# MorAI at QIAS 2025: Collaborative LLM via Voting and Retrieval-Augmented Generation for Solving Complex Inheritance Problems

**Jihad R'baiti[1]*, Chouaib EL Hachimi[2], Youssef Hmamouche[1], Amal El Fallah Seghrouchni[1,3]**

[1] International Artificial Intelligence Center of Morocco (Ai movement), Mohammed VI Polytechnic University (UM6P), Rabat, Morocco.
[2] Research Centre for Artificial Intelligence in Geomatics (RCAIG), Department of Land Surveying and Geo-Informatics (LSGI), The Hong Kong Polytechnic University (PolyU), Kowloon, Hong Kong SAR.
[3] LIP6 - UMR 7606 CNRS, Sorbonne University, Paris, France.

## Abstract

Collaborative approaches have proven effective in addressing complex problems, from human and socio-economic challenges to multi-agent systems. These methods rely on the principle that combining perspectives enhances problem-solving. In this paper, we propose a collaborative large language models (LLM) framework to solve Islamic inheritance problems, which demand precise mathematical reasoning and strict adherence to legal rules for fair distribution among heirs. The system implements a collaborative voting mechanism involving multiple LLMs, namely ALLaM-7B-Instruct-preview, Deepseek-reasoner, and Gemini-2.5-Flash. Each independently answered multiple-choice inheritance questions. The final answer is determined by majority vote. To improve accuracy and domain grounding, we integrate Retrieval-Augmented Generation (RAG). A curated database of solved inheritance cases in JSON format is indexed using TF-IDF. For each query, the most similar cases are retrieved and appended as contextual information to the prompt before being submitted to the LLMs. Experimental results demonstrate that this collaborative RAG-enhanced framework outperforms individual LLMs. The ensemble achieved 88% accuracy, surpassing the best-performing single models: the fine-tuned ALLaM-7B-Instruct-preview (79.50%), Deepeek-reasoner (71.80%), and Gemini-2.5-Flash (83.50%).

## 1 Introduction

LLMs have rapidly gained a prominent role in Natural Language Processing (NLP), transforming how machines understand and generate human-like language. From general-purpose systems like GPT (Achiam et al., 2023) and Gemini (Comanici et al., 2025) to Arabic-focused models such as Fanar (Team et al., 2025) and ALLAM (Bari et al., 2024b), these models have demonstrated remarkable proficiency across a wide range of tasks (Demidova et al., 2024; Singhal et al., 2025; Miah et al., 2024). They are now being used to solve open-domain problems, answer complex questions, and support more specialized areas that require structured knowledge and contextual understanding—such as legal, medical, or religious domains. One such domain is Islamic inheritance, which is based on detailed rules for distributing assets among heirs. Answering questions in this area requires an understanding of Islamic sources and the ability to perform precise calculations involving predetermined shares assigned to each heir. In many cases, small changes in family composition can lead to entirely different outcomes. This makes it a valuable challenge for testing how well LLMs can handle structured reasoning, numerical logic, and Arabic-language understanding in a religious context. To support research in this area, the Question-and-Answer in Islamic Studies Assessment Shared Task (QIAS 2025) (Bouchekif et al., 2025a) was introduced as a benchmark for evaluating the reasoning capabilities of LLMs in the domain of Islamic knowledge. Our focus is on Subtask 1: Islamic Inheritance Reasoning, which presents multiple-choice questions (MCQs) in Arabic across three levels of difficulty: beginner, intermediate, and advanced. The questions are designed to evaluate an LLM's understanding of Islamic inheritance principles and its ability to apply them accurately to solve practical cases. In our final submission, we developed an ensemble-based system that combines RAG with multiple pretrained LLMs to tackle the task's multiple-choice reasoning challenges. We used five models—ALLAM, Fanar, Qween, Gemini, and Deepseek—and applied prompting with RAG during inference. Each model independently predicts an answer (from A to F), and a majority voting strategy is used to select the final response. This setup leverages the con-

*Corresponding author: jihad.rbaiti@um6p.ma



*Proceedings of The Third Arabic Natural Language Processing Conference*, pages 947–952
November 8-9, 2025 

textual strength of RAG and the diverse reasoning capabilities of the models, enabling more robust performance across different question types and difficulty levels. Our system ranked 5th in Subtask 1 of the QIAS 2025 shared task. The full implementation is available at [1]. This paper is organized as follows:
Section 2 provides background relevant to our work. Section 3 describes the dataset, Section 4 presents our method. Section 5 detail the experimental setup used for evaluation. Section 6 presents the result obtained, and Section 7 concludes our work.

## 2 Background

### 2.1 Related Work

Recent work has explored the use of LLMs in religious domains, particularly in the Islamic context. (Mohammed et al., 2025) have proposed an advanced RAG approach using a re-ranker. This approach has proven effective, providing increased response stability, eliminating hallucinations, and obtaining a more accurate answer compared to both base LLM and LLM with RAG methodologies. Similarly, Alan et al. (2025) presented 'MufassirQAS', a system proposed to improve LLM transparency and accuracy using RAG. This system presented relevant sections from the retrieved database alongside the LLM's answers. While Akkila and Naser (2016) developed an expert system designed to simplify and automate the calculation of Islamic inheritance shares based on Sharia law, replacing the traditional method of calculation, aiming to reduce human error and disputes among heirs. In (Bouchekif et al., 2025b), the authors assess LLMs' reasoning capabilities in Islamic inheritance law. The results reveal o3 and Gemini 2.5 as the more accurate models, surpassing ALLaM, Fanar, and Mistral in terms of accuracy.

### 2.2 Islamic inheritance law

Islamic inheritance law (Ilm al-Mawārīth) is a rule-based, mathematical framework derived primarily from Surah An-Nisā' in the Qur'an, which specifies fixed fractional shares for eligible heirs. The framework considers three key factors when distributing shares:

- Degree of kinship: Closer relatives inherit larger shares regardless of gender;
- Generational position: Younger generations preparing for life's responsibilities receive larger shares than older generations relinquishing them;
- Financial responsibility: The only case where gender-based difference applies, when sons inherit twice the share of daughters due to lifelong financial support to their wives and families, while daughters retain their inheritance solely for themselves without any spending obligation.

Verses 4:11–12 and 4:176 outline concise, efficient, and generalizable distribution principles. Children inherit with males receiving the share of two females; parents receive one-sixth each if the deceased has children, with the mother's share adjusted in the presence of siblings; spouses inherit fixed portions depending on the presence of children; and siblings inherit in kalālah, which is the case when there are no ascendants or descendants, with males receiving twice the share of females. These rules are applied only after paying debts and executing valid bequests, which cannot exceed one-third of the estate.

## 3 Data

The approach uses two data sources (Bouchekif et al., 2025a): a structured dataset and a semi-structured dataset. The first one consists of MCQs. It was constructed by converting religio-ethical advice 'fatwas' collected from IslamWeb [2] into a structured format. The preprocessing included reviewing the MCQ by four experts in Islamic studies, rephrasing ambiguous questions, and eliminating semantic and numerical redundancies. Each MCQ has six answer options (A to F), with only one correct. The dataset design uses an increasing complexity of three levels of difficulty—beginner, intermediate, and advanced—to assess LLMs' capability across different levels of expertise. The dataset contains approximately 20000 MCQs for training, 1000 for validation, and 1000 for testing. On the other hand, the semi-structured dataset consists of 4 JSON files containing all necessary information about the problem statements 'fatwas', including ID, URL, category, Gregorian and Hijri dates, question, and answer. It was sourced from IslamWeb and provides about 3065 resolved inheritance problem statements. This dataset was used as

[1] https://shorturl.at/Fev1p

[2] https://www.islamweb.net/

an external source of knowledge to retrieve pertinent similar inheritance cases in our proposed RAG system.

## 4 System Overview

The proposed framework integrates a Multi-LLM Voting Framework with a Retrieval Module to address Arabic Islamic inheritance MCQs (Figure 1). First, a curated JSON knowledge base of solved inheritance cases 'Fatwas' is built and indexed using TF-IDF. After storing the index, each problem statement is vectorized as a query for the RAG system and projected in the index vector space model to be compared with the corpus's documents using cosine similarity, retrieving the top-k most similar cases. These retrieved cases are appended to the original problem statement prompt (Figure 2). This prompt augmentation enriches the context during the inference phase. The augmented prompt is then passed to multiple LLMs, which are the fine-tuned ALLaM-7B, Gemini-2.5-Flash, and DeepSeek-R1. The models independently generate their answers. Finally, the outputs are processed through a voting mechanism that selects the majority answer as the final prediction. In cases where no majority is reached, priority is given to Gemini, as it demonstrated the most reliable performance during experiments.

## 5 Experimental Setup

The goal of the shared task was to evaluate the ability of various LLMs to solve Islamic inheritance reasoning problems, which involve applying strict legal mathematical rules. Multiple submission strategies (Table 1) were developed and evaluated. The first approach involved supervised fine-tuning of the 7 billion version of the ALLaM model, which was selected for its reported strong performance in Arabic (Bari et al., 2024a). The training data consisted of inheritance cases formatted in a question-answering style described in Section 3. The model was trained using standard cross-entropy loss to directly predict the correct inheritance distribution, using the listed hyperparameters (Figure 2). The second approach uses zero-shot learning, where a prompt-based inference was applied without additional fine-tuning. Here, multiple pretrained LLMs were tested, including variants of ALLaM, Fanar, Gemini, Qween, and DeepSeek. Each model was prompted using a fixed template describing the inheritance case and requesting a response in a structured format. Next, a RAG setup was implemented to provide contextual fatwa-based information. Relevant fatwas were retrieved for each case and appended to the original prompt. Two retrieval methods were compared: one using a Neural Embedding for semantic search, and another using TF-IDF for keyword-based retrieval. The same prompting strategy was applied in both cases. Finally, based on model performance, the study chose 3 models to construct the pool of voting to implement the proposed majority voting collaborative strategy.

During implementation of RAG, we experimented with different values of top-k voting using k $\in \{3, 5, 7\}$, to evaluate the effect of the number of retrieved documents (k) and retrieval method (TF-IDF vs. Neural Embeddings). The combination of k=3 using TF-IDF was selected for a balance between computation, prompt input length, and model performance. For the inference, we applied parameter-efficient tuning via LoRA with rank set to 8, alpha set to 16, and dropout set to 0.05, using a maximum input length of 3000 tokens. Decoding was performed with beam search (num_beams=5) combined with sampling (temperature=0.6, top_p=0.9), and the maximum number of generated tokens was limited to 20.

This study uses the Mohammed VI Polytechnic University's high-performance computing (HPC) called 'TOUBKAL', which provides a cluster of various types of computational nodes equipped with NVIDIA A100-SXM4-80GB GPUs. To access the HPC, the MobaXterm software is used as an SSH client for establishing connections to the HPC. The programming language used is Python 3.10, along with Anaconda, to manage packages and dependencies in both local and remote environments.

## 6 Results and Discussion

Table 1 summarizes the performance of our submissions in terms of accuracy. Initial experiments with Arabic-specific models, such as Fanar and ALLaM, showed limited effectiveness in handling Islamic inheritance reasoning. The fine-tuned ALLaM model produced similar results. Subsequent submissions explored prompt-based inference with larger multilingual models. The reasoning-focused version of DeepSeek (Deepseek-Reasoner via API) achieved a significant performance gain, and Gemini-2.5-Flash further improved accuracy to 78.10%, high-

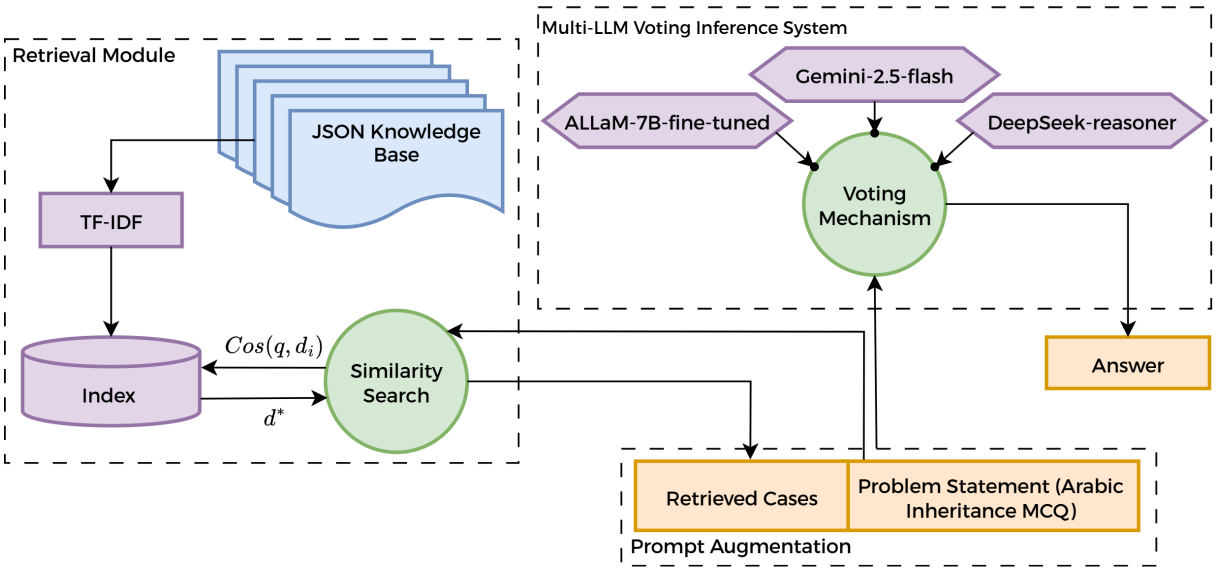


Figure 1: Overall workflow of the implemented collaborative LLMs framework

```
System:
You are a specialist in Islamic sciences. Your task is to answer multiple-choice
questions by selecting the correct option.

User:
Question: {question}

{options_text}

The following fatwas may assist you: {context_text}

Please respond using only one English letter from the following: {valid_letters}
Do not write any explanation or additional text.
```

Figure 2: Prompt structure in the Collaborative LLM Framework; {question} denotes the original inheritance problem statement, {options_text} lists answer choices (one correct), {context_text} contains RAG-retrieved Fatwas for prompt augmentation, and {valid_letters} specifies the expected output format

lighting the benefits of multilingual models over those optimized for a single or limited set of languages. The proposed collaborative voting strategy exploits both the diversity of models (ALLaM, Gemini, DeepSeek), with RAG. This approach achieved the highest accuracy (88.00%), demonstrating the effectiveness of model combination and knowledge retrieval in handling the complex reasoning required for Islamic inheritance cases. A limitation of our voting strategy arises when the three models produce three different answers. In such cases, we default to Gemini's output, given its stronger capability compared to the candidate models. While this rule provided a practical solution in our experiments, it reduces the neutrality of the ensemble. Future work will explore more robust strategies, such as weighted voting, where models' contributions are scaled according to their accuracy.

## 7 Conclusion

In this work, we introduced a collaborative LLM approach augmented with RAG to tackle Islamic inheritance problems in Arabic. By aggregating heterogeneous models, namely Gemini-2.5-Flash, DeepSeek, and ALLaM-7B, all augmented with RAG, through a majority-vote approach, the system was 88.00% accurate, surpassing the top-performing single model, while achieving stronger robustness to model-specific faults. Our experiments showed that retrieval configuration affects model performance, and TF-IDF at k = 3 performs best, surpassing neural embedding methods under this task. These results report that, for such

Table 1: Accuracy of individual models compared to the collaborative voting approach.

| Submission | Model | Accuracy (%) |
|---|---|---|
| 1 | Fanar-1-9B | 36.10 |
| 2 | ALLaM-7B-Instruct-preview | 26.50 |
| 3 | Fine-tuned ALLaM-7B-Instruct-preview | 79.50 |
| 4 | Deepseek-chat | 50.90 |
| 5 | Qwen3-1.7B with RAG | 26.10 |
| 6 | Gemini-2.5-Flash | 78.10 |
| 7 | Gemini-2.5-Flash with RAG | 83.50 |
| 8 | Deepseek-reasoner with RAG | 71.80 |
| **9** | Collaborative Voting | **88.00** |

Table 2: Finetuning Hyperparameters

| Hyperparameter | Value |
|---|---|
| Max training steps | 300 |
| Batch size | 2 |
| Eval batch size | 8 |
| Learning rate | 5e-5 |
| Max sequence length | 1024 |
| Logging frequency | 50 steps |
| Checkpoint frequency | 200 steps |
| Random seed | 42 |
| LoRA rank ($r$) | 8 |
| LoRA alpha | 16 |
| LoRA dropout | 0.05 |
| Optimizer | AdamW |
| Gradient clipping | 1.0 |
| Precision | bfloat16 (GPU) |
| Max new tokens (eval) | 20 |
| Sampling strategy | Greedy |

highly structured legal reasoning problems with explicit rules, conventional lexical retrieval can be more successful, and that multi-model collaborative LLMs are more trustworthy in output than solitary models. Further work involves incorporating more extensive and heterogeneous Arabic-specific models, increasing the dataset, and testing the approach for low-resource language translation.

## References


Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Alaa N Akkila and Samy S Abu Naser. 2016. Proposed expert system for calculating inheritance in islam.

Ahmet Yusuf Alan, Enis Karaarslan, and Ömer Aydın. 2025. Improving llm reliability with rag in religious question-answering: Mufassirqas. *Turkish Journal of Engineering*, 9(3):544–559.

M. Saiful Bari, Yazeed Alnumay, Norah A. Alzahrani, Nouf M. Alotaibi, Hisham A. Alyahya, Sultan Al-Rashed, Faisal A. Mirza, Shaykhah Z. Alsubaie, Hassan A. Alahmed, Ghadah Alabduljabbar, Raghad Alkhathran, Yousef Almushayqih, Raneem Alnajim, Salman Alsubaihi, Maryam Al Mansour, Majed Alrubaian, Ali Alammari, Zaki Alawami, Abdulmohsen Al-Thubaity, and 6 others. 2024a. Allam: Large language models for arabic and english. *13th International Conference on Learning Representations, ICLR 2025*, pages 59235–59270.

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

Abdessalam Bouchekif, Samer Rashwani, Mohammed Ghaly, Mutaz Al-Khatib, Emad Mohamed, Wajdi Zaghouani, Heba Sbahi, Shahd Gaben, and Aiman Erbad. 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 Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Anastasiia Demidova, Hanin Atwany, Nour Rabih, and Sanad Sha'ban. 2024. Arabic train at nadi 2024 shared task: Llms' ability to translate arabic dialects into modern standard arabic. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 729–734.

Md Saef Ullah Miah, Md Mohsin Kabir, Talha Bin Sarwar, Mejdl Safran, Sultan Alfarhood, and MF Mridha. 2024. A multimodal approach to cross-lingual sentiment analysis with ensemble of transformer and llm. *Scientific Reports*, 14(1):9603.

Marryam Yahya Mohammed, Sama Ayman Ali, Salma Khaled Ali, Ayad Abdul Majeed, and Ensaf Hussein Mohamed. 2025. Aftina: enhancing stability and preventing hallucination in ai-based islamic fatwa generation using llms and rag. *Neural Computing and Applications*, pages 1–26.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, and 1 others. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950.

Fanar Team, Ummar Abbas, Mohammad Shahmeer Ahmad, Firoj Alam, Enes Altinisik, Ehsannedin Asgari, Yazan Boshmaf, Sabri Boughorbel, Sanjay Chawla, Shammur Chowdhury, and 1 others. 2025. Fanar: An arabic-centric multimodal generative ai platform. *arXiv preprint arXiv:2501.13944*.