CVPD at QIAS 2025 Shared Task: An Efficient Encoder-Based Approach for Islamic Inheritance Reasoning

Salah Eddine Bekhouche[1] **Abdellah Zakaria Sellam**[2] **Hichem Telli**[3]
Cosimo Distante[2] **Abdenour Hadid**[4]

[1]University of the Basque Country UPV/EHU, San Sebastian, Spain
[2]Institute of Applied Sciences and Intelligent Systems – CNR, Lecce, Italy
[3]Laboratory of LESIA, University of Biskra, Algeria
[4]Sorbonne University Abu Dhabi, UAE

Abstract

Islamic inheritance law (*'Ilm al-Mawārīth*) requires precise identification of heirs and calculation of shares, which poses a challenge for AI. In this paper, we present a lightweight framework for solving multiple-choice inheritance questions using a specialised Arabic text encoder and Attentive Relevance Scoring (ARS). The system ranks answer options according to semantic relevance, and enables fast, on-device inference without generative reasoning. We evaluate Arabic encoders (MARBERT, ArabicBERT, AraBERT) and compare them with API-based LLMs (Gemini, DeepSeek) on the QIAS 2025 dataset. While large models achieve an accuracy of up to 87.6%, they require more resources and are context-dependent. Our MARBERT-based approach achieves 69.87% accuracy, presenting a compelling case for efficiency, on-device deployability, and privacy. While this is lower than the 87.6% achieved by the best-performing LLM, our work quantifies a critical trade-off between the peak performance of large models and the practical advantages of smaller, specialized systems in high-stakes domains.

1 Introduction

Large Language Models (LLMs) such as GPT-4 (OpenAI, 2023), Gemini (Team et al., 2023), and Deepseek-v3 (Liu et al., 2024) have advanced natural language processing, and show strong reasoning capabilities on many topics. However, as they were mainly trained on general web data, they often struggle in specialised domains with high accuracy (Bubeck et al., 2023). Islamic inheritance law (*'Ilm al-Mawārīth*) is one such area, which is based on fixed rules from the Qur'an and Sunnah and requires a precise understanding of the law and accurate mathematical proportion calculations (Esmaeili, 2012; Phillips and Wilson, 1995). The complexity arises from rules such as *farā'iḍ* (fixed shares), *'awl* (reduction of shares if more than one), and *radd* (increase of shares if less than one) (El-Far, 2011), where errors can cause serious legal and financial problems. General LLMs often fail at such tasks due to the multi-step reasoning and strict numerical precision, especially in Arabic contexts (Arabi and Hassan, 2023). Reinforcing this point, a recent comprehensive study by (Bouchekif et al., 2025b) specifically assessed LLMs on Islamic inheritance law, providing empirical evidence of their limitations in this domain. Therefore, the *QIAS 2025* SubTask 1 becomes a valuable benchmark for the assessment (Bouchekif et al., 2025a). This paper presents a lightweight framework developed for Islamic inheritance reasoning to address these challenges. Our approach combines a pre-trained Arabic text encoder with an Attentive Relevance Scoring (ARS) module. Instead of generating step-by-step generative answers, the system measures how strongly each possible answer relates to the question. The ARS module then ranks the options and selects the correct legal and mathematical outcome. This design focuses on accuracy and efficiency, providing a more feasible solution than large LLMs requiring high computational resources. We compare our specialised model with several leading general-purpose LLMs, including Gemini and DeepSeek, using the official QIAS 2025 dataset. Our experiments show that large models are prone to certain types of errors, especially under specific inference conditions. Our targeted approach, while not perfect, presents an alternative with a different performance and error profile, prioritizing consistency and efficiency. The primary contributions of this work are threefold:

1. We present an efficient, specialized framework that applies an attentive relevance scoring mechanism to pre-trained Arabic encoders for Islamic inheritance reasoning.

2. We provide a comparative analysis comparing our specialized model with SOTA general-

Proceedings of The Third Arabic Natural Language Processing Conference, pages 929–934
November 8-9, 2025 ©2025 Association for Computational Linguistics

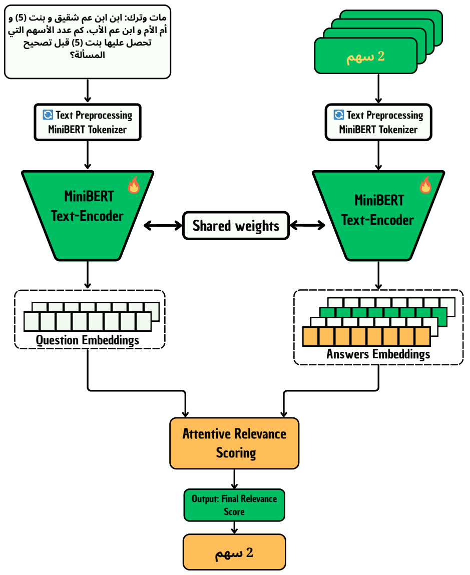

Figure 1: The proposed architecture. Parallel Text Encoders convert a question and answer into Question Embeddings and Answer Embeddings. An Attentive Relevance Scoring module then compares these embeddings to output a Final Relevance Score

purpose LLMs, highlighting the significant impact of inference strategies (batched vs. single input) on LLM performance.

3. We provide empirical evidence of the practical advantages (efficiency, privacy, deployability) of domain-specific models, offering a viable alternative to resource-intensive LLMs despite a performance trade-off.

The remainder of this paper is organised as follows: Section 2 describes our approach in detail; Section 3 presents results and discussion; and Section 4 concludes with future research directions.

2 Methodology

This section describes our proposed hybrid architecture that combines an Arabic text encoder with a scoring mechanism called Attentive Relevance Scoring (ARS) (Bekhouche et al., 2025). We evaluate several Arabic text encoders in this setup. The method aims to improve question answering for Islamic inheritance law by capturing complex semantic relationships while keeping computation lightweight, making it suitable for low-resource environments and edge devices without cloud access. A key design choice is that our approach does not rely on explicit reasoning. Instead, it focuses on providing a fast and low-cost inference solution directly on the device. The text encoder processes both the question and candidate answers, producing dense vector embeddings. The ARS module then scores each candidate answer by assigning higher weights to terms that are contextually important, enabling the system to capture fine-grained details in legal terminology. As shown in Figure 1, the system operates in two stages: (1) the encoder generates semantic embeddings for the question and answers, and (2) ARS refines the ranking by computing a final relevance score. It is important to clarify that our approach does not perform explicit, step-by-step symbolic reasoning. Instead, it is designed to solve this complex reasoning task by learning to identify the candidate answer with the highest semantic relevance to the question.

2.1 Text Encoder

We experiment with five Arabic text encoders: ArabicBERT-Mini (Safaya et al., 2020), ArabicBERT (Safaya et al., 2020), AraBERT (Antoun et al., 2020), MARBERT (Abdul-Mageed et al., 2020), and QARiB (Abdelali et al., 2021). Given a question q and a candidate answer c, the encoder processes each independently, producing two types of representations: (1) **Sequence-level representations**, denoted as $\mathbf{H}_q \in \mathbb{R}^{B \times L \times d}$ and $\mathbf{H}_c \in \mathbb{R}^{B \times L \times d}$, which capture contextual embeddings for each token in the question and the answer. (2) **Pooled representations** from the final-layer [CLS] token.

Here, B is the batch size, L is the input length, and d is the hidden dimension of the model. For all models, $L = 512$. The hidden size d is 256 for ArabicBERT-Mini and 768 for the other encoders. For global semantic representation, we extract the [CLS] token embedding from the final layer and apply ℓ_2 normalization:

$$\begin{aligned} \mathbf{q}_{\text{emb}} &= \text{Norm}(E(q)_{[\text{CLS}]}) \in \mathbb{R}^d, \\ \mathbf{c}_{\text{emb}} &= \text{Norm}(E(c)_{[\text{CLS}]}) \in \mathbb{R}^d, \end{aligned} \tag{1}$$

where $\text{Norm}(\cdot)$ is ℓ_2 normalization. This normalization projects embeddings onto the unit hypersphere, improving stability in similarity computations.

2.2 Attentive Relevance Scoring

The ARS module (Bekhouche et al., 2025) computes adaptive semantic similarity between the question and candidate embeddings via a trainable interaction model. First, both embeddings are projected into a shared latent space:

$$\mathbf{h}_q = W_q \mathbf{q}_{\text{emb}}, \quad \mathbf{h}_c = W_c \mathbf{c}_{\text{emb}}, \tag{2}$$

where $W_q, W_c \in \mathbb{R}^{h \times d}$ are learnable projection matrices and h is the shared hidden dimensionality. Next, element-wise multiplication is applied, followed by a non-linear activation to compute the interaction vector $\mathbf{v}_{\text{int}}$:

$$\mathbf{v}_{\text{int}} = \tanh(\mathbf{h}_q \odot \mathbf{h}_c), \tag{3}$$

where $\odot$ denotes element-wise multiplication and $\tanh(\cdot)$ is the hyperbolic tangent function. Finally, the relevance score r is obtained using an attention vector $w_{\text{att}} \in \mathbb{R}^h$:

$$r = \sigma\left(w_{\text{att}}^\top \mathbf{v}_{\text{int}}\right), \tag{4}$$

where $\sigma(\cdot)$ is the sigmoid function.

2.3 Training Objective

To train the model effectively, we employ a composite training objective designed to optimize for both semantic representation and accurate ranking. This objective is composed of three distinct loss functions, each with a specific goal:

- *Contrastive Loss* ($\mathcal{L}_{\text{cons}}$): Aligns the embeddings of correct question-answer pairs while pushing them apart from incorrect pairs.
- *Dynamic Relevance Loss* ($\mathcal{L}_{\text{dyn}}$): Directly supervises the final ARS scores to ensure the model produces confident and well-calibrated rankings.
- *Relevance Score Logit Regularization* ($\mathcal{L}_{\text{reg}}$): Stabilizes training by encouraging variance in the pre-activation logits, preventing score collapse.

The total loss, $\mathcal{L}_{\text{total}}$, is a weighted sum of these components, formulated as:

$$\mathcal{L}_{\text{total}} = \alpha\mathcal{L}_{\text{cons}} + \beta\mathcal{L}_{\text{dyn}} + \gamma\mathcal{L}_{\text{reg}} \tag{5}$$

We empirically set the balancing weights to $\alpha = 0.4$, $\beta = 0.4$, and $\gamma = 0.2$. A detailed mathematical formulation for each component is provided in Appendix A.

3 Results and Discussion

3.1 Dataset

The dataset in this study is from the official release of SubTask 1: Islamic Inheritance Reasoning in the QIAS 2025 challenge. It covers the rule-based field of Islamic inheritance law, where systems must understand scenarios, identify heirs, apply fixed-share rules, handle diminution and radd return, and calculate exact shares. All questions are multiple-choice with one correct answer, grouped into Beginner, Intermediate, and Advanced levels. The training set has 9,446 samples (5,095 Beginner, 3,431 Intermediate, 920 Advanced), the validation set has 1,000 samples (500 Beginner, 300 Intermediate, 200 Advanced), and the test set has 1,000 samples (500 Beginner, 500 Advanced, no labels). Training and validation have six labels (A–F), with C most common; the test set is unlabeled. Beginner questions involve simple share identification, Intermediate include adjusted shares after radd, and Advanced require full monetary distribution. This dataset is well-suited for testing both language understanding and precise numerical reasoning in Islamic law.

3.2 Experimental Setup

Experiments were performed on a system with seven NVIDIA L4 GPUs, each with 24 GB of VRAM, using a distributed multi-GPU training strategy. Mixed-precision training was not used, and the gradient accumulation step was set to 1 for stability. Optimization was done with the AdamW optimizer, starting at 1×10^{-4} and $\epsilon = 1 \times 10^{-8}$. A cosine annealing scheduler was employed to adjust the learning rate, which was warmed up to 10% of its target before decaying. Gradient clipping with a maximum norm of 0.5 was applied for numerical stability.

3.3 Results and Discussion

Table 1 summarizes the performance and computational costs of various Arabic text encoders in our framework. MARBERT achieved the highest validation and test sets accuracy, showcasing its strong ability to capture the linguistic and domain-specific nuances needed for Islamic inheritance reasoning. Previous research supports that MARBERT, which is trained on extensive Arabic social media data, effectively handles complex morphology and semantic variations. While this analysis primarily compares our models with state-of-the-art (SOTA) large language models (LLMs), future work should benchmark against traditional non-neural baselines (e.g., TF-IDF with cosine similarity) to quantify the advantages of deep learning methods, especially for lower-parameter encoders. We also tested API-based LLMs using two inference strategies to assess the impact of context size on performance. The

Model	Params (M) ↓	GFlops ↓	Results	
			Valid ↑	Test ↑
ArabicBERT-Mini (Safaya et al., 2020) + ARS	11.6	10.3	65.62%	64.23%
ArabicBERT (Safaya et al., 2020) + ARS	110.7	71.1	69.08%	67.19%
AraBERT (Antoun et al., 2020) + ARS	135.3	96.2	73.85%	68.46%
MARBERT (Abdul-Mageed et al., 2020) + ARS	162.9	124.5	**77.32%**	**69.87%**
QARiB (Abdelali et al., 2021) + ARS	135.3	96.2	74.18%	68.63%

Table 1: Performance and computational cost of various Arabic text encoders within our proposed framework on the QIAS 2025 SubTask 1 validation and test sets. MARBERT achieves the highest accuracy, demonstrating its superior ability to handle the linguistic nuances of Islamic inheritance law. Bold values indicate the best performance in each column.

primary method involved a batched approach with 50 questions in a single prompt, which proved efficient but created a large context window. By contrast, the single-question method (used for testing Gemini-2.5-flash) improved accuracy significantly, from 68.65% to 87.60%. This indicates that larger context windows can lead to errors due to cross-question interference. Although API-based models like Gemini and DeepSeek variants outperform our locally trained models regarding accuracy, their high computational requirements prevent direct deployment on edge devices. While running them through cloud services is viable, it entails recurring costs, latency issues, and privacy concerns, making local solutions more attractive in constrained or sensitive environments. Ultimately, these findings reveal a trade-off between performance and deployability. A model with around 70% accuracy is best suited as an assistive tool for legal experts rather than an autonomous decision-maker, facilitating rapid analysis or verification of simple cases, while human oversight remains essential. This positions such models as efficient assistants for on-device or offline scenarios where cloud access is not feasible. Additionally, our experiments demonstrate that inference setup and input structuring significantly impact model behavior, highlighting the importance of evaluation settings when comparing LLM-based systems.

Base Model	**Reasoning**	**ACC ↑**
deepseek-chat	No	66.40%
deepseek-reasoner	Yes	69.40%
gemini-2.0-flash	No	60.44%
gemini-2.5-flash	Yes	68.65%
gemini-2.5-flash*	Yes	87.60%

Table 2: Performance of API-based LLMs. All models were evaluated using a batched input of 50 questions, except where noted by an asterisk (*).

4 Conclusion

We presented a lightweight framework for automated Islamic inheritance reasoning (*'Ilm al-Mawārīth*), combining a specialized Arabic text encoder with an Attentive Relevance Scoring (ARS) mechanism for multiple-choice questions. Our local model, using MARBERT, achieved a test accuracy of 69.87%, which, while promising, is notably lower than the 87.60% reached by leading API-based LLMs like Gemini. This performance difference stems from our model's core design, which forgoes explicit, step-by-step symbolic reasoning in favor of efficient semantic matching. Despite this accuracy trade-off, our approach offers significant advantages in computational efficiency, on-device deployability, and data privacy, making it a viable solution for resource-constrained or offline applications.

These results highlight a critical trade-off between peak performance and practical usability. The current accuracy level positions our system as a valuable **assistive tool** for legal experts rather than a fully autonomous decision-maker, underscoring the necessity of human oversight in such high-stakes, rule-based domains. This demonstrates that lightweight, domain-adapted models remain highly relevant for specific use cases. Future work will directly aim to close the accuracy gap by integrating symbolic reasoning capabilities to handle the precise calculations inherent in inheritance law. We will also explore hybrid approaches that combine the efficiency of our lightweight model with the reasoning power of large models to achieve an optimal balance of performance and practicality.

Acknowledgement

The authors thank Mr. Arturo Argentieri from CNR-ISASI Italy for his technical contribution to the multi-GPU computing facilities. This research was partially funded by the Italian Ministry of University and Research (MUR) with the project "Future Artificial Intelligence Research —FAIR" Grant number PE0000013 CUP B53C22003630006.

References

Ahmed Abdelali, Sabit Hassan, Hamdy Mubarak, Kareem Darwish, and Younes Samih. 2021. Pre-training bert on arabic tweets: Practical considerations. *arXiv preprint arXiv:2102.10684*.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2020. Arbert & marbert: Deep bidirectional transformers for arabic. *arXiv preprint arXiv:2101.01785*.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.

Khalid Arabi and Samira Hassan. 2023. Large language models in the arabic-speaking world: A case study on domain-specific challenges. *Journal of Natural Language Processing Arabia*, 5(2):45–61.

Salah Eddine Bekhouche, Azeddine Benlamoudi, Yazid Bounab, Fadi Dornaika, and Abdenour Hadid. 2025. Enhanced arabic text retrieval with attentive relevance scoring. *arXiv preprint arXiv:2507.23404*.

Abdessalam Bouchekif, Samer Rashwani, Emad Mohamed, Mutaz Al-Khatib, Heba Sbahi, Shahd Gaben, Wajdi Zaghouani, Aiman Erbad, and Mohammed Ghaly. 2025a. Qias 2025: Overview of the shared task on islamic inheritance reasoning and knowledge assessment. In *Proceedings of The Third Arabic Natural Language Processing Conference, ArabicNLP 2025, Suzhou, China, November 5–9, 2025*. Association for Computational Linguistics.

Abdessalam Bouchekif, Samer Rashwani, Heba Sbahi, Shahd Gaben, Mutaz Al-Khatib, and Mohammed Ghaly. 2025b. Assessing large language models on islamic legal reasoning: Evidence from inheritance law evaluation. In *Proceedings of The Second Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, and 1 others. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Ibrahim A. El-Far. 2011. The islamic rules of inheritance. *Journal of Islamic Accounting and Business Research*, 2(1):7–23.

Dr. Abedeen Esmaeili. 2012. *The Islamic Law of Succession: A Practical Guide to the Laws of Faraid*. AS Noordeen.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

OpenAI. 2023. Gpt-4 technical report.

Arthur Phillips and Roland Knyvet Wilson. 1995. *A Treatise on the Muhammadan Law*. Kegan Paul International.

Ali Safaya, Moutasem Abdullatif, and Deniz Yuret. 2020. Kuisail at semeval-2020 task 12: Bert-cnn for offensive speech identification in social media. *arXiv preprint arXiv:2007.13184*.

Gemini Team and 1 others. 2023. Gemini: A family of highly capable multimodal models.

A Detailed Training Objective

This section provides the detailed mathematical formulation of the three loss components used in our training objective. The total loss is defined as:

$$\mathcal{L}_{\text{total}} = 0.4 \cdot \mathcal{L}_{\text{cons}} + 0.4 \cdot \mathcal{L}_{\text{dyn}} + 0.2 \cdot \mathcal{L}_{\text{reg}} \quad (6)$$

A.1 Contrastive Loss ($\mathcal{L}_{\text{cons}}$)

We use an InfoNCE-based contrastive loss on the [CLS] token embeddings. This loss aims to pull the question embedding ($\mathbf{q}$) closer to the correct answer embedding ($\mathbf{c}^+$) and push it away from the five incorrect answer embeddings ($\mathbf{c}^-$).

$$\mathcal{L}_{\text{cons}} = -\frac{1}{B}\sum_{i=1}^{B}\log\left(\frac{e^{\text{sim}(\mathbf{q}_i,\mathbf{c}_i^+)}}{e^{\text{sim}(\mathbf{q}_i,\mathbf{c}_i^+)}+\sum_{j=1}^{5} e^{\text{sim}(\mathbf{q}_i,\mathbf{c}_{i,j}^-)}}\right) \quad (7)$$

where $\text{sim}(\mathbf{a}, \mathbf{b}) = (\mathbf{a}^\top \mathbf{b})/\tau$. Here, $\mathbf{q}_i$ and $\mathbf{c}_i$ are the embeddings for the question and answers, and τ is a trainable temperature parameter.

A.2 Dynamic Relevance Loss ($\mathcal{L}_{\text{dyn}}$)

This loss directly supervises the final ARS scores (r) to ensure they are well-calibrated. It maximizes the score for the correct answer and minimizes the score for a randomly selected incorrect answer.

$$\mathcal{L}_{\text{dyn}} = -\frac{1}{B}\sum_{i=1}^{B}\left[\log(r_i^+ + \epsilon) + \log(1 - r_i^- + \epsilon)\right] \quad (8)$$

Here, r_i^+ and r_i^- are the sigmoid-activated ARS scores for the correct and a randomly chosen incorrect answer. The constant ϵ ensures numerical stability.

A.3 Relevance Score Logit Regularization ($\mathcal{L}_{\text{reg}}$)

To improve training stability, we apply a regularization loss on the raw, pre-sigmoid relevance scores (logits, s). This loss maximizes the variance of the logits within a batch, encouraging the model to use a wider dynamic range for its scores.

$$\mathcal{L}_{\text{reg}} = -(\mathrm{Std}(s_{\text{batch}}^+) + \mathrm{Std}(s_{\text{batch}}^-)) \tag{9}$$

where s_{batch}^+ and s_{batch}^- are the sets of logits for all correct and incorrect answers across the batch. We minimize the negative standard deviation, which is equivalent to maximizing the standard deviation.