implementation remains patchy and information often remains inaccessible. In recent years, the rapid emergence of legal-technology platforms (such as Juridia) has been reshaping access to justice and legal services. Despite this progress, there remains a striking absence of publicly-available benchmark datasets aligned with the domestic legal context (Arabic, French, Moroccan regulatory and case-law mix) that industrial systems can use for rigorous evaluation, model comparison and continuous improvement. Our proposed dataset fills this gap by offering domain-specific, openly reusable data tuned to Morocco's legal ecosystem, thereby enabling legal-tech developers, law firms and regulators to benchmark model performance, identify bias or errors. In doing so, it supports the deployment of robust, scalable NLP systems in real-world industrial settings.

Limitations

This work is a domain-specific first step in Arabic legal evaluation, focused on Moroccan law. Limitations include: (i) coverage bias from a finite set of law categories and imbalance across them; (ii) limited real-world complexity, as even reasoning-based, multi-answer items can oversimplify legal interpretation; and (iii) reliance on MCQs, which do not fully capture professional reasoning. Following prior work (Guha et al., 2023; Fei et al., 2024), MizanQA is scoped to parametric knowledge only (no retrieval, prompt engineering, or tool use) to isolate memorized legal-term understanding and provide a clean baseline for future retrieval-augmented and rationale-required tracks.

References

M Saiful Bari, Yazeed Alnumay, Norah A Alzahrani, Nouf M Alotaibi, Hisham A Alyahya, Sultan AlRashed, Faisal A Mirza, Shaykhah Z Alsubaie, Hassan A Alahmed, Ghadah Alabduljabbar, and 1 others. 2024. Allam: Large language models for arabic and english. *arXiv preprint arXiv:2407.15390*.

Mohamed Bayan Kmainasi, Ali Ezzat Shahroor, and Amani Al-Ghraibah. 2025. Can large language models predict the outcome of judicial decisions? *arXiv e-prints*, pages arXiv–2501.

Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. 2024. Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis*, 16(1):64–93.

Mouath Abu Daoud, Chaimae Abouzahir, Leen Kharouf, Walid Al-Eisawi, Nizar Habash, and Farah E Shamout. 2025. Medarabiq: Benchmarking large language models on arabic medical tasks. *arXiv preprint arXiv:2505.03427*.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Alan Huang, Songyang Zhang, Kai Chen, Zhixin Yin, Zongwen Shen, and 1 others. 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.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, and 1 others. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*, 36:44123–44279.

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.

Chakib Lebaidi Ismail Mellouki. 2021. Issues of equivalence in the moroccan legal text. *Journal of University Studies for Inclusive Research*, pages 1456–1478.

Haitao Li, Junjie Chen, Jingli Yang, Qingyao Ai, Wei Jia, Youfeng Liu, Kai Lin, Yueyue Wu, Guozhi Yuan, Yiran Hu, and 1 others. 2024. Legalagentbench: Evaluating llm agents in legal domain. *arXiv preprint arXiv:2412.17259*.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Nicholas Pipitone and Ghita Houir Alami. 2024. Legalbench-rag: A benchmark for retrieval-augmented generation in the legal domain. *arXiv preprint arXiv:2408.10343*.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, William Marshall, Gurpreet Gosal, Cynthia Liu, Zhiming Chen, and 1 others. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *arXiv preprint arXiv:2308.16149*.

Mikolaj Sitarz. 2022. Extending f1 metric, probabilistic approach. *arXiv preprint arXiv:2210.11997*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Meta Llama Team. 2025. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation — ai.meta.com. https://ai.meta.com/blog/llama-4-multimodal-intelligence/. [Accessed 05-05-2025].

Lin Yang, Shawn Xu, Andrew Sellergren, Timo Kohlberger, Yuchen Zhou, Ira Ktena, Atilla Kiraly, Faruk Ahmed, Farhad Hormozdiari, Tiam Jaroensri, and 1 others. 2024. Advancing multimodal medical capabilities of gemini. *arXiv preprint arXiv:2405.03162*.

Ethics Statement

This work presents MizanQA, a research-oriented legal QA benchmark based on Moroccan law, constructed from official public-domain sources while excluding sensitive data. One legal expert and four researchers (PhD/postdoc) volunteered to verify MCQs and correct answers (Appendix D). Verification guidelines required content-faithful transcriptions, option-order checks, and answer parity with the source. No personal data were used. We include license, compensation (volunteer), and conflict-resolution procedures in Appendix D.

A Construction process

The construction process of MizanQA is semi-automated. It is composed out of multiple steps, some of which are automated while others require human intervention. We observed that a significant number of documents are based on outdated legislation; consequently, to remove these documents, **Step 2** was included. The motivation behind **steps 3** and **4** is the problems faced by annotators when copying and pasting Arabic text from PDFs. The vast majority of documents, when copied and pasted, produce unreadable information. Consequently, optical character recognition (OCR) was essential to automate the extraction. Although the automated extraction is highly accurate, the LLM produces some mistakes (e.g. not listing all the right answers, etc). To eliminate these issues **step 5** is conducted for manual verification. In the last step, MCQs are categorised depending on the original documents from which they were extracted, and the categories are normalised to remove any redundancies made by the annotators. In what follows, we give more details about the construction process.

A.1 Step 1: Collection

The data is collected from a plethora of documents that are generally PDFs or Word documents. The MCQs are structured in various formats inside the documents: single MCQ per page (Figure 4), multiple MCQ per page (Figure 5), etc.

Figure 4: An example of a document page.

Figure 5: An example of a document page.

A.2 Step 2: Temporal curation

The raw documents are given to a legal expert to evaluate the recency of the legislation that they are based on. The documents based on outdated legislation are not considered for further processing.

A.3 Step 3: Organisation

The chosen documents are then either manually or automatically transformed into images containing batches of MCQs. The automatic process takes advantage of the structured nature of some documents to gather them in batches. On the other hand, for more irregular documents (e.g. pages contain a varying number of MCQs, MCQs that are not completely expressed on the same page, the answers for MCQs are in a separate page, etc). In this case, the MCQs are screened one by one manually and concatenated into a document. The pages of the documents are then turned into images. Figure 6 shows an example of these images.

A.4 Step 4: Extraction

After organising the MCQs to images, where each image contains a batch of MCQs, the images are fed to a vision LLM (Gemini-2.0-Flash) to structure the MCQs in a machine-readable format automatically. Figure 7 represents the prompt used to extract MCQs. The extraction LLM is used only to transcribe/structure public-domain MCQs; all items/answers were manually verified (A.5), and extraction does not influence evaluation.

A.5 Step 5: Verification

In this step, the MCQs are manually verified by annotators. The verification step follows the following guidelines:

Figure 6: A batch of MCQs concatenated one below the other.

QA Pairs extraction from images

#Instructions: - This is a list of multiple-choice questions in Arabic.
- Extract the different MCQS in the following format:

```
[

"question": "",
"options":
"A": "",
"B": "",
"C": "",
"D": ""
,
"answer": "option index letter",
"hint":"",
"source": ""
,
...
]
```

 # Response:

Figure 7: Prompt for extracting MCQs from the organised images of MCQs obtained in step 3.

- Check if the question is identical to the original question.
- Check if the options are correct.
- Check if the order of options is the same.
- Check if the answers are identical to the source answers.

A.6 Step 6: Categorisation

The annotators are tasked to use the original documents from which the MCQs are extracted to categorise the different law texts that they are based on (e.g. Criminal Law, Constitution, etc.). These categories are explored and normalised to remove any redundancy.

B Benchmarking

MizanQA is tested on many multilingual and Arabic language models to assess their knowledge of Moroccan law. Figure 8 shows the prompt for prompting the different LLMs. 9 gives an English translation of the prompt. We map reported confidences to [0,1] by dividing by 100.

B.1 Performance vs. Law Category

Table 5 summarises the results of the different models by law category. The models are assessed across several Moroccan law categories: Civil Procedure, Criminal Law, Family Code, Family Law, Law of Obligations and Contracts, The Judicial System of the Kingdom, The Justice Sector, and The Moroccan Constitution. Per-category analyses exclude the 'Exam' bucket (mixture of topics) to isolate category effects. Across the models, there is a general trend of improvement in performance

Prompt (AR)

- لقد تم إعطاؤك سؤال حول القانون المغربي.
- أجب عن السؤال باختيار مؤشر الخيارات الصحيح.
- يمكنك اختيار خيارات متعددة تعتقد أنها صحيحة.
- تأكد من اختيار الخيارات الصحيحة فقط وإلا ستتعرض للعقوبة.
- أعطي درجة ثقتك من ١ إلى ١٠٠ في كل خيار تختاره.
- جب أن يكون الناتج الخاص بك بالتنسيق التالي فقط لا غير «(درجة الثقة، الخيار ١)، (درجة الثقة، الخيار ٢) ...».

سؤال:
<QUESTION>
خيارات:
<OPTIONS>
#إجابة:

Figure 8: Instructions used to prompt various LLMs to answer MizanQA questions.

Prompt (EN)

- You have been given a question about Moroccan law.
- Answer the question by choosing the correct option indicator.
- You can choose multiple options that you think are correct.
- Make sure to choose only the correct options or you will be penalized.
-Give your confidence score from 1 to 100 for each option you choose.
- Your output must be in the following format only [("Confidence Score", "Option 1"), ("Confidence Score", "Option 2")...].

Question:
<QUESTION>
Options:
<OPTIONS>
#Answer:

Figure 9: English translation of instructions used to prompt various LLMs to answer MizanQA questions.

from Allam-2 (7b) to Gemini-2.0-Flash, with the Gemini models generally outperforming the Llama models. For specific law categories, Law of Obligations and Contracts and the Moroccan Constitution tend to have higher scores across most metrics and models, indicating that these areas may be easier for the LLMs to handle. This may reflect greater alignment with internationally standardised concepts and terminology. Conversely, Family Code and Criminal Law often exhibit lower performance scores, suggesting these domains pose a greater challenge. These domains combine Islamic jurisprudence with modern human-rights norms, increasing doctrinal complexity. The calibration errors (ECE_{opt} and ECE_{set}) vary across models and categories, with no clear pattern of consistency, indicating differences in the models' confidence and accuracy alignment.

B.2 Performance vs. Number of options

In addition, figures 10, 11,12, and 13 represent the stratified results by the number of answer options (2–12) for Gemini-2.0-Flash, Gemini-1.5-Flash, Llama-3.3 (70b) and Llama-4-Maverick (17b) respectively. We report ACC, F1-like(1/2), PMPA(1/0.5), and calibration (ECE_{opt}, ECE_{set}) per bin. All metrics are shown on a 0–100 scale; ECE values are plotted as percentages. Across Gemini-2.0-Flash, Gemini-1.5-Flash, Llama-3.3-70B, and Llama-4-Maverick-17B, we observe the same qualitative pattern: as the number of options per question increases, accuracy and the selection-sensitive metrics (F1-like and PMPA) decrease, while calibration errors—both option-level ECE and set-level ECE—increase. The decline is most pronounced for F1-like(2) and PMPA(1), which penalise extra selections more heavily; ACC falls more gently, reflecting its insensitivity to partial credit. Set-level calibration (ECE_{set}) grows faster than option-level (ECE_{opt}), consistent with compounding uncertainty when models distribute probability mass over longer option lists. Collectively, these curves indicate rising over-selection risk and worsening confidence alignment as choice sets grow.

The relative ranking of models on top-line metrics largely persists across option-count bins, but gaps widen at high option counts, where selection penalties and joint-confidence calibration matter most. This analysis pinpoints choice-set size as a dataset-level difficulty factor and clarifies why selection-aware metrics and set-level calibration are essential for multi-answer legal MCQ.

C Technical setup

All the experiments are conducted using either the Groq API or the Gemini API. All the models are incorporated in Groq except Gemini-2.0-Flash and Gemini-1.5-Flash. We use Python to access the APIs, prompt the models, process and save their outputs.

D Annotators

This dataset was annotated by volunteers. The group of volunteers contained one legal expert, three PhD students and one postdoctoral student, supervised by a professor. These participants agreed to volunteer for free due to the importance of the dataset in the assessment of legal knowledge in LLMs, which is a first step towards democratising access to legal support in Morocco. These annotators belong to a diverse set of demographic and socioeconomic backgrounds. Dataset license: CC BY-NC-SA 4.0; source texts are public-domain official materials.

E Use of AI

AI has been used in the extraction process. It was also evaluated using our dataset. During the writing of the paper, it was used for editing and grammar and style correction.

Model	Category	PMPA(1)	PMPA(0.5)	F1-Like(1)	F1-Like(2)	ACC	ECE_{opt}	ECE_{set}
Allam-2 (7b)	Civil Procedure	27.70	35.34	46.28	42.98	10.87	21.09	52.88
	Criminal Law	26.73	32.90	40.94	38.00	17.95	33.02	50.62
	Family Code	20.61	25.00	33.60	31.62	7.89	37.83	67.03
	Family Law	31.69	39.71	50.13	47.33	13.64	27.36	56.62
	Law of Obligations and Contracts	31.08	38.51	46.76	44.14	18.92	19.45	52.75
	The Judicial System of the Kingdom	17.61	27.46	36.66	32.90	6.82	27.61	48.98
	The Justice Sector	27.35	35.68	47.48	43.13	17.95	35.02	57.66
	The Moroccan Constitution	41.67	54.64	64.40	59.69	28.57	24.62	40.50
Gemini-1.5-Flash	Civil Procedure	40.50	50.66	61.79	56.72	25.85	19.02	43.48
	Criminal Law	29.55	35.99	44.19	40.48	19.45	47.02	53.85
	Family Code	48.68	54.61	63.51	59.52	34.21	22.21	51.46
	Family Law	39.07	50.77	63.16	57.04	18.18	21.81	52.12
	Law of Obligations and Contracts	70.27	79.05	84.41	80.77	62.16	14.94	22.15
	The Judicial System of the Kingdom	39.32	47.44	54.07	50.06	29.49	29.44	47.19
	The Justice Sector	42.31	55.56	62.54	56.49	30.77	26.31	50.79
	The Moroccan Constitution	49.75	60.61	66.85	62.60	40.91	17.10	32.25
Gemini-2.0-Flash	Civil Procedure	56.63	62.65	69.35	66.70	40.09	12.94	40.54
	Criminal Law	48.37	51.86	58.72	56.04	39.55	40.25	44.13
	Family Code	62.28	64.91	69.04	67.54	55.26	17.44	34.49
	Family Law	60.23	66.91	73.51	70.72	40.91	13.13	41.31
	Law of Obligations and Contracts	73.42	77.48	81.62	80.18	64.86	11.35	25.74
	The Judicial System of the Kingdom	52.49	57.71	63.92	61.18	39.08	21.74	43.95
	The Justice Sector	53.42	67.09	74.67	68.21	38.46	18.64	37.88
	The Moroccan Constitution	69.76	75.12	80.29	78.00	58.57	11.43	30.61
Llama-3.3 (70b)	Civil Procedure	48.29	53.37	61.47	58.97	29.57	22.63	61.61
	Criminal Law	44.24	47.38	57.52	53.79	33.29	44.85	60.60
	Family Code	47.37	52.63	57.98	55.96	34.21	30.00	60.50
	Family Law	42.75	49.43	56.20	53.21	21.21	25.82	66.62
	Law of Obligations and Contracts	66.67	69.82	73.40	72.12	59.46	17.96	37.40
	The Judicial System of the Kingdom	42.33	46.92	53.74	50.92	29.55	32.00	60.75
	The Justice Sector	58.12	65.49	73.99	69.12	43.59	24.01	51.21
	The Moroccan Constitution	59.05	62.14	67.46	65.98	45.71	17.88	48.85
Llama-4-Maverick (17b)	Civil Procedure	53.86	59.98	67.61	65.17	35.15	7.55	33.10
	Criminal Law	46.16	50.90	63.01	58.32	35.70	26.38	28.35
	Family Code	56.14	60.75	65.18	63.33	42.11	9.84	30.08
	Family Law	47.78	54.75	63.47	60.43	24.24	13.12	37.67
	Law of Obligations and Contracts	72.97	78.38	82.52	80.36	64.86	5.92	26.88
	The Judicial System of the Kingdom	46.31	53.03	59.78	56.70	34.09	13.48	37.28
	The Justice Sector	51.92	61.11	68.64	63.93	41.03	9.72	23.45
	The Moroccan Constitution	61.90	67.98	72.86	70.67	54.29	8.35	29.89
Llama-4-Scout (17b)	Civil Procedure	52.26	57.20	64.96	62.62	34.78	22.09	57.10
	Criminal Law	36.94	41.68	56.18	50.63	26.09	48.03	68.15
	Family Code	50.00	55.26	60.18	58.25	39.47	29.33	56.42
	Family Law	44.44	49.65	57.30	54.95	25.76	26.41	69.21
	Law of Obligations and Contracts	69.82	74.32	78.17	76.17	59.46	16.62	35.33
	The Judicial System of the Kingdom	38.92	43.37	49.47	47.30	26.14	32.22	61.12
	The Justice Sector	44.66	56.20	67.20	60.67	33.33	31.06	55.47
	The Moroccan Constitution	65.10	69.44	75.25	73.22	55.07	16.94	41.67

Table 5: The results of different models on MizanQA, stratified by Moroccan law categories. This excludes questions from the "Exam" category, which mixes categories. The exam category was excluded to study the effects of different categories in isolation.

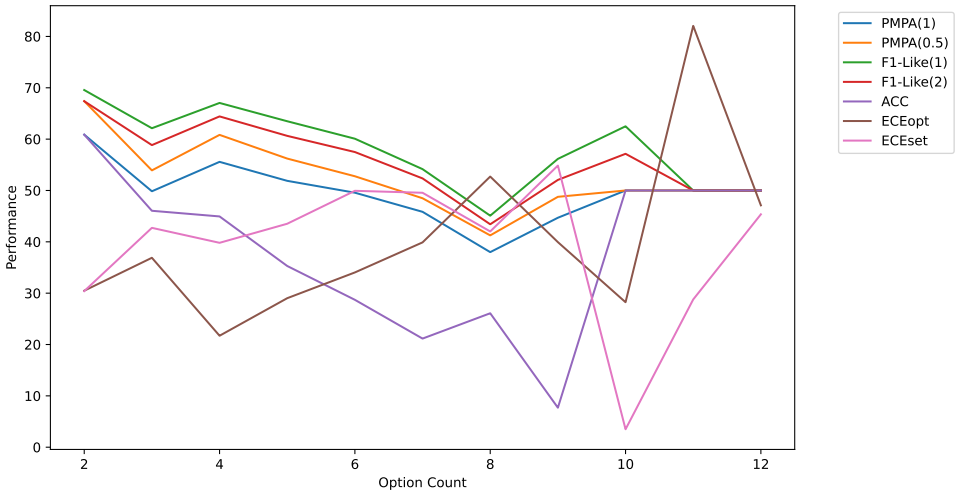

Figure 10: Performance vs. option count for Gemini-2.0-Flash on MizanQA.

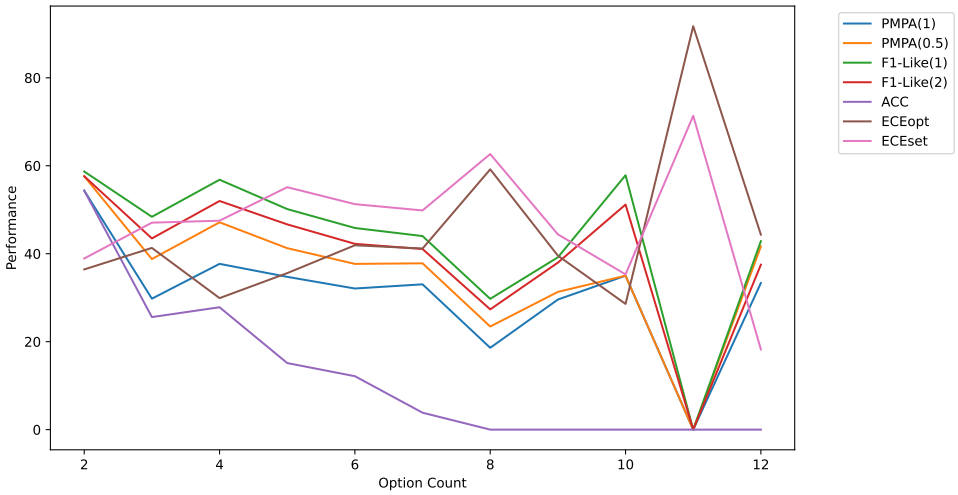

Figure 11: Performance vs. option count for Gemini-1.5-Flash on MizanQA.

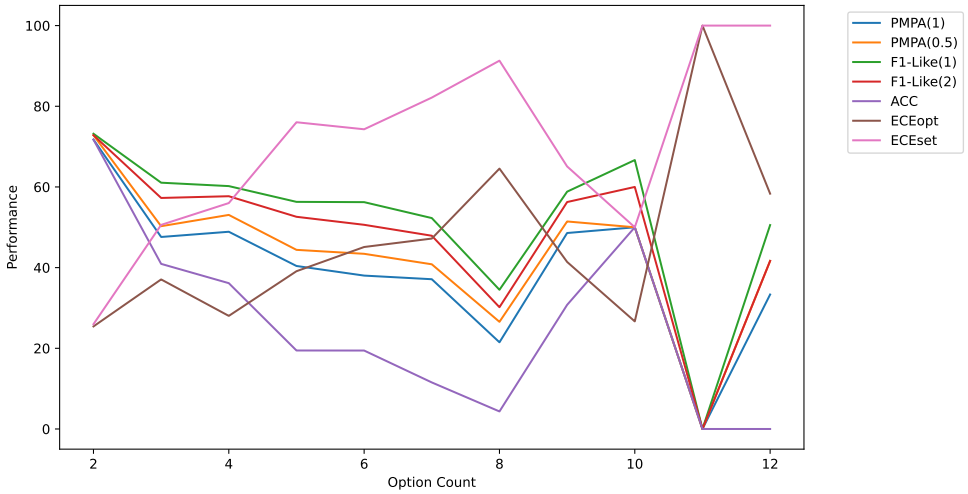

Figure 12: Performance vs. option count for Llama-3.3 (70b) on MizanQA.

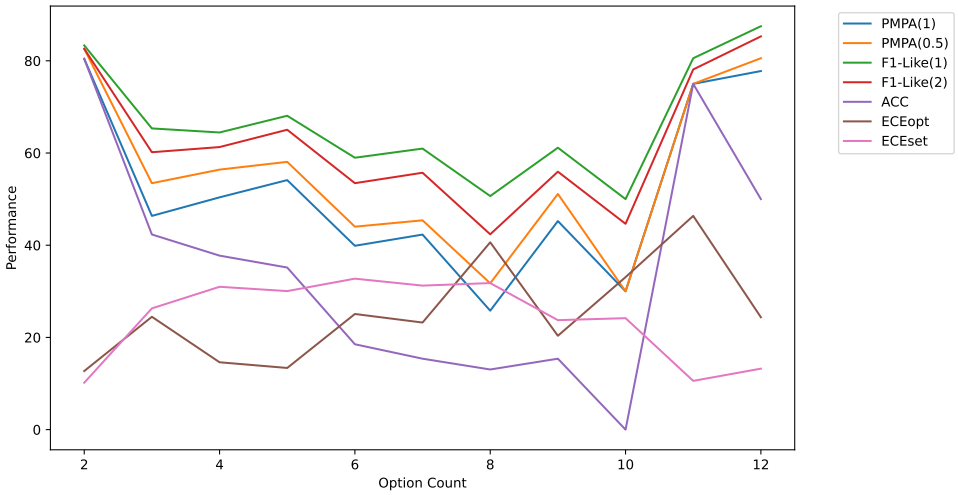

Figure 13: Performance vs. option count for Llama-4-Maverick (17b) on MizanQA.