QIAS 2025: Overview of the Shared Task on Islamic Inheritance Reasoning and Knowledge Assessment

Abdessalam BOUCHEKIF[1], **Samer RASHWANI**[1], **Emad MOHAMED**[2], **Mutaz AL-KHATIB**[1], **Heba SBAHI**[1], **Shahd GABEN**[1], **Wajdi ZAGHOUANI**[3], **Aiman ERBAD**[4], **Mohammed GHALY**[1]

[1]**Hamad Bin Khalifa University, Qatar** [2]**Nazarbayev University, Kazakhstan**
[3]**Northwestern University, Qatar** [4]**Qatar University, Qatar**
abouchekif, srashwani, mghaly, malkhatib, sgaben, hsbahi@hbku.edu.qa
emad.mohamed@nu.edu.kz wajdi.zaghouani@northwestern.edu aerbad@qu.edu.qa

Abstract

This paper provides a comprehensive overview of the QIAS 2025 shared task, organized as part of the ArabicNLP 2025 conference and co-located with EMNLP 2025. The task was designed for the evaluation of large language models in the complex domains of religious and legal reasoning. It comprises two subtasks: *(1)* Islamic Inheritance Reasoning, requiring models to compute inheritance shares according to Islamic jurisprudence, and *(2)* Islamic Knowledge Assessment, which covers a range of traditional Islamic disciplines. Both subtasks were structured as multiple-choice question answering challenges, with questions stratified by varying difficulty levels. The shared task attracted significant interest, with 44 teams participating in the development phase, from which 18 teams advanced to the final test phase. Of these, 6 teams submitted entries for both subtasks, 8 for Task 1 only, and two for Task 2 only. Ultimately, 16 teams submitted system description papers. Herein, we detail the task's motivation, dataset construction, evaluation protocol, and present a summary of the participating systems and their results.

1 Introduction

The emergence of Large Language Models (LLMs) has transformed NLP, enabling state-of-the-art performance in tasks requiring deep linguistic understanding, complex reasoning, and coherent text generation. Trained on large-scale general-purpose corpora, LLMs have demonstrated strong performance across a variety of benchmarks, including question answering, summarization, and dialogue. However, LLMs still face challenges in specialized domains, particularly those requiring high information accuracy, and sensitivity to cultural or religious contexts. In the Islamic contexts, LLMs must reason over authoritative and structured sources such as the Qur'an, Hadith, and fatwas. They must also consider differences in interpretation across schools of thought, including variations within Sunni Islam across the four major legal schools: Ḥanafī, Mālikī, Shāfi'ī, and Ḥanbalī.

To evaluate LLMs' capabilities in both Islamic legal reasoning and specialized religious knowledge, we introduce the QIAS 2025 Shared Task. This benchmark presents a diverse set of question-answering challenges across multiple domains, difficulty levels, and jurisprudential perspectives. The task includes two subtasks: (1) Islamic Inheritance Reasoning, which requires precise, rule-based reasoning grounded in classical Islamic jurisprudence. Task 2 focuses on general Islamic knowledge, incorporating questions curated by experts from key disciplines. Each question is labeled by difficulty and assesses knowledge of religious concepts, legal reasoning, and interpretive differences.

In this paper, we present an overview of the QIAS 2025[1] Shared Task, which represents an important step toward developing NLP models capable of addressing complex challenges in Islamic knowledge. This includes inheritance calculation tasks requiring precise reasoning and rule-based computation grounded in Islamic jurisprudence. To our knowledge, no previous dataset has been specifically designed for fine-tuning models on Islamic inheritance reasoning at this scale. The second task focuses on question answering covering diverse areas of Islamic scholarship. Unlike many existing datasets relying on general cultural or surface-level questions, our dataset is curated and annotated by domain experts to reflect a deeper understanding of jurisprudential and theological concepts.

2 Related Work

Recent LLMs such as GPT-4 (Achiam et al., 2023), Gemini2.5 (Comanici et al., 2025), and DeepSeek-R1 (Guo et al., 2025) have achieved state-of-the-art performance across diverse stan-

[1] https://sites.google.com/view/qias2025/

Proceedings of The Third Arabic Natural Language Processing Conference, pages 851–860
November 8-9, 2025

dard NLP benchmarks. In parallel, several Arabic-focused LLMs have been developed to better capture linguistic, cultural, and domain-relevant needs of Arabic-speaking communities, including Falcon(Almazrouei et al., 2023), Jais (Sengupta et al., 2023), AceGPT(Huang et al., 2023), ArabianGPT (Koubaa et al., 2024), ALLaM (Bari et al., 2024), and Fanar (Abbas et al., 2025). These efforts have motivated growing interest in applying LLMs to tasks involving Islamic content and knowledge.

The application of LLMs to Islamic texts has recently gained increasing attention within the NLP community. (Malhas et al., 2022) (Malhas et al., 2023) organized shared tasks focused on advancing Islamic information retrieval, with a particular focus on understanding Qur'anic passages. These tasks included a Qur'anic passage retrieval task—requiring models to retrieve relevant verses from the Qur'an given a question, and a reading comprehension task, where expected to extract accurate answers from a provided passage. More recently, (Sayeed et al., 2025) explored QA systems for *ibb nabawī* (Prophetic medicine) using LLaMA-3, Mistral-7B, and Qwen-2 combined with RAG, while (Alan et al., 2024) proposed MufassirQAS, a RAG-based system trained on Turkish Islamic texts to improve transparency and reduce hallucinations in religious QA. (Rizqullah et al., 2023) introduced QASiNa QA dataset, derived from *Sirah Nabawiyah* texts in Indonesian, comparing traditional multilingual transformers (XLM-R, mBERT, IndoBERT) with GPT-3.5 and GPT-4. (Qamar et al., 2024) introduced a dataset of 73,000 question–answer pairs has been introduced, focusing on non-factoid QA for Quranic Tafsir and Hadith. The study revealed a critical gap between automatic evaluation metrics (such as ROUGE) and human judgments. These results show that automatic evaluation metrics alone are not sufficient, and highlight the need for more robust evaluation methods that can better reflect the complexity and interpretive nature of Islamic religious texts. In (Aleid and Azmi, 2025), the authors released Hajj-FQA, a benchmark of 2,826 QA pairs extracted from 800 expert-annotated fatwas concerning the Hajj pilgrimage. Despite these efforts, several studies have identified significant limitations in this LLMs. For instance, (Mohammed et al., 2025) show that even advanced models like GPT-4 tend to produce factually incorrect or misleading answers when applied to Islamic content. They identify three main issues: *(i)* misinterpretation of religious context, *(ii)* generation of answers that are unclear or not based on reliable Islamic sources like the Qur'an or Hadith, and *(iii)* high sensitivity to slight variations in question phrasing, leading to inconsistent responses. Similarly, (Alnefaie et al., 2023) observed that GPT-4 has difficulty answering Quranic questions accurately, due to difficulties with classical arabic, semantic ambiguity, and misinterpretation of contextual meaning.

Early research on automating Islamic inheritance began with expert systems focused on calculating basic inheritance shares (Akkila and Naser, 2016). Later works incorporated intricate adjustments such as *ḥajb*, ʿawl, and radd (Tabassum et al., 2019). (Zouaoui and Rezeg, 2021) proposed a Arabic ontology for identifying heirs and d calculating their inheritance shares (Tabassum et al., 2019). Most recently, (Bouchekif et al., 2025) evaluated seven LLMs on Islamic inheritance. The results reveal that models with strong reasoning capabilities, such as Gemini 2.5 and o3, achieved high performance, with accuracy rates of 90.6% and 93.4%, respectively. In contrast, models lacking advanced reasoning abilities—such as Jais, Mistral, and LLaMA—performed significantly worse, with accuracy rates below 50%, highlighting their limitations in handling complex legal reasoning tasks.

3 Task1: Islamic Inheritance Reasoning

3.1 Task Description

The task1 focuses on the domain of *'lm* al-mawārīth, the Islamic science of inheritance. The goal is to assess the ability of LLMs to accurately apply Islamic inheritance rules in realistic scenarios. Solving inheritance problems requires a combination of cognitive, legal, and computational skills, including:

1. Identifying familial relationships and considering legal conditions such as debts, bequests, and the sequence of deaths among relatives.
2. Determining eligible heirs, including fixed-share heirs (*aṣḥāb al-furūḍ*) and residuaries (*'aṣabāt*), and correctly applying exclusion rules (*ḥajb*) based on valid justifications and authentic scriptural evidence.
3. Computing shares by deriving a common denominator and adjusting the distribution when necessary:
   - *Radd* (redistribution) is used when a surplus remains after initial allocation. This surplus is proportionally redistributed among

the heirs, excluding spouses. — *Example:* Wife $(1/4)$ and full sister $(1/2)$, leaving a surplus of $1/4$; after redistribution, the wife receives $(1/4)$ and the sister receives $(3/4)$.
- ʿAwl (proportional reduction) is applied when the sum of assigned shares exceeds the estate. All shares are scaled down proportionally. — *Example:* Father $(1/6)$, mother $(1/6)$, wife $(1/8)$, and four daughters $(2/3)$; the total exceeds 1. The denominator is adjusted to 27, and then the wife receives $3/27 = 1/9$.

4. Addressing complex and exceptional cases, such as consecutive death (*munāsakha*) or juristic disputes like the *Akdariyya* case involving grandparents and siblings.
5. Numerical precision in the final distribution, including the correct adjustment and fractional allocation[2].

3.2 Data

The dataset contains 22,000 MCQs, including 10,446 generated from IslamWeb fatwas and 11,554 constructed from inheritance case resolutions using the calculator of the *Almwareeth* website[3], offers a specialized tool that algorithmically solves all types of mirath (Islamic inheritance) problems. The IslamWeb-based MCQs were derived from Islamic religio-ethical rulings (fatwas)[4], which were automatically converted into question-answer format using Gemini 2.5 Pro. Each generated question was then reviewed by four experts in Islamic studies to ensure both legal soundness and linguistic clarity. As part of the preprocessing phase, ambiguous questions were rephrased to guarantee a single, unambiguous interpretation. The answer choices were also revised to eliminate semantic and numerical redundancies, such as equivalent options (*e.g* $1/2$ and $2/4$). The dataset has two levels of difficulty: **Beginner** and **Advanced**, reflecting increasing complexity in both legal reasoning and mathematical computation.

Participants are also provided a collection of 3,165 fatwas (question–answer pairs) from IslamWeb is available. These fatwas cover a broad spectrum of Islamic legal, ethical, and social issues and can serve as a valuable supplementary knowledge base.

[2]For more details about the terminology and rules of Islamic inheritance law, see "*Irth*," in *Al-Mawsūʾa al-Fiqhiyya* (The Kuwaitan Encyclopedia of Fiqh). Kuwait: *Wazārat al-Awqāf* wa-al-Shuʾūn al-Islamiyya. 45 Vols. 1984-2007. Vol. 3, Pp. 17-79.

[3]https://almwareeth.com/

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

Example – Level Beginner

توفي عن أب، و2 أخ شقيق، و1 ابن أخ شقيق، و2 عم شقيق للأب، وأم، و2 بنت، و1 زوجة، ما هو نصيب الأم؟

He was survived by his father, two full brothers, one nephew (son of a full brother), two paternal uncles, his mother, two daughters, and his wife. What is the share of the mother?

(One-third)	الثلث	□
(One-quarter)	الربع	□
(One-sixth)	السدس	■
(One-eighth)	الثمن	□
(One-half)	النصف	□
(Nothing)	لا شيء	□

Example – Level Advanced

توفي عن زوجة وبنتين وأخ شقيق، والتركة 12000 درهم. ما هو النصيب النهائي لكل وارث من التركة؟

He was survived by his wife, two daughters, and one full brother. The estate is 12,000 dirhams. What is the final share of each heir from the estate?

■ الزوجة: 1500 درهم، البنتان: 8000 درهم، الأخ الشقيق: 2500 درهم

Wife: 1500 dirhams, Two daughters: 8000 dirhams, Full brother: 2500 dirhams

□ الزوجة: 3000 درهم، البنتان: 8000 هم، الأخ الشقيق: 1000 درهم

Wife: 3000 dirhams, Two daughters: 8000 dirhams, Full brother: 1000 dirhams

□ الزوجة: 1500 درهم، البنتان: 6000 درهم، الأخ الشقيق: 4500 درهم

Wife: 1500 dirhams, Two daughters: 6000 dirhams, Full brother: 4500 dirhams

□ الزوجة: 1500 درهم، البنتان: 8000 درهم، الأخ الشقيق: 3000 درهم

Wife: 1500 dirhams, Two daughters: 8000 dirhams, Full brother: 3000 dirhams

□ الزوجة: 2000 درهم، البنتان: 7500 درهم، الأخ الشقيق: 2500 درهم

Wife: 2000 dirhams, Two daughters: 7500 dirhams, Full brother: 2500 dirhams

□ الزوجة: 1000 درهم، البنتان: 8500 درهم، الأخ الشقيق: 2500 درهم

Wife: 1000 dirhams, Two daughters: 8500 dirhams, Full brother: 2500 dirhams

4 Task2: Islamic Assessment

4.1 Task Description

The task2 evaluates general Islamic knowledge across a wide range of topics within Islamic knowledge, including ʿulūm al-Qurʾān (Quranic studies), ʿulūm al-Ḥadīth (hadith criticism), *fiqh* (jurisprudence), uṣūl al-fiqh (legal theory), *sīrah* (Prophetic Biography). It is organized into three progressively

Task	Split	Levels			Total
		Beg.	Int.	Adv.	
Task 1					
	Training	10000	—	10000	20000
	Dev	500	—	500	1000
	Test	500	—	500	1000
	Total	**11000**	—	**11000**	**22000**
Task 2					
	Training	—	—	—	—
	Dev	350	175	175	700
	Test	700	150	150	1000
	Total	**1050**	**325**	**325**	**1700**

Table 1: Unified distribution of MCQs across dataset splits and difficulty levels for Task 1 (Inheritance Reasoning) and Task 2 (Islamic Knowledge Assessment). "—" indicates not available.

challenging difficulty levels: beginner, intermediate, and advanced.

4.2 Data

The dataset was constructed from collection of 25 relevant classical Islamic books that are widely recognized by scholars as authoritative. It consists of 1,400 MCQs (700 for training and 700 for testing), all rigorously reviewed and validated by five experts in Islamic studies. Each question has been carefully designed to elicit a single, unambiguous correct answer, thereby ensuring clarity and consistency in the evaluation process.

The answers to the MCQs in the validation and test sets are derived from a selection of classical Islamic texts, which we provide to participants. As such, this corpus can be leveraged either as part of a Retrieval-Augmented Generation (RAG) system to enhance the model's ability to generate accurate and contextually grounded responses, or to fine-tune language models on Islamic studies.

Example of MCQ Level Beginner

ما مدة المسح على الخفين للمقيم؟

What is the duration of wiping over the leather socks for a resident?

- ■ يوم وليلة — One day and one night
- □ ثلاثة أيام بلياليهن — Three days and their nights
- □ يومان وليلتان — Two days and two nights
- □ أسبوع كامل — A full week

Example of MCQ Level Intermediate

من شروط الأصل في القياس؟

Which of the following is a condition for the base case (al-aṣl) in analogical reasoning (*qiyās*)?

- □ أن يكون الأصل فرعًا لأصلٍ آخر
  That the base case (al-aṣl) is itself a branch (farʿ) of another base case.
- □ ألا يكون الحكم ثابتًا في الأصل بطريقٍ سمعيٍّ شرعي
  That the ruling in the base case is *not* established by a revealed textual proof.
- ■ ألا يكون الأصلُ فرعًا لأصلٍ آخر
  That the base case (al-aṣl) is *not* a branch (farʿ) of another base case.
- □ ألا تُعرَف طريقةُ الاستنباط
  That the method of derivation is unknown.

Example of MCQ Level Advanced

ما هو طريق الحكماء لإثبات وجود الواجب؟

What is the method of the philosophers to prove the existence of the necessary Being (*al-Wājib*)?

- □ عن طريق اعتبار العالم قديمًا.
  By positing the world as eternal.
- □ عن طريق إثبات أن العالم واجب لذاته.
  By claiming the world is necessary in itself.
- ■ عن طريق امتناع التسلسل والدور.
  By the impossibility of infinite regress (*tasalsul*) and circular causation.
- □ عن طريق إثبات حدوث العالم.
  By demonstrating that the world is originated.

Team Name	Task 1	Task 2	Affiliations
Gumball (Elrefai et al., 2025)	✔	✔	Alexandria University, Ain Shams University, Benha University
PuxAI (Phuc and Đặng Văn, 2025)	✔	✔	VNU□HCM University of Information Technology
NYUAD (AlDahoul and Zaki, 2025)	✔		New York University Abu Dhabi
HIAST (Hamed et al., 2025)	✔	✔	Higher Institute for Applied Sciences and Technology
MorAI (R'baiti et al., 2025)	✔		Mohammed VI Polytechnic University
CVPD (Bekhouche et al., 2025)	✔		University of the Basque Country, Sorbonne University Abu Dhabi
QU-NLP (AL-Smadi, 2025)	✔	✔	Qatar University
CIS-RG (Zaki et al., 2025)	✔		Sinai University
ANLPers (Sibaee et al., 2025)	✔	✔	Prince Sultan University
Athar (Noureldien et al., 2025)	✔	✔	University of Khartoum, University Malaysia
SHA (Altammami, 2025)	✔		King Saud University
SEA (Alowaidi et al., 2025)	✔		University of Leeds
HAI (Hossain and Afli, 2025)	✔		ADAPT Centre
IWAN	✔		King Saud University
Transform_Tafsir (Abu Ahmad et al., 2025)	✔		University of Osnabrück, German Research Center for Artificial Intelligence
N&N (Alangari and AlShenaifi, 2025)		✔	King Saud University
Teams60		✔	MBZUAI
Tokenizers United (Samy et al., 2025)		✔	Nile University, Ain Shams University

Table 2: The participating teams: tasks and affiliations.

5 Results and Discussion

A total of 17 teams participated in the Test phase. Among these, 6 teams submitted systems for both subtasks, 7 teams participated in Task 1 only, and 2 teams in Task 2 only. Table 2 summarizes the participating teams and their affiliations. The Dev phase lasted approximately one and a half months, followed by a 5-day test phase. During the test phase, participants made a total of 127 submissions for Task 1 and 50 for Task 2. We use accuracy to evaluate models, calculated as the percentage of questions for which the model's prediction exactly matches the correct answer. We provide a baseline implementation using Fanar, a modern Arabic large language model accessible via API. This baseline relies exclusively on prompting techniques, without any fine-tuning. The goal is to provide a simple yet effective reference point for evaluating model performance. The dataset and baseline code are publicly available. [5]

5.1 Participating Teams and Results

Table 3 presents the leaderboard rankings and accuracy scores for both subtasks. In Subtask 1 (Islamic Inheritance Reasoning), the best-performing sys-

[5] https://gitlab.com/islamgpt1/qias_shared_task_2025

Task 1			Task 2		
Rank	**Team**	**Accuracy**	**Rank**	**Team**	**Accuracy**
1	Gumball	0.972	1	PuxAI	0.9369
2	PuxAI	0.957	2	Athar	0.9272
3	NYUAD	0.927	3	HIAST	0.9259
4	HIAST	0.895	4	N&N	0.8984
5	MorAI	0.880	5	Tokenizers United	0.8738
6	CVPD	0.876	6	SEA	0.8601
7	QU-NLP	0.859	7	Teams60	0.8491
8	CIS-RG	0.763	8	Transformer_Tafsir	0.7970
9	ANLPers	0.707	9	CIS-RG	0.7874
10	Athar	0.704			
11	SHA	0.624			
12	SEA	0.599			
13	HAI	0.547			
14	Baseline	0.515			
15	IWAN	0.496			
16	Transform_Tafsir	0.447			

Table 3: Accuracy performance of teams on Task 1 and Task 2.

tem reached an accuracy of 97.2%, showcasing strong capabilities in handling complex jurisprudential computations. In Subtask 2 (Islamic Knowledge Assessment), the top score was 93.7%, reflecting the broader challenge of covering multiple Islamic disciplines.

The ***Gumball*** team (Elrefai et al., 2025) secured first place in Subtask 1 with a Qwen3-4B model fine-tuned through a two-stage pipeline combining classical inheritance texts with supervised MCQ training. Their system achieved 97.2% accuracy, outperforming all other submissions.

The ***PuxAI*** team (Phuc and Đặng Văn, 2025), ranked second, introduced a hybrid multi-agent architecture. For inheritance, they developed a *Virtual Inheritance Expert* pipeline combining fatwa retrieval with rule-based reasoning. For general knowledge, they designed a *Proponent–Critic Debate* pipeline, where agents engaged in adversarial reasoning before synthesis. Their system reached 95.7% on Subtask 1 and 93.7% on Subtask 2.

The ***NYUAD*** team (AlDahoul and Zaki, 2025), in third place, evaluated a diverse set of models, including open-source Arabic LLMs (Falcon3, Fanar, Allam), proprietary systems (GPT-4o, GPT-o3, Gemini Flash 2.5, Gemini Pro 2.5), and fine-tuned variants. While Arabic open-source models remained below 40% accuracy, proprietary models achieved up to 92.3%. Their final ensemble system (GPT-o3, Gemini Flash 2.5, Gemini Pro 2.5) reached 92.7%.

The ***HIAST*** team (Hamed et al., 2025) implemented a lightweight RAG pipeline based on Claude 4 Opus, retrieving top-ranked sources (often IslamWeb) and appending them to Arabic few-shot prompts. This approach improved inheritance reasoning, achieving 89.5% accuracy.

The ***MorAI*** team (R'baiti et al., 2025) proposed a collaborative LLM framework combining majority voting with retrieval-augmented generation. Their system integrated ALLaM-7B, DeepSeek-Reasoner, and Gemini-2.5-Flash, each independently generating predictions, with a voting mechanism selecting the final answer. Augmented with TF-IDF retrieval over a curated inheritance case database, their ensemble achieved 88.0% on Subtask 1, compared to 79.5% for ALLaM-7B, 71.8% for DeepSeek-Reasoner, and 83.5% for Gemini-2.5-Flash.

The ***CVPD*** team (Bekhouche et al., 2025) developed an encoder-based approach using Arabic text encoders with an Attentive Relevance Scoring

(ARS) module. Their best configuration, MARBERT with ARS, achieved 69.9% accuracy, while commercial LLMs such as Gemini reached up to 87.6%.

The ***QU-NLP*** team (AL-Smadi, 2025) fine-tuned Fanar-1-9B with LoRA and integrated it into a FAISS-based RAG pipeline. Their system achieved 85.8% accuracy, outperforming GPT-4.5 (74.0%), LLaMA-3 (48.8%), Mistral (44.5%), ALLaM-7B (42.9%), and the Fanar base model (48.1%).

The ***CIS-RG*** team (Zaki et al., 2025) combined fine-tuning, chain-of-thought prompting, and retrieval-augmented generation across multiple models, including Fanar, LLaMA, Gemini, and Mistral. Their hybrid system achieved 76.3% accuracy on Subtask 1, demonstrating competitive reasoning on basic inheritance cases but struggling with complex scenarios such as *ʿawl* and *ḥajb*.

The ***N&N*** team (Alangari and AlShenaifi, 2025) developed a system based on few-shot chain-of-thought prompting combined with ensemble methods and retrieval-augmented re-prompting (R^2P). Their pipeline consisted of (i) few-shot CoT prompting with standardized Arabic templates; (ii) a majority-vote ensemble over GPT-4o, Gemini 2.5, DeepSeek, and Qwen-plus; and (iii) retrieval-augmented re-prompting when the ensemble failed to agree. This design achieved 89.9% accuracy on Subtask 2, ranking them second overall in this task.

The ***ANLPers*** team (Sibaee et al., 2025) focused on Chain-of-Thought prompting, testing Claude 3.7 Sonnet and GPT-4o with direct-answer and step-by-step reasoning. Structured reasoning improved accuracy from 67.0% to 81.0% on Claude 3.7 and from 63.0% to 74.0% on GPT-4o. Error analysis revealed persistent difficulties with *tasheeh* (integer normalization of shares).

The ***Athar*** team (Noureldien et al., 2025) explored both subtasks with distinct strategies. For Subtask 1, they employed a zero-shot DeepSeek-R1 pipeline with constrained prompting and regex-based label extraction, achieving 70.4% accuracy. For Subtask 2, they designed a three-stage hybrid RAG pipeline combining BM25 and dense retrieval with GPT-based reranking, reaching 92.7% and ranking second overall. Their analysis highlighted sensitivity to question length and answer option complexity in inheritance reasoning, and retrieval errors as the main limitation in broader knowledge assessment.

The ***SHA*** team (Altammami, 2025) integrated static and dynamic few-shot prompting with retrieval-augmented generation. Although some models showed performance drops when augmented with additional context, their best configuration—Gemini with RAG and dynamic prompting—achieved 62.3% accuracy.

The ***SEA*** team (Alowaidi et al., 2025) designed an Islamic RAG framework with three stages: (i) knowledge resource preparation, preprocessing fatwas and Islamic books into 500-token chunks indexed in FAISS; (ii) retrieval using similarity search, Keyword-Augmented Two-Stage Retrieval (K2R), or Multi-Query Reformulation (MQR-K); and (iii) answer generation with structured prompting and post-generation validation. Their system achieved 60.0% accuracy on Subtask 1 and 86.0% on Subtask 2.

The ***ADAPT–MTU HAI*** team (Hossain and Afli, 2025) introduced a dual-expert architecture based on ALLaM-7B, combining a LoRA fine-tuned inheritance specialist with its base model. A constrained decoding mechanism enforced valid outputs (A–F). Their system achieved 54.7% accuracy, improving substantially over the 42.9% ALLaM-7B zero-shot baseline.

The ***Transformer Tafsir*** team (Abu Ahmad et al., 2025) developed a hybrid RAG pipeline combining sparse (BM25) and dense retrieval with cross-encoder reranking. While gains in inheritance reasoning were modest (Fanar: 44.0% → 45.0%; Mistral: 35.0% → 39.0%), Subtask 2 showed substantial improvements (Fanar: 55.0% → 80.0%; Mistral: 69.0% → 79.0%).

The ***Tokenizers United*** team (Samy et al., 2025) proposed a Retrieval-Augmented Generation (RAG) pipeline that combined *Muffakir* embeddings for domain-specific retrieval with the Gemini 2.5 Flash Lite model for lightweight generative reasoning. Their design prioritized efficiency, opting for direct similarity search (Top-K = 8–10) rather than complex reranking mechanisms. On the development set, performance varied between 44.3% and 84.3%, depending on the configuration. Their best-performing setup—Qdrant with cosine similarity, a chunk size of 400 characters, and Muffakir embeddings—achieved 87.4% accuracy on the official test set, ranking 5th out of 10 participating teams in Task 2.

6 Conclusions and Future work

In this paper, we presented the QIAS 2025 Shared Task, designed to evaluate the capabilities of large

language models in understanding and reasoning within Islamic knowledge domains. The task was divided into two subtasks: Islamic Inheritance Reasoning and Islamic Knowledge Assessment, both formulated as multiple-choice question answering problems with varying levels of difficulty. The submitted systems revealed significant performance gaps between open-source and commercial LLMs, with commercial models showing notably stronger results.

As a future direction, we plan to organize a follow-up edition of the shared task focused more deeply on Islamic inheritance. Unlike the current multiple-choice setup, the next edition will involve end-to-end problem solving—from identifying eligible heirs based on a given scenario to computing their exact shares. This approach will better reflect real-world applications and offer a more rigorous benchmark for legal reasoning tasks. In this context, we will also encourage researchers to use small and open-source language models. These models are easier to deploy, more accessible, and promote better transparency and reproducibility. We hope this will empower researchers—especially in low-resource settings—to develop useful tools and contribute to the field of Islamic studies.

7 Acknowledgments

We thank all participating teams for their contributions, as well as the organizers of the Arabic NLP 2025 Conference and the broader community for their support. We are grateful to our annotation team for ensuring the authenticity and appropriateness of our resources. This shared task was funded by the ARG grant ARG01-0524-230318, awarded by the Qatar Research, Development, and Innovation Council (QRDI). The efforts of Dr. Wajdi Zaghouani were partially supported by the NPRP14C-0916-210015 grant from QRDI.

References

Ummar Abbas, Mohammad Shahmeer Ahmad, Firoj Alam, Enes Altinisik, Ehsannedin Asgari, Yazan Boshmaf, Sabri Boughorbel, Sanjay Chawla, Shammur A. Chowdhury, Fahim Dalvi, Kareem Darwish, Nadir Durrani, Mohamed Elfeky, Ahmed K. Elmagarmid, Mohamed Y. Eltabakh, Masoomali Fatehkia, Anastasios Fragkopoulos, Maram Hasanain, Majd Hawasly, and 22 others. 2025. Fanar: An arabic-centric multimodal generative AI platform. *arXiv preprint*, arXiv:2501.13944.

Muhammad Abu Ahmad, Mohamad Ballout, Raia Abu Ahmad, and Elia Bruni. 2025. Transformer tafsir at qias 2025 shared task: Hybrid retrieval-augmented generation for islamic knowledge question answering. In *Proceedings of The Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics.

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. *International Journal of Advanced Research in Computer Science*.

Mohammad AL-Smadi. 2025. Qu-nlp at qias 2025 shared task: A two-phase llm fine-tuning and retrieval-augmented generation approach for islamic inheritance reasoning. In *Proceedings of The Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics.

Ahmet Yusuf Alan, Enis Karaarslan, and Ömer Aydin. 2024. A rag-based question answering system proposal for understanding islam: Mufassirqas llm. *arXiv preprint arXiv:2401.15378*.

Nourah Alangari and Nouf AlShenaifi. 2025. N&n at qias 2025: Chain-of-thought ensembles with retrieval-augmented framework for classical arabic islamic. In *Proceedings of The Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics.

Nouar AlDahoul and Yasir Zaki. 2025. Nyuad at qias shared task: Benchmarking the legal reasoning of llms in arabic islamic inheritance cases. In *Proceedings of The Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics.

Hayfa A Aleid and Aqil M Azmi. 2025. Hajj-fqa: A benchmark arabic dataset for developing question-answering systems on hajj fatwas: H. aleid and a. azmi. *Journal of King Saud University Computer and Information Sciences*, 37(6):135.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, and 1 others. 2023. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*.

Sarah Alnefaie, Eric Atwell, and Mohammad Ammar Alsalka. 2023. Is gpt-4 a good islamic expert for answering quran questions? In *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*, pages 124–133.

Sanaa Alowaidi, Eric Atwell, and Mohammed Ammar Alsalka. 2025. Sea-team at qias 2025: Enhancing llms for question answering in islamic texts. In *Proceedings of The Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics.

Shatha Altammami. 2025. Sha at the qias shared task: Llms for arabic islamic inheritance reasoning. In *Proceedings of The Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics.

M. Saiful Bari, Yazeed Alnumay, Norah A. Alzahrani, Nouf M. Alotaibi, Hisham A. Alyahya, Sultan Alrashed, Faisal Mirza, Shaykhah 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. 2024. Allam: Large language models for arabic and english. *arXiv preprint arXiv:2407.15390*.

Salah Eddine Bekhouche, Abdellah Zakaria Sellam, Hichem Telli, Cosimo Distante, and Abdenour Hadid. 2025. Cvpd at qias 2025 shared task: An efficient encoder-based approach for islamic inheritance reasoning. In *Proceedings of The Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics.

Abdessalam Bouchekif, Samer Rashwani, Heba Sbahi, Shahd Gaben, Mutaz Al-Khatib, and Mohammed Ghaly. 2025. 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, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 43 others. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. *arXiv preprint arXiv:2507.06261*. Rapport technique, Équipe Gemini, Google.

Eman Elrefai, Abdelrahman Ahmad, Aml Hassan Esmail, and Mohamed Lotfy Elrefai. 2025. Gumball at qias 2025: Arabic llm automated reasoning in islamic inheritance. In *Proceedings of The Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Mohamed Motasim Hamed, Nada Ghneim, and Riad Sonbol. 2025. Hiast at qias 2025: Retrieval-augmented llms with top-hit web evidence for arabic islamic reasoning qa. In *Proceedings of The Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics.

Shehenaz Hossain and Haithem Afli. 2025. Adapt–mtu hai at qias2025: Dual-expert llm fine-tuning and constrained decoding for arabic islamic inheritance reasoning. In *Proceedings of The Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics.

Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Dingjie Song, Zhihong Chen, Abdulmohsen Alharthi, Bang An, Juncai He, and 1 others. 2023. Acegpt, localizing large language models in arabic. *arXiv preprint arXiv:2309.12053*.

Anis Koubaa, Adel Ammar, Lahouari Ghouti, Omar Najar, and Serry Sibaee. 2024. Arabiangpt: Native arabic gpt-based large language model. *arXiv preprint arXiv:2402.15313*.

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 *Proceedings of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools*, 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.

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.

Yossra Noureldien, Hassan Suliman, Farah Attallah, Abdelrazig Mohamed, and Sara Abdalla. 2025. Athar at qias2025: Mcqs-based question answering systems for islamic inheritance and classical knowledge. In *Proceedings of The Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics.

Nguyen Xuan Phuc and Thìn Đặng Văn. 2025. Puxai at qias 2025: Multi-agent retrieval-augmented generation for islamic inheritance and knowledge reasoning. In *Proceedings of The Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics.

Faiza Qamar, Seemab Latif, and Rabia Latif. 2024. A benchmark dataset with larger context for non-factoid question answering over islamic text. *arXiv preprint arXiv:2409.09844*.

Jihad R'baiti, Chouaib El Hachimi, Youssef Hmamouche, and Amal Seghrouchni. 2025. Morai at qias 2025: Collaborative llm via voting and retrieval-augmented generation for solving complex inheritance problems. In *Proceedings of The Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics.

Muhammad Razif Rizqullah, Ayu Purwarianti, and Alham Fikri Aji. 2023. Qasina: Religious domain question answering using sirah nabawiyah. In *2023 10th International Conference on Advanced Informatics: Concept, Theory and Application (ICAICTA)*, pages 1–6.

Mohamed Samy, Mayar Boghdady, Marwan El Adawi, Mohamed Nassar, and Ensaf Hussein. 2025. Tokenizers united at qias 2025: Rag-enhanced question answering for islamic studies by integrating semantic retrieval with generative reasoning. In *Proceedings of The Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics.

Mohammad Amaan Sayeed, Mohammed Talha Alam, Raza Imam, Shahab Saquib Sohail, and Amir Hussain. 2025. From rag to agentic: Validating islamic-medicine responses with llm agents. *arXiv preprint arXiv:2506.15911*.

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

Serry Sibaee, Mahmoud Reda, Omer Nacar, Yasser Alhabashi, Adel Ammar, and Wadii Boulila. 2025. Anlpers at qias 2025: Cot for islamic inheritance. In *Proceedings of The Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics.

Sadia Tabassum, AHM Hoque, Sharaban Twahura, and Mohammad Osiur Rahman. 2019. Developing an islamic farayez system applying software engineering. *Jurnal Kejuruteraan*, 31(1):25–38.

Osama Zaki, Asmaa Badawy, Nada Elgewily, and Ahmed Sharaf. 2025. Cis-rg at qias 2025: Assessing large language models on islamic legal reasoning and mathematical calculations. In *Proceedings of The Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics.

Samia Zouaoui and Khaled Rezeg. 2021. Islamic inheritance calculation system based on arabic ontology (arafamonto). *Journal of King Saud University-Computer and Information Sciences*, 33(1):68–76.