SHA at the QIAS Shared Task: LLMs for Arabic Islamic Inheritance Reasoning

Shatha Altammami
King Saud University, Saudi Arabia
shaltammami@ksu.edu.sa

Abstract

This paper presents our system for SubTask 1: Islamic Inheritance Reasoning in the QIAS 2025 Shared Task, which evaluates large language models (LLMs) on (*ilm al-mawārīth*) (Islamic science of inheritance) using a benchmark of Arabic multiple-choice questions (MCQs) derived from expert-reviewed fatwas. We explore static and dynamic few-shot prompting, retrieval-augmented generation (RAG) with a large fatwa corpus, and a progressive n-gram overlap retrieval method. The n-gram method is applied both to select the top five most similar MCQs for dynamic prompting and to retrieve the most relevant fatwa answer as additional context during inference. We evaluate proprietary and open-source LLMs individually and in ensemble form. Results show that dynamic prompting and RAG consistently improve accuracy across our best performing model, Gemini, achieving 62.26% accuracy on the test set.

1 Introduction

Large Language Models (LLMs) have achieved impressive advances in reasoning and problem solving(Plaat et al., 2024; Zhao et al., 2023), yet their performance often varies across languages and domains (Matarazzo and Torlone, 2025). While most prior work has focused on English, a growing body of research has examined Arabic, revealing mixed results in comprehension and complex reasoning (Khondaker et al., 2024).

One underexplored domain is (*ilm al-mawārīth*) (Islamic science of inheritance), which requires mapping textual descriptions of heirs to precise share distributions — a task demanding multi-step reasoning and domain-specific accuracy. To enable systematic evaluation in this underexplored domain, the QIAS 2025 Shared Task offers a large-scale benchmark of Arabic multiple-choice questions (MCQs) on (*ilm al-mawārīth*) (Bouchekif et al., 2025a,b).

In this paper, we describe our system for the shared task, which integrates static and dynamic few-shot prompting, retrieval-augmented generation (RAG) using a large fatwa corpus, and a progressive n-gram overlap retrieval method. The n-gram method is employed in two ways: (1) to retrieve the top five most similar MCQs from the training set for dynamic prompting, and (2) to identify the most relevant fatwa question and extract its answer as contextual input for inference.

We evaluate both proprietary and open-source models, individually and in an ensemble configuration. Results show that dynamic prompting and RAG provide consistent improvements, with our best-performing model, Gemini, achieving 62.26% accuracy on the test set.

2 Related Work

LLMs have demonstrated strong performance on text-based multiple-choice questions (MCQs), particularly in factual recall and reading comprehension tasks(Matarazzo and Torlone, 2025). Proprietary models such as GPT-4 and Gemini consistently achieve high accuracy on knowledge-based and standardized exam questions, with documented success on domains such as the Dental Admission Test (DAT) (Hou et al., 2025). Similarly, large open-source models like LLaMA3-70B perform competitively in natural sciences and reading comprehension domains (Hou et al., 2025). However, these models consistently struggle with higher-order cognitive skills, multi-step reasoning, and advanced mathematical problem solving, with hallucination remaining a persistent issue, especially in complex reasoning scenarios (Saxena et al., 2024).

Although most prior work has focused on evaluating models in English, some studies have examined Arabic. Existing research shows that LLMs demonstrate mixed performance in comprehend-

Proceedings of The Third Arabic Natural Language Processing Conference, pages 867–872

ing Arabic content and solving complex reasoning tasks. Proprietary models such as GPT-4 and GPT-3.5 perform competitively but are often outperformed by smaller, fine-tuned Arabic models on domain-specific tasks (Khondaker et al., 2023). In contrast, open-source models like LLaMA-3-70B still lag behind both ChatGPT and specialized Arabic models, partly due to limited Arabic representation in large pretraining corpora and high sensitivity to input phrasing (Khondaker et al., 2024).

The closest relevant evaluation is the Qur'an Question Answering shared task (Malhas et al., 2022, 2023), which addressed Machine Reading Comprehension (MRC) over Classical Arabic text. It highlighted the Qur'an's linguistic complexity and topic diversity. Their results emphasize the gap between general Arabic NLP progress and the sensitive religious domains..

Despite this growing body of work, there is a clear research gap: no empirical studies have systematically evaluated LLMs' accuracy in answering questions across diverse areas of Islamic scholarship. Current literature focuses on general NLP benchmarks and professional examinations, leaving domain-specific tasks such as Islamic jurisprudence (fiqh) and inheritance law (*ilm al-mawarith*) largely unexplored.

3 Task Description

The shared task focuses on evaluating LLMs in the Islamic domain, with a particular emphasis on their ability to reason about inheritance-related scenarios (*ilm al-mawārīth*). In this subtask, each multiple-choice question (MCQ) presents a specific inheritance case describing a set of heirs, and the proposed model must determine the correct distribution outcome by selecting the right option from a predefined set of answers. The evaluation is based on classification accuracy over a held-out test set, ensuring an objective comparison of model performance.

4 Dataset

Experiments were conducted using the official dataset provided for the Islamic Inheritance Reasoning task. The dataset comprises multiple-choice questions (MCQs) drawn from authentic Islamic jurisprudential sources and are designed to test not only factual recall but also the model's ability to apply complex, rule-based reasoning grounded in Islamic law. Also a supplementary fatwa corpus is provided which we used for the retrieval-augmented generation (RAG)-based inference.

MCQ Dataset

- **Training Set:** 20,000 MCQs, distributed across three difficulty levels: 500 Beginner, 300 Intermediate, and 200 Advanced.
- **Validation Set:** 1,000 MCQs, distributed across three difficulty levels: 500 Beginner, 300 Intermediate, and 200 Advanced.
- **Test Set:** 1,000 MCQs with hidden labels, balanced between 500 Beginner and 500 Advanced questions.

Each MCQ includes 4 to 6 answer options (A–F), with exactly one correct label. The questions span a wide range of inheritance scenarios requiring precise application of Islamic legal principles (*ilm al-mawārīth*).

Fatwa Corpus

In addition to the MCQ dataset, we used a corpus of 3,165 fatwas from IslamWeb to support retrieval-augmented generation (RAG). Stored as JSON files, each fatwa contains a user-submitted question and an expert legal response, offering rich, domain-specific context to enhance model reasoning.

5 Methodology

As baseline models, we fine-tuned the top-performing model from the Qur'an Question Answering shared task (Malhas et al., 2022, 2023), AraBERTv2 (Antoun et al., 2020), on the 20,000-question training set, achieving an accuracy of 47.4% on the validation (development) set. Another baseline(code was provided by the shared task organizers) involved prompting the Fanar LLM with two few-shot MCQ examples, which yielded 49.7% accuracy on the same validation (development) set. Building on the best-performing baseline, we consulted the literature and identified key areas for improvement.

5.1 Few-Shot In-Context Learning

Few-shot prompting has proven effective in eliciting structured reasoning from LLMs (Brown et al., 2020; Kojima et al., 2022). Static few-shot examples provide a general template for reasoning, whereas dynamic example selection can improve

performance by aligning examples with the test instance (Liu et al., 2022). In this work, we explore both static and dynamic prompting strategies. In the static approach, five examples are included in the prompt for every question. In the dynamic approach, the five most similar questions are retrieved from the training set and included in the prompt. Similarity is determined using an n-gram overlap strategy previously introduced in Altammami et al. (2019), originally developed for segmenting and annotating Hadith corpora. The algorithm has been adapted for the current task, as explained in Section 5.3.

5.2 Retrieval-augmented generation

We utilized Retrieval-Augmented Generation (RAG), a widely recognized and effective approach for improving NLP tasks (Lewis et al., 2020; Wu et al., 2024), particularly in knowledge-intensive and domain-specific scenarios (Xiong et al., 2024). RAG consistently enhances answer accuracy, factuality, and adaptability compared to language models that rely solely on pre-trained knowledge (Siriwardhana et al., 2023).

Initial experiments using vector-based semantic similarity methods (e.g., FAISS) yielded suboptimal results. These approaches often failed to distinguish between conceptually distinct heirs (e.g., *son* vs. *daughter*), treating them as similar due to surface-level embedding similarities. This limitation is particularly problematic in the domain of Islamic inheritance law, where precise legal roles carry significant implications.

To address this, our system identifies the most similar fatwa question from a large corpus using the n-gram approach described in Section 5.3. The corresponding fatwa answer is then extracted and incorporated as additional context for the language model during inference.

5.3 Progressive n-gram Overlap

To support both dynamic few-shot prompting and retrieval-augmented generation (RAG), we developed a custom progressive n-gram overlap matching function. This method is used in two key places: (1) to select the most similar five MCQ questions from the training set for dynamic prompting, and (2) to identify the most relevant fatwa question from the Fatwa Corpus in order to extract its answer as additional context during inference.

Given a new inheritance question, the system iterates through the relevant dataset (training set or Fatwa Corpus) and compares the input question against all available questions using the progressive n-gram overlap function. The matching function follows a fallback strategy: It first computes trigram overlap, and if no sufficient match is found, it falls back to bigram or unigram overlap. We assign higher weights to longer n-grams to prioritize more specific matches: trigrams $w_3 = 1.0$, bigrams $w_2 = 0.5$, and unigrams $w_1 = 0.2$.

For dynamic prompting, the top five questions with the highest similarity scores from the training set are selected and included in the prompt. For RAG, the single highest-scoring fatwa question is selected from across all fatwa JSON files, and its `Answer` field is extracted and supplied to the language model as contextual input, as illustrated in Algorithm 1.

This approach ensures that retrieved examples and contextual fatwas are both semantically and structurally aligned with the input question, avoiding misleading matches that often occur with purely embedding-based similarity methods.

6 Experimental Design

6.1 Models

Four LLMs were evaluated in this study. Inference was configured to favor deterministic, short outputs by setting the temperature to 0.0 and the maximum output length to 2 tokens (sufficient to return a single uppercase letter).

- **Gemini-1.5-pro**: Google's generative language model, accessed via the Google Vertex AI API.
- **GPT-4**: OpenAI's GPT-4o model, accessed through the OpenAI API.
- **LLaMA**: Meta's LLaMA-3.3-70b model, accessed via the Groq API.
- **Fanar**: A domain-specific Arabic language model, accessed through a custom API.

6.2 Prompt Engineering

Three prompt engineering strategies were designed to evaluate their impact on model performance in Islamic inheritance reasoning:

- **Trial 1: Static Few-Shot In-Context Learning**
  A baseline configuration using five manually selected MCQ examples from the training

Model	Trial 1		Trial 2		Trial 3	
	Dev	Test	Dev	Test	Dev	Test
Gemini	60.50	57.30	60.10	61.40	61.80	**62.26**
GPT-4	58.30	55.55	57.60	47.90	54.10	46.30
Fanar	57.27	40.24	54.06	38.49	56.27	38.74
LLaMA	46.40	49.40	46.60	46.10	45.65	47.40
Ensemble	62.60	55.80	61.90	56.40	**63.40**	57.30

Table 1: Performance (%) of different models across three trials on the Islamic inheritance MCQ development and test datasets. Best results are shown in **bold**.

Algorithm 1: Progressive N-gram Matching for Fatwa Retrieval

Input : Question Q;
Set of Fatwa Files $\mathcal{F}$ (each containing Question, Answer fields)
Output : Best matching fatwa answer A^*

Initialize:
$best\_score \leftarrow -\infty$, $A^* \leftarrow$ None
foreach *fatwa file* $f \in \mathcal{F}$ **do**
  Load all questions Q_f and answers A_f from f;
  foreach *candidate question* $q \in Q_f$ **do**
    Normalize Q and q by removing punctuation, extra spaces;
    $score \leftarrow 0$;
    for $n \in \{3, 2, 1\}$ **do**
      Extract n-grams from Q and q;
      Compute $overlap \leftarrow$ intersection of n-grams;
      Update $score \leftarrow score + w_n \times |overlap|$;
      Remove matched n-grams from further consideration;
    if $score > best\_score$ **then**
      $best\_score \leftarrow score$;
      $A^* \leftarrow$ corresponding answer to q;
return A^*

set. These examples were appended to each prompt uniformly, without regard to question similarity.

- **Trial 2: Dynamic Few-Shot In-Context Learning**
  Few-shot examples were dynamically selected for each input question using n-gram similarity from the training set. This ensured structural and semantic relevance between the input and the few-shot examples.

- **Trial 3: Dynamic Few-Shot In-Context Learning and RAG**
  In addition to dynamic example selection, the most similar fatwa question was retrieved using n-gram overlap, and the corresponding fatwa answer was appended to the prompt as context.

Model performance was assessed using accuracy of correctly answered 1,000 MCQs testing questions.

6.3 Results

Table 1 reports development and test accuracies across three independent trials. Gemini consistently outperforms other single models, achieving the highest test accuracy in Trial 3 (62.26%). GPT-4 performs competitively in Trial 1 but its accuracy declines sharply in later trials. Fanar and LLaMA lag behind, though LLaMA generally surpasses Fanar on the test set.

The ensemble method, based on majority voting, yields the best development accuracy in Trial 3 (63.40%) and consistently competitive results overall. Its improvements are more pronounced on development data than on test data, reflecting differences in dataset composition: The dev set contains beginner, intermediate, and advanced items, while the test set excludes intermediate items. This mismatch reduces the ensemble's generalization strength.

Gemini's steady gains suggest that it leverages additional retrieved context effectively, whereas GPT-4 appears more prone to "distraction," with the same context introducing noise and lowering accuracy. These contrasting behaviors highlight model-specific sensitivities to retrieval-augmented prompting, and further analysis is needed in future work to better understand how such distractions arise.

7 Conclusion

This paper presented our system for SubTask 1: Islamic Inheritance Reasoning in the QIAS

2025 Shared Task, where we evaluated static few-shot prompting, dynamic few-shot prompting, and dynamic prompting combined with retrieval-augmented generation, supported by a progressive n-gram overlap method. Evaluation on proprietary and open-source LLMs revealed that while some models experienced performance drops—suggesting that additional context can sometimes distract the model—others achieved consistent gains. Our best configuration (Gemini with RAG and dynamic prompting) reached 62.26% accuracy on the test set. Further analysis is required to better understand how retrieval context may distract certain models and how to design strategies that mitigate this effect.

Acknowledgments

I would like to thank the organizers and reviewers of this shared task.

References

Shatha Altammami, Eric Atwell, and Ammar Alsalka. 2019. Text segmentation using n-grams to annotate hadith corpus. In *Proceedings of the 3rd Workshop on Arabic Corpus Linguistics*, pages 31–39.

Wissam Antoun and 1 others. 2020. Arabert: Transformer-based model for arabic language understanding. In *LREC*.

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 Second 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.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Yu Hou, Jay Patel, Liya Dai, Emily Zhang, Yang Liu, Zaifu Zhan, Pooja Gangwani, and Rui Zhang. 2025. Benchmarking of large language models for the dental admission test. *Health Data Science*, 5:0250.

Md Tawkat Islam Khondaker, Numaan Naeem, Fatimah Khan, Abdelrahim Elmadany, and Muhammad Abdul-Mageed. 2024. Benchmarking llama-3 on arabic language generation tasks. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 283–297.

Md. Tawkat Islam Khondaker, Abdul Waheed, El Moatez Billah Nagoudi, and M. Abdul-Mageed. 2023. Gptaraeval: A comprehensive evaluation of chatgpt on arabic nlp. *ArXiv*, abs/2305.14976.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wentau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Rana Malhas, Watheq Mansour, and Tamer Elsayed. 2022. Qur'an QA 2022: Overview of the first shared task on question answering over the holy qur'an. In *Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*, pages 79–87, Marseille, France. European Language Resources Association.

Rana Malhas, Watheq Mansour, and Tamer Elsayed. 2023. Qur'an QA 2023 shared task: Overview of passage retrieval and reading comprehension tasks over the holy qur'an. In *Proceedings of ArabicNLP 2023*, pages 690–701, Singapore (Hybrid). Association for Computational Linguistics.

Andrea Matarazzo and Riccardo Torlone. 2025. A survey on large language models with some insights on their capabilities and limitations. *ArXiv*, abs/2501.04040.

A. Plaat, Annie Wong, Suzan Verberne, Joost Broekens, N. V. Stein, and T. Back. 2024. Reasoning with large language models, a survey. *ArXiv*, abs/2407.11511.

Yash Saxena, Sarthak Chopra, and Arunendra Mani Tripathi. 2024. Evaluating consistency and reasoning capabilities of large language models. In *2024 Second International Conference on Data Science and Information System (ICDSIS)*, pages 1–5. IEEE.

Shamane Siriwardhana, Rivindu Weerasekera, Elliott Wen, Tharindu Kaluarachchi, Rajib Rana, and Suranga Nanayakkara. 2023. Improving the domain adaptation of retrieval augmented generation (rag) models for open domain question answering. *Transactions of the Association for Computational Linguistics*, 11:1–17.

Shangyu Wu, Ying Xiong, Yufei Cui, Haolun Wu, Can Chen, Ye Yuan, Lianming Huang, Xue Liu, Tei-Wei Kuo, Nan Guan, and 1 others. 2024. Retrieval-augmented generation for natural language processing: A survey. *arXiv preprint arXiv:2407.13193*.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6233–6251.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Z. Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, and 3 others. 2023. A survey of large language models. *ArXiv*, abs/2303.18223.