# HIAST at QIAS 2025: Retrieval-Augmented LLMs with Top-Hit Web Evidence for Arabic Islamic Reasoning QA

**Mohamed Motasim Hamed**[*], **Riad Sonbol**[*], **Nada Ghneim**[**]
[*]Higher Institute for Applied Sciences and Technology, Damascus, Syria
[**]Arab International University, Daraa, Syria
{motasim.hamed, riad.sonbol}@hiast.edu.sy, n-ghneim@aiu.edu.sy

## Abstract

We describe our participation in the QIAS 2025 Shared Task on Islamic Studies Question Answering, comprising two subtasks: (1) Islamic Inheritance Reasoning and (2) General Islamic Knowledge Assessment. Both were solved using the Claude 4 Opus LLM via API with tailored prompting. For Subtask 1, we implemented a lightweight Retrieval-Augmented Generation (RAG) pipeline, which retrieves the top Google Search result (often from IslamWeb), preprocesses it, and appends it to a structured few-shot Arabic prompt, thereby boosting reasoning accuracy. For Subtask 2, where web-retrieval was not feasible due to closed-book sources, we applied topic-diverse few-shot prompting to leverage the model's internal knowledge. Our systems achieved 4th/15 (0.895) in Subtask 1 and 3rd/10 (0.9259) in Subtask 2, demonstrating the effectiveness of targeted retrieval in open-web contexts and structured prompting in closed-domain Arabic QA.

## 1 Introduction

The QIAS 2025 Shared Task (*Question and Answer in Islamic Studies Assessment*) serves as a benchmark for evaluating large language models (LLMs) on domain-specific reasoning in Islamic knowledge (Bouchekif et al., 2025a). It consists of two multiple-choice subtasks: (1) Islamic Inheritance Reasoning (ᶜIlm al-Mawārīth) and (2) Islamic Knowledge Assessment (covering Fiqh, al-Ḥadīth, Tafsīr, uṣūl al-fiqh, etc.), with MCQs spanning beginner, intermediate, and advanced levels—designed to assess reasoning accuracy in both retrieval-supported and retrieval-free settings.

The task is conducted entirely in Arabic, reflecting the primary language of Islamic scholarship and presenting a significant challenge for LLMs given the language's morphological richness and syntactic complexity.

The two subtasks differ in their source and the feasibility of web-based retrieval. Subtask 1 draws primarily from online fatāwā, making retrieval from the open web practical and often effective. Subtask 2, by contrast, is based on classical and modern Islamic closed books that are generally not available through open web indexing; although retrieval may be beneficial in this task, we opted to approach it using the internal knowledge of the language model.

Our submission addresses both QIAS subtasks using LLM-based pipelines built around the Claude Opus 4 API. For Subtask 1 (Retrieval-supported QA), we adopted a single-document retrieval approach using the Google Search API, appending the top-ranked result, often from IslamWeb, to a structured few-shot prompt. This provided the model with both contextual exemplars and relevant retrieved knowledge, substantially enhancing rule-based reasoning in inheritance scenarios. Across various strategies and LLMs, supplementing questions with retrieved content from reliable sources yielded accuracy improvements exceeding 14% over using the model's internal knowledge alone.

For Subtask 2 (Zero-retrieval QA), where the data source comprised offline books, no external retrieval was performed. Instead, we employed a structured few-shot prompt with curated exemplars, leveraging the model's internal knowledge and reasoning capabilities. This strategy achieved competitive accuracy, demonstrating the model's ability to recall domain-specific information and perform sophisticated reasoning even in the closed-book setting.

In general, our findings underscore the effectiveness of structured web-retrieval in low-resource domains, such as Islamic law. QIAS offers a valuable benchmark for testing such approaches. We report the leaderboard results [1], along with the implemen-

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

*Proceedings of The Third Arabic Natural Language Processing Conference*, pages 883–891
November 8-9, 2025 ©2025 Association for Computational Linguistics

tation details [2].

## 2 Background

### 2.1 Task Setup

We participated in **both** subtasks of the QIAS 2025 Shared Task on evaluating multilingual LLMs in Islamic reasoning and knowledge, as described in the task overview (Bouchekif et al., 2025a).

**Subtask 1: Islamic Inheritance Reasoning (ʿIlm al-Mawārīth)** Each item in Subtask 1 frames a detailed family fact pattern involving heirs such as wife, parents, full or half siblings, children, or deceased heirs. A multiple-choice question (6 options; exactly one correct) asks for the appropriate heir-share(s) based on fixed-rule Islamic jurisprudence (farāʾiḍ).

**Example of Beginner level in Arabic, from the official dataset:**

توفي عن أب، و2 أخ شقيق، و1 ابن أخ شقيق، و2 عم شقيق للأب، وأم، و2 بنت، و1 زوجة، ما هو نصيب الأم؟
A) الثلث, B) الربع, C) السدس, D) الثمن, E) النصف, F) لا شيء

**Subtask 2: Islamic Knowledge Assessment** Subtask 2 contains knowledge-based exam MCQs on topics such as ʿulūm al-Qurʾān, al-Ḥadīth, fiqh, uṣūl al-fiqh, sīrah, and Aqīdah, designed to elicit doctrinal, doctrinal-reasoned, or interpretive recall. The questions come in 4-option MCQ format, with exactly one correct answer.

**Example of Beginner level in Arabic, from the official dataset:**

ما مدة المسح على الخفين للمقيم؟
A)يوم وليلة, B)ثلاثة أيام بلياليهن, C) يومان وليلتان, D) أسبوع كامل

### 2.2 Dataset Details

**Subtask 1:** The dataset comprises ~20,000 training, 1,000 validation, and 1,000 test MCQs (six options each), generated from curated IslamWeb fatwas via Gemini 2.5 and validated by domain experts. Pre-processing involved deduplication and disambiguation. An auxiliary corpus of 3,165 original fatwas was also provided as an optional knowledge base (Bouchekif et al., 2025a).

**Subtask 2:** A collection of classical Islamic texts was provided as unsupervised data to fine-tune or as part of a Retrieval-Augmented Generation (RAG) (Lewis et al., 2020). From 25 reference texts, 1,400 MCQs (700 validation, 700 test) were created in seven disciplines (beginner-advanced), each with four answer choices, and validated by five experts.

### 2.3 Prior Work

Recent advances in Arabic LLMs, such as Jais (Sengupta et al., 2023), AceGPT (Huang et al., 2023), ALLaM (Bari et al., 2024), Kuwain (Hennara et al., 2025), and Fanar (Abbas et al., 2025) have expanded the ability to understand religious texts by pretraining large-scale corpora, including Quran, Hadith, and fatwa archives. Shared tasks have emerged to benchmark Islamic NLP, including Quranic QA (Malhas et al., 2022, 2023), Islamic knowledge retrieval (Qamar et al., 2024), and most recently QIAS 2025 (Bouchekif et al., 2025a), which evaluates LLMs on Islamic inheritance reasoning and general religious knowledge. Prior work in automating Islamic inheritance (IRTH) largely relied on expert systems (Akkila and Naser, 2016; Tabassum et al., 2019; Zouaoui and Rezeg, 2021) that encoded symbolic rules, or on RAG-based QA pipelines for Islamic texts (Alan et al., 2024; Sayeed et al., 2025). However, the performance of LLM on Islamic content remains constrained by factual errors, misinterpretation of context, and sensitivity to question phrasing (Mohammed et al., 2025; Alnefaie et al., 2023; Bouchekif et al., 2025b).

In Subtask 1, we used a lightweight, single-document retrieval-augmented generation setup, grounding Claude Opus 4 with authoritative sources retrieved via the Google Search API. This lightweight approach avoids handcrafted rules and instead provides juristic context for in-context reasoning. By contrasting this retrieval-supported configuration with a zero-retrieval baseline in Subtask 2, we enable a controlled comparison of retrieval-augmented versus unsupported Islamic-domain reasoning.

## 3 System Overview

Our QIAS 2025 submission adopts two distinct configurations, each tailored to its subtask, both powered by the Claude Opus 4 API.

### 3.1 Subtask 1: Islamic Inheritance Reasoning (Lightweight RAG)

We implemented a single-document Retrieval-Augmented Generation (RAG) pipeline. Each question was paired with the top-ranked Google Search

[2] https://gitlab.com/Moatasem444/qias2025-hiast-submission/

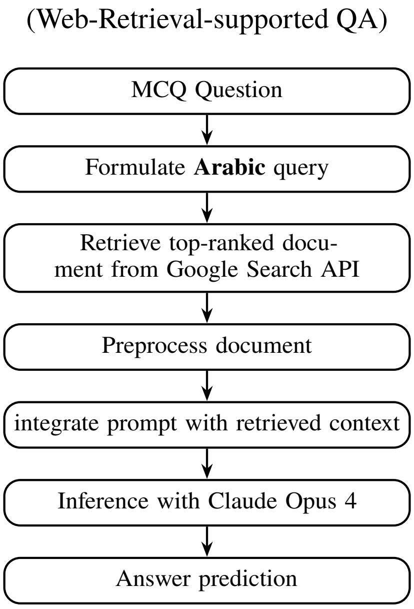
(Web-Retrieval-supported QA)
MCQ Question
Formulate **Arabic** query
Retrieve top-ranked document from Google Search API
Preprocess document
integrate prompt with retrieved context
Inference with Claude Opus 4
Answer prediction

Figure 1: Web-Retrieval pipeline for Subtask 1.

result (primarily from IslamWeb[3], pre-processed to remove boilerplate and appended (in Arabic) to the questions and answer choices. This method provided high-quality domain-specific grounding while avoiding the latency and complexity of multi-document vector retrieval. The complete retrieval pipeline is provided in figure 1.

### 3.2 Subtask 2: Islamic Knowledge Assessment (Zero-Retrieval Few-Shot)

Since web-retrieval was not possible, we adopted a few-shot prompting strategy. Three to four representative MCQs, covering different topics and difficulty levels, were inserted into a fixed Arabic prompt template before the test question. This few-shot configuration leveraged the LLM's internal knowledge for doctrinal and interpretive reasoning. Detailed work examples are provided in figure 2. The detailed prompt formulations corresponding to the two subtasks are provided in Appendix A.

### 3.3 Challenges Addressed

We address several challenges in our approach, including Arabic morphology and orthography, where queries preserve diacritics to improve retrieval precision; input length control, with retrieved passages truncated to $L_{\max} = 2000$ characters to fit model limits; and knowledge coverage, where few-shot examples are selected for topical diversity to reduce bias toward frequent topics.

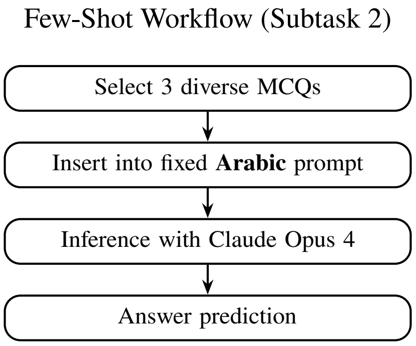
Few-Shot Workflow (Subtask 2)
Select 3 diverse MCQs
Insert into fixed **Arabic** prompt
Inference with Claude Opus 4
Answer prediction

Figure 2: Few-shot pipeline for Subtask 2.

## 4 Experimental Setup

### 4.1 Data Sources and Preprocessing

For **Subtask 1**, we performed real-time retrieval from publicly available web sources, primarily IslamWeb, using the Google Custom Search API, instead of using the 3,165 IslamWeb fatwas provided. The retrieved documents were cleaned by removing HTML tags and boilerplate text and truncated to $L_{\max} = 2000$ UTF-8 characters.

For **Subtask 2**, the large corpus of classical Islamic texts was not used. Instead, we applied a few-shot prompt with $k = 3$ examples drawn from the validation set (to prevent data leakage). The few-shot MCQs were normalized to a consistent template format comprising an Arabic question stem and four labeled options.

### 4.2 Task Evaluation Metrics

We followed the official QIAS 2025 evaluation protocol (Bouchekif et al., 2025a), using only the validation and test sets provided. The precision in the test set, calculated as the proportion of exact matches between the predictions and the gold answers, was the only ranking metric. The outputs were normalized to choice letters. The source code is available online[4].

### 4.3 Implementation Details

**Hyperparameters.** Both configurations used the default Claude Opus 4 API parameters, with temperature set to 0.0 for deterministic outputs and max_tokens fixed at 1000. For Subtask 2, the number of few-shot exemplars was set to $k = 3$ based on preliminary validation and the number of difficulty levels.

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

[4] https://gitlab.com/Moatasem444/qias2025-hiast-submission/

**External Tools and Libraries.** We employed the Claude Opus 4 API (Anthropic, May 2025 release) as the primary LLM, accessed via its paid subscription tier. Live web-retrieval for Subtask 1 was performed using the Google Custom Search API[5], also under a paid usage plan. All API integration and pre-processing were implemented in Google Colab.

## 5 Results

### 5.1 Official Leaderboard Performance

We achieved strong results in test data for both subtasks, placing 4th/15 in Subtask 1 (Accuracy: 0.895) and 3rd/10 in Subtask 2 (Accuracy: 0.9259). Detailed scores and rankings are shown in Table 5 (Appendix B).

### 5.2 Comparative Analysis and Error Patterns

For Subtask 1, we compared our lightweight Google Search API pipeline with:

(a) the same LLM without retrieval,

(b) the same LLM with its built-in "web search" mode.

Our approach consistently outperformed all baseline methods in a wide range of models, including closed-source systems such as Claude 4 Opus (C4O), GPT 4.1 Mini (G4.1M), and Gemini 2.5 Flash (G2.5F), as well as the open-source model like Fanar (Table 1). In particular, even minimal, yet high-quality retrieval yielded substantial gains in accuracy.

In contrast, for Subtask 2, where sources are closed-book, web-retrieval offered no benefit; structured few-shot prompting proved most effective (Table 2).

Table 1: Validation results for Subtask 1 under different models and retrieval settings.

| Model | No Ret. | Built-in WS | Ours |
|---|---|---|---|
| C4O | 0.785 | 0.812 | **0.924** |
| G2.5F | 0.700 | N/A | 0.871 |
| G4.1M | 0.580 | 0.690 | 0.822 |
| Fanar | 0.574 | N/A | 0.645 |

Error analysis on 50 random misclassifications per task revealed:

- **Task 1:** Failures occurred when the retrieval was missing or contained partial matches, forcing reliance on internal knowledge of LLM. Some models (e.g., Gemini) deviated from the format by including explanations. Other causes included ambiguous fatwā phrasing, missing numeric details, and context length limits ($L_{\max} = 2000$).
- **Task 2:** Most errors stemmed from fine-grained doctrinal differences and narrations that required exact recall. It should be noted that error rates are distributed inversely between difficulty levels.

Table 2: Validation results for Subtask 2 on Gemini 2.5 flash, showing no gain from web-retrieval (WebR).

| Method | Acc. | Δ vs. Few-Shot |
|---|---|---|
| Few-Shot only | 0.875 | — |
| Few-Shot + WebR | 0.805 | −7% |

### 5.3 Similarity and Prediction Accuracy

We computed the similarity score between text and the result of the top hits on the Web using the Muffakir Embedding model [6]. We then analyzed the relationship between the similarity score and the prediction precision. The results do not indicate significant relevance: the average similarity for incorrect predictions was 0.645, while for correct predictions it was 0.653, with a correlation of only 0.027. Importantly, it also indicates that retrieving only the top-ranked search result is sufficient: If the answer is present in the retrieved context, it is most likely in the first result, and additional results are unlikely to improve accuracy. This further supports the effectiveness of Google Search's ranking in providing the most relevant information for this task.

### 5.4 Error Analysis

We examined the relationship between the availability of the retrieved web context and the system error rates (Table 3). Although the number of cases with the retrieved web context (923) is substantially higher than those without (77), the relative error rate is lower (7.37% vs. 9.09%). This demonstrates the effectiveness of web-retrieval; If comparable contextual information had been available for the remaining instances, the overall error rate could have been reduced by up to 1.7%. In contrast, the absence of such context correlates with increased error rates. In particular, the system fails to retrieve relevant context in approximately 8.3% of the total cases, which directly limits the attainable performance limit. Reducing this retrieval failure rate is

[5] https://developers.google.com/custom-search

[6] https://huggingface.co/mohamed2811/Muffakir_Embedding

therefore critical to achieving consistently higher accuracy.

Table 3: Subtask 1 error rates with/without web context (NW = No Web).

| Context | #Q | #Wrong | Err. (%) |
|---|---|---|---|
| With Web | 923 | 68 | 7.37 |
| NW | 77 | 7 | 9.09 |

We also analyzed incorrect predictions by difficulty level of the questions in the validation sets for both tasks (Table 4). In task 1, the majority of errors occurred at the advanced level (45 errors, 60%), followed by the beginner level (30 errors, 40%). In Task 2, errors were more evenly distributed: the beginner level questions accounted for 40 errors (43%), intermediate for 31 errors (33.3%), and advanced for 22 errors (23.7%). These results suggest that in Task 1, advanced-level questions are disproportionately challenging, while in Task 2, errors are less skewed toward a single difficulty level, indicating a more balanced difficulty distribution.

Table 4: Wrong predictions by difficulty level in validation data. Percentages relative to total wrong predictions per task.

| Level | Task 1 | | Task 2 | |
|---|---|---|---|---|
| | Wrong | % | Wrong | % |
| Beginner | 30 | 40.0 | 40 | 43.0 |
| Intermediate | – | – | 31 | 33.3 |
| Advanced | 45 | 60.0 | 22 | 23.7 |

Additional error samples are shown in Appendix C.

## 6 Conclusion

We addressed the QIAS 2025 Shared Task using large language models with task-specific prompting strategies. For Subtask 1, live Google Search retrieval achieved 0.895 accuracy, while for Subtask 2, few-shot prompting reached 0.9259 accuracy. The main limitations include the dependence on the quality of the retrieval, the doctrinal differences, and the dependence on the closed-source Claude Opus 4 API. Future work will fine-tune Arabic-specific models and employ domain-restricted RAG over curated texts to mitigate coverage gaps and ambiguity.

## 7 Acknowledgments

We acknowledge MISRAJ.AI for supporting parts of the development process and anonymous reviewers for their constructive comments.

## References

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

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

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*.

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.

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

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

Khalil Hennara, Sara Chrouf, Mohamed Motaism Hamed, Zeina Aldallal, Omar Hadid, and Safwan AlModhayan. 2025. Kuwain 1.5 b: An arabic slm via language injection. *arXiv preprint arXiv:2504.15120*.

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*.

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

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.

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.

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*.

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*.

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.

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.

# 8 Appendix

## A Prompt Templates

### Subtask 1 Prompt

These are a few-shot examples for the task: answering multiple-choice questions by selecting the correct option.

**Example 1:**
**Question:** توفي عن أب، وأخوين شقيقين، وابن أخ شقيق، وعمين شقيقين، وأم، وبنتين، و زوجة، فما نصيب الأم؟
A) الثلث
B) الربع
C) السدس
D) الثمن
E) النصف
F) لا شيء

**Answer:** C

**Example 2:**
**Question:** توفي عن أخ من الأم، وبنت، وزوجة، وأختين من الأم: كم عدد أسهم البنت بعد الرد؟
A) سهم واحد
B) سهمان
C) ثلاثة أسهم
D) أربعة أسهم
E) سبعة أسهم
F) ثمانية أسهم

**Answer:** E

**Context for the question:** {context}

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

**Question:** {question} {options_text}

Please respond using **only one English letter** from the following: A, B, C, D, E, F.
Do not write any explanation or additional text.

### Subtask 2 Prompt

These are a few-shot examples for the task: answering multiple-choice questions by selecting the correct option.

**Example 1:**
**Question:** ما مدة المسح على الخفين للمقيم؟
A) يوم وليلة
B) ثلاثة أيام بلياليهن
C) يومان وليلتان
D) أسبوع كامل

**Answer:** A

**Example 2:**

**Question:** من شروط الأصل في القياس؟
A) أن يكون الأصل فرعا لأصل آخر
B) أن لا يكون الحكم ثابتا
في الأصل بطريق سمعي شرعي C) ألا يكون الأصل فرعا لأصل آخر
D) ألا تعرف طريقة الاستنباط

**Answer:** C

**Example 3:**
**Question:** ما هو طريق الحكماء لإثبات وجود الواجب؟
A) عن طريق اعتبار العالم قديمًا.
B) عن طريق إثبات أن العالم واجب لذاته.
C) عن طريق امتناع التسلسل والدور.
D) عن طريق إثبات حدوث العالم.

**Answer:** C

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

**Question:** {question} {options_text}

Please respond using **only one English letter** from the following: A, B, C, D.
Do not write any explanation or additional text.

## B Supplementary Results

Table 5: QIAS 2025 official leaderboards. Our system **HIAST** ranked **4th** in Subtask 1 (Acc. 0.895) and **3rd** in Subtask 2 (Acc. 0.9259) on the test set.

**Subtask 1: Islamic Inheritance Reasoning (ranked by test Accuracy)**

| Rank | Team | Accuracy | Affiliation(s) |
|---|---|---|---|
| 1 | Gumball | 0.972 | Alexandria University, Ain Shams University |
| 2 | PuxAI | 0.957 | VNU-HCM University of Information Technology |
| 3 | NYUAD | 0.927 | New York University Abu Dhabi |
| **4** | **HIAST** | **0.895** | **Higher Institute for Applied Sciences and Technology** |
| 5 | MorAI | 0.880 | International Center for AI, Mohammed VI Polytechnic University |

**Subtask 2: Islamic Knowledge Assessment (ranked by test Accuracy)**

| Rank | Team | Accuracy | Affiliation(s) |
|---|---|---|---|
| 1 | PuxAI | 0.9369 | VNU-HCM University of Information Technology |
| 2 | Athar | 0.9272 | University of Khartoum, International Islamic University Malaysia |
| **3** | **HIAST** | **0.9259** | **Higher Institute for Applied Sciences and Technology** |
| 4 | N&N | 0.8984 | King Saud University |
| 5 | Tokenizers United | 0.8738 | Nile University |

## C Error Examples

These examples illustrate that Subtask 2 errors arise from doctrinal differences across Islamic disciplines (e.g., Sufism) as well as reference/attribution issues that require reliable sourcing.

Table 6: Examples of wrong predictions in Subtask 2 validation data, categorized by difficulty level and error type.

| Level | Question (Arabic) | Correct Answer | Model Prediction | Error Type |
|---|---|---|---|---|
| Beginner | ما الذي يقصد بالفناء الصوفي؟ | A) استغراق النفس في الروح الإلهي | B) موت النفس | Comprehension/Disambiguation |
| Intermediate | ما هو الحرف الناقص في آية سورة ص مقارنة بآية سورة البقرة؟ | B) حرف "الواو" | A) حرف "أبى" | Comprehension/Disambiguation |
| Advanced | ما هو القول القديم للشافعي في صوم أيام التشريق؟ | B) يجوز صومها للمتمتع إذا عدم الهدي عن الأيام الثلاثة الواجبة في الحج | A) لا يجوز صومها مطلقاً | Doctrinal Variance |
| Beginner | ما المصادر غير الإسلامية التي يرى ترمنجهام أن التصوف يمت إليها بصلة طفيفة؟ | C) الحياة الصوفية الزهدية للمسيحية الشرقية | B) الفلسفة الهندية | Reference/Attribution Error |
| Advanced | ما هو الاعتراض الذي قد يورد على مقدمة إمكان مخالفة أحد الإلهين للآخر؟ | C) أن مخالفة أحدهما للآخر وإرادة ضده ليست ممكنة دائمًا | B) أن المخالفة بين الإلهين مستحيلة | Doctrinal Variance |
| Beginner | هل تعرض الإمام البخاري في "التاريخ الكبير" للجرح والتعديل؟ | B) تعرض أحياناً | C) تعرض دائماً | Reference/Attribution Error |
| Intermediate | ما معنى "اللمزة" في اللغة؟ | B) الذي يعيب الناس سرًا ويؤذيهم | A) الذي يشتم الرجل علانية ويكسر عينيه عليه | Comprehension/Disambiguation |