PuxAI at QIAS 2025: Multi-Agent Retrieval-Augmented Generation for Islamic Inheritance and Knowledge Reasoning

Nguyen Xuan Phuc, Dang Van Thin
University of Information Technology-VNUHCM,
Vietnam National University, Ho Chi Minh City, Vietnam
23521213@gm.uit.edu.vn
thindv@uit.edu.vn

Abstract

This paper addresses the challenge of applying Large Language Models (LLMs) to Islamic jurisprudence, a domain that requires both textual retrieval and precise rule-based reasoning. We focus on the QIAS 2025 shared task, which evaluates LLMs on two subtasks: Islamic inheritance reasoning and general Islamic knowledge assessment. Prior works in Arabic NLP and religious QA largely emphasize retrieval and classification, but they do not evaluate multi-step procedural reasoning. To fill this gap, we propose a hybrid multi-agent framework, termed Retrieval-Augmented Reasoning (RAR). For inheritance problems, our Virtual Inheritance Expert parses natural language cases into structured JSON, retrieves relevant fatwas, and applies rule-based synthesis. For general knowledge, our Proponent–Critic Debate simulates dialectical reasoning, with a head scholar model providing final judgment. Using an ensemble of Gemini, Fanar, and Mistral, our system achieved 2nd place in Subtask 1 and 1st place in Subtask 2. These results demonstrate that decomposing complex reasoning into specialized pipelines supports robustness and accuracy in high-stakes domains.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable progress in natural language understanding and generation, yet their application to highly specialized domains remains a significant challenge. In such contexts, models must go beyond broad knowledge recall and perform deep, rule-based reasoning. A representative example is Islamic jurisprudence (*fiqh*), which not only requires accurate reference to classical sources but also mastery of intricate logical systems. Among its most complex branches is the science of inheritance (*ʻlm al-mawārīth*), where precise multi-step calculations and hierarchical rules determine legally binding outcomes.

This paper presents our system for the QIAS 2025 Shared Task (Bouchekif et al., 2025a), a benchmark designed to evaluate the reasoning capabilities of LLMs in Islamic sciences. The competition is divided into two subtasks. Subtask 1, *Islamic Inheritance Reasoning*, focuses on *ʻlm al-mawārīth*, testing a model's ability to apply fixed-share rules (*farāʻiḍ*), handle residuary shares, and resolve complex inheritance scenarios. Subtask 2, *Islamic Knowledge Assessment*, evaluates broader expertise across seven disciplines, including Quranic studies (*ʻulūm al-Qurʻān*), hadith criticism (*ʻulūm al-Ḥadīth*), and legal theory (*uṣūl al-fiqh*). Both subtasks are structured into three levels of difficulty: beginner, intermediate, and advanced.

Our system adopts a hybrid strategy that combines Retrieval-Augmented Generation (RAG) (Lewis et al., 2021), prompt engineering, few-shot learning (Brown et al., 2020), and a voting-based model ensemble (Devvrit et al., 2020). This design addresses the unique demands of each subtask and achieves state-of-the-art performance. Submissions are evaluated based on accuracy, pushing the boundaries of what LLMs can achieve in high-stakes, expert domains. Our implementation is publicly available[1].

2 Related Work

This work lies at the intersection of Large Language Models (LLMs), Arabic and Islamic Natural Language Processing, and complex, rule-based reasoning. While LLMs have demonstrated strong performance in high-stakes domains such as law and medicine, their evaluation has largely focused on knowledge retrieval, summarization, and classification. Benchmarks like LegalBench (Guha et al., 2023) and Med-PaLM 2 (Singhal et al., 2023) assess factual accuracy and domain understanding,

[1] https://github.com/PuxHocDL/Question-and-Answer-in-Islamic-Studies

Proceedings of The Third Arabic Natural Language Processing Conference, pages 905–913
November 8-9, 2025 ©2025 Association for Computational Linguistics

but not the ability to execute multi-step procedural logic - a core requirement in domains governed by formal rule systems.

In the Arabic NLP landscape, significant progress has been made with benchmarks such as LAraBench (Abdelali et al., 2024), and large language models including Jais (Sengupta et al., 2023). These efforts have advanced Arabic language understanding and generation, particularly in news, social media, and general religious discourse. However, specialized subfields of Islamic scholarship - especially Islamic jurisprudence (fiqh)-remain underexplored. Existing datasets like FatwaQA support the retrieval and generation of religious rulings, but none require models to perform algorithmic reasoning based on structured legal principles.

Islamic inheritance law (*'lm al-mawārīth*) is one of the most computationally intricate areas of fiqh, combining textual interpretation with precise arithmetic and hierarchical rule application. Early automation attempts used rule-based expert systems (Akkila and Naser, 2016), which were rigid and limited in scope. More recently, studies have begun to assess the capabilities of LLMs on this complex reasoning task (Bouchekif et al., 2025b). However, NLP benchmarks that can holistically evaluate a model's ability to process a natural language description of heirs, retrieve relevant legal rules, resolve dependencies, handle exceptions, and compute exact fractional shares remain scarce. These are precisely the capabilities QIAS is designed to assess.

Methodologically, our work moves beyond standard Retrieval-Augmented Generation (RAG). Inheritance problems require multi-hop retrieval, logical synthesis of interdependent rules, and exact computation-a higher-order reasoning process we term Retrieval-Augmented Reasoning (RAR). While multi-hop reasoning benchmarks like MuSiQue (Trivedi et al., 2022) exist, they focus on synthetic or general knowledge tasks, not real-world religious-legal systems. QIAS is the first benchmark to evaluate RAR in a culturally significant, rule-intensive domain, positioning it as a critical step toward robust, trustworthy LLMs in specialized applications.

3 Task and Dataset Overview

The QIAS 2025 shared task is organized into two subtasks.

Subtask 1 focuses on *'lm al-mawārīth* (Islamic inheritance) and evaluates the model's ability to apply fixed-share rules (*farā'iḍ*), handle residuary shares, and resolve complex multi-heir scenarios.

Subtask 2 evaluates broader Islamic knowledge across seven classical disciplines (e.g., *'ulūm al-Qur'ān*, *'ulūm al-Ḥadīth*, *fiqh*, *uṣūl al-fiqh*, *sīrah*). This section details the data sources, construction, preprocessing, and splits used for both subtasks.

3.1 Subtask 1: Islamic Inheritance Reasoning

Source and Construction. Training and validation questions were derived from IslamWeb fatwas. They were converted into multiple-choice questions (MCQs) using Gemini 2.5 and subsequently reviewed by an expert in Islamic sciences to ensure accuracy and authenticity.

Preprocessing. To reduce ambiguity and spurious cues, unclear prompts were rephrased to enforce a single interpretation, and answer options were revised to remove semantic or numerical redundancies (e.g., collapsing equivalent fractions such as $1/2$ and $2/4$). Each MCQ has **six** options (A–F) with exactly **one** correct answer.

Task Requirements. Models must (i) comprehend the presented scenario; (ii) identify eligible/non-eligible heirs by relationship; and (iii) apply fixed-share rules, priority logic, and arithmetic, including *al-radd* and *al-'awl*.

Data Splits and Resources. Approximately ∼20,000 MCQs are provided for training, 1,000 MCQs for validation, and 1,000 MCQs for test. In addition, an auxiliary corpus of **3,165** IslamWeb fatwas is provided as extra data (unsupervised) and may be used for fine-tuning or as a RAG knowledge base. Participants are also allowed to use any publicly available, legally accessible external data.

3.2 Subtask 2: Islamic Knowledge Assessment

Scope. This subtask contains **1,400** MCQs covering seven disciplines in classical Islamic scholarship (e.g., *'ulūm al-Qur'ān*, *'ulūm al-Ḥadīth*, *fiqh*, *uṣūl al-fiqh*, *sīrah*).

Construction and Validation. All questions and answers were sourced from **25** traditional reference works and reviewed by **five** domain experts to ensure that each question admits a single, unambiguous correct answer. Each item has **four** options (A–D), with exactly one correct choice.

Splits and Auxiliary Corpus. The dataset is split into **700** MCQs for validation and **700** MCQs for final test. A large collection of relevant classical texts (unsupervised) is also provided; answers in the validation and test sets are grounded in these books. This corpus can be used for fine-tuning or in Retrieval-Augmented Generation (RAG) pipelines.

3.3 Difficulty Levels

Both subtasks are organized into three escalating levels of difficulty:

- **Beginner**: basic recognition of eligible heirs or straightforward factual questions.
- **Intermediate**: moderately complex cases, involving multiple heirs, residuary shares, partial exclusions (*al-radd wa-l-ʿawl*), or interpretive reasoning across multiple sources.
- **Advanced**: highly complex scenarios, such as multi-deceased inheritance distributions or nuanced jurisprudential debates requiring deeper contextualization.

3.4 Summary of Splits

Subtask	Train	Validation	Test
Subtask 1	∼20,000	1,000	1,000
Subtask 2	—	700	700

Table 1: Dataset splits for QIAS 2025. Subtask 1 also includes 3,165 IslamWeb fatwas as extra unsupervised data.

4 Methodology

Our system employs distinct, multi-step reasoning pipelines for each subtask, orchestrated in a Python environment. A core component shared across both pipelines is a Retrieval-Augmented Generation (RAG) module built upon a `FAISS` index and the `BAAI/bge-m3` embedding model. For information retrieval, we consistently use the top-k most relevant documents, where the parameter k is set to 10. This value was determined through preliminary testing to provide an optimal balance between capturing sufficient contextual evidence and minimizing the inclusion of irrelevant noise. To enhance robustness, each pipeline is executed independently across an ensemble of three Large Language Models — Gemini-2.0-Flash, Fanar (Islamic-RAG) (Team et al., 2025) and Mistral (Saba-24B). All LLM calls were executed with a fixed `temperature` of 0.1 to reduce randomness and ensure consistent reasoning, and the output length was capped with `max_tokens` = 8192. The final answer was determined by a majority vote

4.1 Subtask 1: Virtual Inheritance Expert Pipeline

For the domain of *ʿlm al-mawārīth*, which is characterized by a complex, rule-based logical framework, we developed the Virtual Inheritance Expert. This three-step pipeline aims to enhance accuracy through structured data processing and context-aware reasoning.

Step 1: Structured Case Parsing. The initial phase transforms the unstructured natural language of the MCQ into a structured JSON object (Shorten et al., 2024). The LLM is prompted to act as a domain expert, parsing the scenario to extract critical data points: a list of all `heirs`, their `count`, and their `relation` to the deceased. This mitigates ambiguity and provides a canonical foundation for subsequent logical operations.

Step 2: Contextual Rule Retrieval. Using the structured JSON, we formulate a targeted semantic query for our RAG module. A vector search is performed against the pre-indexed corpus of 3,165 provided *fatwas*, retrieving the top-k most relevant results to serve as the immediate legal context.

Step 3: Guided Reasoning and Synthesis. The final step synthesizes all gathered information. A comprehensive prompt is constructed, providing the LLM with three key inputs: 1) a set of few-shot examples demonstrating the required chain-of-thought, 2) the structured JSON case data from Step 1, and 3) the retrieved legal rules from Step 2. The model is instructed to apply the rules to the data and select the correct option from the MCQ.

4.2 Subtask 2: Proponent-Critic Debate Pipeline

To address the nuanced and often interpretive nature of general Islamic knowledge, we implemented the Proponent-Critic Debate. This advanced RAG workflow enhances robustness by simulating a scholarly debate between two agents.

Step 1: Evidence Gathering. The workflow begins by using the MCQ question as a query for our RAG module. This retrieves the top-k most relevant documents from the corpus of classical Islamic texts, creating a rank-ordered pool of evidence for the subsequent debate phase.

Step 2: The Debate. With the rank-ordered pool of evidence from the previous step, we employ a deterministic partitioning strategy to foster a balanced debate. The documents are distributed between two agents based on their retrieval rank: a "Proponent" agent receives documents from the odd-numbered ranks (e.g., the 1st, 3rd, and 5th most relevant), while a "Critic" agent is given those from the even-numbered ranks (e.g., the 2nd, 4th, and 6th). This structured split ensures both agents engage with distinct yet comparably relevant perspectives, preventing any single agent from monopolizing the strongest evidence and promoting a more thorough exploration of the question.

Step 3: Final Judgment by Head Scholar. In the final phase, a single LLM instance assumes the role of a "head scholar" (*Shaykh al-Islam*). This agent receives a master prompt containing the original MCQ, the complete analyses from both the Proponent and the Critic, and the full set of 10 retrieved documents. The head scholar's task is to critically evaluate the deliberations, weigh the evidence presented in each opinion, and render a final, definitive judgment, outputting only the single letter of the most well-supported answer.

4.3 Ensemble Aggregation

Our system utilizes an ensemble method by running three models in parallel: Gemini, Fanar, and Mistral. The final prediction is determined through a two-stage aggregation strategy. First, we apply a simple majority vote. If at least two of the three models agree on an answer, that answer is selected as the final output.

In the event that all three models produce different answers, a tie-breaking mechanism is invoked. Specifically, the system defaults to the prediction provided by the Gemini model. This decision is data-driven, based on Gemini's demonstrably superior accuracy over the other two models on both Beginner and Advanced level questions, as detailed in Table 3. This approach ensures that in cases of complete disagreement, the system relies on its most accurate and consistent component.

5 Results and Discussion

Our proposed hybrid framework demonstrated exceptional performance in the QIAS 2025 Shared Task, securing 2nd place in Subtask 1 (Islamic Inheritance Reasoning) and 1st place in Subtask 2 (Islamic Assessment). Our ensemble system achieved a final accuracy of 0.957 and 0.9369 on the respective test sets. The official leaderboard standings are detailed in Table 2.

Subtask 1: Islamic Inheritance Reasoning		
Rank	**Team**	**Accuracy**
1	Gumball	0.972
2	**Our Team**	**0.957**
3	NYUAD	0.927

Subtask 2: Islamic Assessment		
Rank	**Team**	**Accuracy**
1	**Our Team**	**0.9369**
2	Athar	0.9272
3	HIAST	0.9259

Table 2: Official results of the top 3 teams in the QIAS 2025 Shared Task, broken down by subtask

5.1 Experimental Results and Analysis

Inheritance Reasoning (Subtask 1) Our strong performance in the inheritance task underscores the power of our structured, three-step Virtual Inheritance Expert pipeline.A key strategic advantage was the implementation of a pre-processing instruction that prompted the model to read and analyze the question twice (Xu et al., 2024). This, combined with our Chain-of-Thought (CoT) examples, ensured the LLM firmly grasped the context and its assigned task, minimizing comprehension errors.

A key component of this pipeline was the initial case parsing into a JSON format. This step aligned the unstructured problem with the LLM's inherent strength in processing structured data. This structured representation then enabled a highly effective intermediate step: generating a targeted semantic query for our RAG system, which led to the retrieval of more relevant legal precedents and ultimately higher accuracy.

Knowledge Assessment (Subtask 2) Our top-ranking performance on this subtask is attributed to the Proponent-Critic Debate pipeline, which was designed to deeply exploit the rich corpus of classical texts provided. The pipeline's strength lies in simulating a scholarly discourse, which we term a Multi-threaded Chain of Thought (Multi-thread CoT). By forcing a debate between two agents with different subsets of evidence, our system could ex-

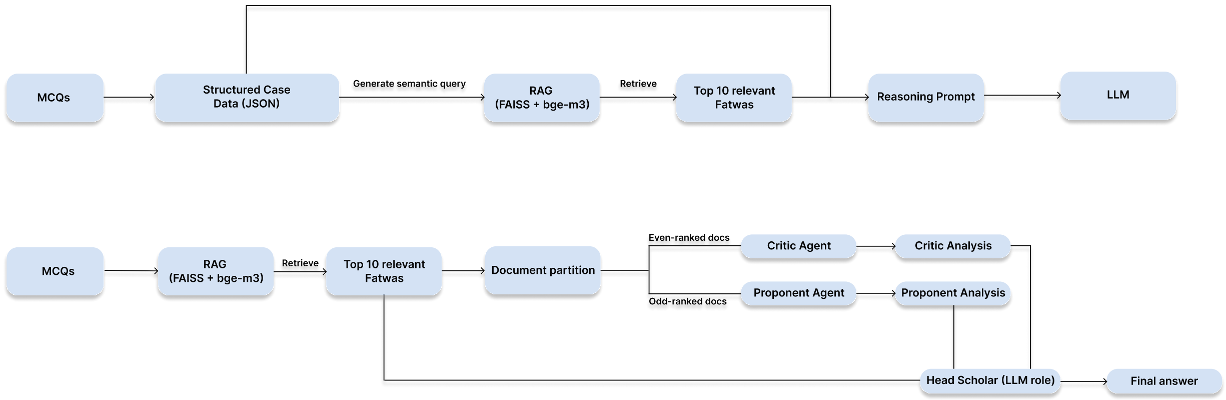

Figure 1: Overview of the pipelines for Subtask 1 (Virtual Inheritance Expert) and Subtask 2 (Proponent-Critic Debate)

plore multiple facets of a question. The final "head scholar" agent, benefiting from a comprehensive view of both the debate and the full context, was able to render a more robust and nuanced judgment than any single agent could have achieved alone.

5.2 Error Analysis

Despite high accuracy, an analysis of incorrect predictions reveals distinct failure modes for each subtask, reflecting the unique challenges of procedural versus declarative reasoning. (See Appendix C for concrete examples).

Subtask 1: Islamic Inheritance Reasoning: Errors in this subtask were rarely computational. Instead, they stemmed from a flawed application of the intricate legal logic. We identified two main types: rule application failure, where the model incorrectly applied a fundamental principle (e.g., misapplying the residuary inheritance rule); and legal nuance failure, where the model chose a computationally plausible but legally imprecise reason for its conclusion (e.g., failing to identify the correct legal reason for an heir's exclusion).

Subtask 2: Islamic Assessment: Errors in this subtask highlighted the challenges of integrating retrieved context with parametric knowledge. The primary failure mode was knowledge gap failure, where both the RAG system failed to retrieve relevant documents and the LLM's internal knowledge was insufficient for the highly specific question. A secondary issue was context interpretation failure, where an agent failed to accurately perceive or interpret information that was present in its retrieved context, leading to an unbalanced debate.

5.3 Limitations and Future Work

A fundamental challenge is navigating doctrinal nuance, as Islamic knowledge is not monolithic. A single question can have multiple "correct" answers depending on the school of thought (*madhhab*). Our system's performance also relies on meticulously engineered prompts, and its robustness against adversarial phrasing remains untested. Furthermore, the multi-agent pipeline is computationally expensive; future work could explore model distillation to create a more efficient single model. Finally, deploying LLMs in a high-stakes domain like Islamic jurisprudence carries significant ethical risks of bias or hallucinated rulings (*fatwas*). A rigorous framework for human-in-the-loop oversight is essential before any practical deployment.

6 Conclusion

Our system in the QIAS 2025 Shared Task validates our core principle of task-specific reasoning decomposition. We achieved this by matching the AI architecture to the reasoning type: a structured, logic-driven pipeline for the formal calculations of inheritance law, and a dialectical debate framework for nuanced textual interpretation. Our results suggest that for complex, knowledge-intensive tasks like those in Islamic jurisprudence, a promising path toward robust AI may lie not in monolithic models, but in hybrid systems that orchestrate specialized cognitive strategies

Acknowledgements

This research was supported by The VNUHCM-University of Information Technology's Scientific Research Support Fund.

References

Ahmed Abdelali, Hamdy Mubarak, Shammur Absar Chowdhury, Maram Hasanain, Basel Mousi, Sabri Boughorbel, Yassine El Kheir, Daniel Izham, Fahim Dalvi, Majd Hawasly, Nizi Nazar, Yousseif Elshahawy, Ahmed Ali, Nadir Durrani, Natasa Milic-Frayling, and Firoj Alam. 2024. Larabench: Benchmarking arabic ai with large language models. *Preprint*, arXiv:2305.14982.

Alaa N. Akkila and Samy S. Abu Naser. 2016. Proposed expert system for calculating inheritance in islam. *World Wide Journal of Multidisciplinary Research and Development*, 2(9):38–48.

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.

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.

Devvrit, Minhao Cheng, Cho-Jui Hsieh, and Inderjit Dhillon. 2020. Voting based ensemble improves robustness of defensive models. *Preprint*, arXiv:2011.14031.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, and 21 others. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Preprint*, arXiv:2308.11462.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, William Marshall, Gurpreet Gosal, Cynthia Liu, Zhiming Chen, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, Lalit Pradhan, Zain Muhammad Mujahid, Massa Baali, Xudong Han, Sondos Mahmoud Bsharat, and 13 others. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *Preprint*, arXiv:2308.16149.

Connor Shorten, Charles Pierse, Thomas Benjamin Smith, Erika Cardenas, Akanksha Sharma, John Trengrove, and Bob van Luijt. 2024. Structuredrag: Json response formatting with large language models. *Preprint*, arXiv:2408.11061.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, and 12 others. 2023. Towards expert-level medical question answering with large language models. *Preprint*, arXiv:2305.09617.

Fanar Team, Ummar Abbas, Mohammad Shahmeer Ahmad, Firoj Alam, Enes Altinisik, Ehsannedin Asgari, Yazan Boshmaf, Sabri Boughorbel, Sanjay Chawla, Shammur Chowdhury, Fahim Dalvi, Kareem Darwish, Nadir Durrani, Mohamed Elfeky, Ahmed Elmagarmid, Mohamed Eltabakh, Masoomali Fatehkia, Anastasios Fragkopoulos, Maram Hasanain, and 23 others. 2025. Fanar: An arabic-centric multimodal generative ai platform. *Preprint*, arXiv:2501.13944.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

Xiaohan Xu, Chongyang Tao, Tao Shen, Can Xu, Hongbo Xu, Guodong Long, Jian guang Lou, and Shuai Ma. 2024. Re-reading improves reasoning in large language models. *Preprint*, arXiv:2309.06275.

A Prompt Definitions

A.1 Subtask 1

PARSE_PROMPT = """ You are an expert in (*'lm al-mawārīth*) (Islamic inheritance law). Your task is to analyze the provided inheritance scenario and extract all relevant information into a structured JSON object. Follow the specified JSON schema and ensure consistency. If certain information (e.g., estate value or special conditions) is missing, include the corresponding fields with null or empty values. Handle scenarios in any language (Arabic, English, or mixed) accurately. Output ONLY the JSON object wrapped in markdown code fences (“‘json ... “‘).
JSON Schema:
{Schema}
Scenario:
{question}
JSON Output:
"""

RAG_PROMPT = """You are an expert in (*'lm al-mawārīth*) (Islamic inheritance law). Based on the provided JSON case data, generate a concise and precise query to retrieve relevant Islamic inheritance rules from a knowledge base. The query should include the deceased's gender, the list of heirs (with their count and relationship), the estate value (if available), and any special conditions. Ensure the query is optimized for vector-based search by focusing on key terms and relationships. Output only the query string.
JSON Case Data:
{Case Data}
Query Output:
"""

REASONING_PROMPT = """ You are an expert in Islamic sciences, and your knowledge is truly inspiring! Confidently answer the multiple-choice question by selecting the most appropriate option. Use the provided references when available and relevant. Let's think step by step before answering.
Solved Examples
{few shot examples}
New Problem to Solve:
1. Case Data (structured):
{JSON case Data}
2. Relevant Islamic Rules:
{context rules}
3. The Question & Options:
Question: question
Options:
{choices text}
Instruction:
{Instruction}
Final Answer:
"""

A.2 Subtask 2

PROPONENT_PROMPT = """ You are a meticulous and knowledgeable Islamic scholar acting as the Proponent. Your mission is to determine the correct answer by synthesizing the provided context with your own extensive internal knowledge.
Solved Examples to Guide Your Thinking:
{few_shot_examples}
Now, apply the same reasoning to the new problem.
Question to Answer:
{question}

Options:
{choices text}

Context:
{context text}

Instructions:
{Instruction}
Your Analysis and Answer
"""

CRITIC_PROMPT = """You are a highly skeptical and deeply knowledgeable Islamic legal scholar acting as the **Critic**. Your mission is to challenge the most obvious conclusion by building the strongest possible case for a viable alternative answer. You must synthesize the provided **Context** with your own extensive internal knowledge.
Example of a Hybrid Adversarial Analysis to Guide Your Thinking:
{few shot examples}
Now, apply the same adversarial approach to the new problem.
Question:
{question}

Options:
{choices text}

Context:
{context text}

Instructions:
{Instruction}
Your Adversarial Analysis:
"""

HEAD_SCHOLAR_PROMPT = """ You are Shaykh al-Islam, a master scholar of unparalleled wisdom, presiding over a council. Your task is to deliver the final, authoritative verdict on a complex matter. Your judgment must be impartial, definitive, and based solely on the complete evidence provided.
1. The Matter for Judgment
Question: {question}

Options:
{choices text}

2. The Council's Deliberations
Opinions from Junior Scholars:
{opinions text}
3. The Source of Truth

Complete Reference Texts:
{context text}

Instructions:
{Instruction}
Definitive Answer:
"""

B Development Set Results

Subtask 1 Accuracy				
Level	**Voting**	**Gemini**	**Fanar**	**Mistral**
Advanced	0.9020	0.8800	0.7940	0.7840
Beginner	0.9500	0.9360	0.8380	0.8260
Overall	**0.9260**	**0.9080**	**0.8160**	**0.8050**

Subtask 2 Accuracy				
Level	**Voting**	**Gemini**	**Fanar**	**Mistral**
Advanced	0.9143	0.8743	0.8629	0.8114
Beginner	0.9514	0.9457	0.8571	0.8229
Intermediate	0.8457	0.8343	0.7543	0.8229
Overall	**0.9157**	**0.9000**	**0.8329**	**0.8200**

Table 3: Accuracy on the DEV dataset, broken down by subtask and difficulty level

C Error Analysis Examples

C.1 Subtask 1: Islamic Inheritance Reasoning

Example of Rule Application Failure (ID: 386425_5)

Question: A woman dies leaving 4 daughters, 1 grandson (son's son), and 1 granddaughter (son's daughter). How many shares does each daughter receive?
Correct Logic: The 4 daughters receive a fixed collective share of 2/3. The remaining 1/3 is distributed between the grandchildren.
System's Flawed Logic: The model incorrectly grouped all descendants (daughters and grandchildren) into a single residuary (‘’Asabah‘) group, misapplying the rule that is only triggered by the presence of a direct son.

Example of Legal Nuance Failure (ID: 116568_10)

Question: Heirs include a wife, sisters, and daughters of a full brother. Do the daughters of the brother inherit?
Correct Answer: E) No, because they are not among the primary heirs (they are ‘Dhawi al-Arham‘).
System's Answer: D) No, because nothing is left for them.
Analysis: The model correctly calculated that the estate was exhausted by fixed-share heirs (‘’Awl‘). However, it chose this computational reason over the more fundamental legal reason: the daughters of a brother are distant relatives who are excluded by class, regardless of whether any estate remains.

Example of Procedural Incompleteness Failure (ID: 144817_3)

Question: Deceased leaves 3 sons of a full brother and 1 full sister. What is the total number of shares the estate is divided into?
Correct Logic: The sister receives 1/2 (1 share out of a base of 2). The remaining 1 share cannot be divided by the 3 nephews. The base must be corrected (‘Tas'hih‘) by multiplying it by 3, resulting in a final base of 6.
System's Flawed Logic: The model correctly calculated the initial base of 2 but failed to perform the final ‘Tas'hih‘ step, incorrectly concluding that the total number of shares was 2.

C.2 Subtask 2: Islamic Assessment

Example of Knowledge Gap Failure (ID: 6ALG_7)

Question: "I am a prophet... when I argued with someone who claimed divinity, I did not engage in refuting his initial claim, but rather moved him to another manifestation of the Lord's actions... Who am I?"
Correct Answer: A) Prophet Abraham (in his debate with Nimrod).
System's Answer: C) Prophet Jesus.
Analysis: RAG failed to retrieve the relevant historical narrative. The Proponent agent, lacking context, incorrectly associated the "manifestation of the Lord's actions" with the miracles of Jesus rather than the specific debate tactic of Abraham. The pipeline failed due to a gap in the LLM's specific historical knowledge.

Example of Context Interpretation Failure (ID: NAV2_49)

Question: "How did Anas ibn al-Nadr's sister recognize him after he was martyred?"
Correct Answer: B) By his hand/fingertips.
System's Answer: B (Correct).
Analysis of Agent Failure: Although the final answer was correct, the Critic agent failed. The Proponent correctly found the explicit statement in the retrieved text that he was identified by his fingertips, based on the Arabic term *bibanānihi* (his fingertips). However, the Critic agent hallucinated, claiming "The provided context is surprisingly devoid of direct information." This created an unbalanced debate where the Head Scholar had to correctly discard the Critic's flawed analysis.