N&N at QIAS 2025: Chain-of-Thought Ensembles with Retrieval-Augmented framework for Classical Arabic Islamic MCQs

Nourah Alangari
King Saud University
nmalangari@ksu.edu.sa

Nouf AlShenaifi
King Saud University
noalshenaifi@ksu.edu.sa

Abstract

We present our system developed for the Question-and-Answer in Islamic Studies Assessment Shared Task on Evaluating LLMs for Islamic Knowledge (QIAS 2025), which focuses on answering Arabic multiple-choice questions (MCQs) derived from classical Islamic texts. Our methodology integrates few-shot chain-of-thought prompting across multiple LLMs, enhanced by a majority-vote ensemble mechanism. In situations of ensemble uncertainty, we deploy a retrieval-augmented re-prompting module that extracts contextually relevant passages from digitized Islamic sources to refine model predictions. Our final system achieves an accuracy of **89.8%** on the hidden test set.

1 Introduction

Recent advancements in large language models (LLMs) have significantly enhanced their capabilities in understanding and reasoning across diverse knowledge domains. However, their performance on specialized, culturally-rich content such as classical Islamic texts remains less explored. Classical Islamic texts—covering jurisprudence, creed, exegesis, and hadith—pose distinctive challenges: they are primarily in Arabic, employ specialized terminology, encode subtle doctrinal distinctions across legal schools, and often require multi-step reasoning (e.g., analogical and numerical reasoning in inheritance) to reach a correct answer (Bouchekif et al., 2025b). In this context, we present our system for the Question-and-Answer in Islamic Studies Assessment Shared Task on Evaluating LLMs for Islamic Knowledge (QIAS 2025) (Bouchekif et al., 2025a), which involves answering Arabic multiple-choice questions (MCQs) drawn specifically from classical Islamic literature. Our proposed approach integrates few-shot chain-of-thought prompting across several prominent LLMs, coupled with a robust majority-vote ensemble strategy. When the ensemble fails to reach consensus, our retrieval-augmented re-prompting (R^2P) module dynamically retrieves relevant textual evidence from digitized Islamic resources, enabling models to produce refined and contextually grounded predictions. Our final submission achieves an accuracy of **89.8%** on the hidden test set. The remainder of this paper is structured as follows: Section 2 reviews related work. Section 3 describes the QIAS 2025 task and dataset. Section 4 presents the system overview. Section 5 details the experimental setup. Section 6 reports and analyzes results. Section 7 concludes and outlines future work. An Appendix includes prompt templates and additional examples.

2 Related work

Large Language Models (LLMs) have shown remarkable advancements in zero-shot and few-shot reasoning tasks (Al Nazi et al., 2025) (Meshkin et al., 2024). Chain-of-thought (CoT) prompting has emerged as a powerful strategy to guide LLMs through intermediate reasoning steps before producing a final answer. Introduced by Wei et al. (Wei et al., 2022), CoT prompting significantly improved performance on tasks requiring logical reasoning, arithmetic, and commonsense inference. Later, Wang et al. (Wang et al., 2022) enhanced this framework with self-consistency sampling, where multiple reasoning paths are sampled, and the most consistent final answer is selected resulting in more robust predictions. While these techniques have been extensively evaluated on general-domain tasks in English, their application to Arabic particularly domain-specific Arabic such as classical Islamic jurisprudence and theology remains limited. Retrieval-augmented generation (RAG), introduced by Lewis et al. (Lewis et al., 2020), combines external document retrieval with generation-based models to inject relevant background knowl-

Proceedings of The Third Arabic Natural Language Processing Conference, pages 878–882

edge into the reasoning process. RAG has shown utility in open-domain QA, but few studies have adapted this method to classical Arabic corpora with domain-specific embeddings and passage reranking (Omoush and Ghnemat, 2025) (Bazzi and Gaith, 2025). Our contribution is novel in its application of CoT ensembles with on-demand retrieval for domain-specific Arabic MCQs—a setting that requires precise integration of theological and jurisprudential sources. This combination of retrieval-augmented CoT and ensemble majority voting is particularly impactful for advanced questions requiring deeper contextual grounding.

3 Task Description

3.1 Task setup

The QIAS 2025 Subtask 2 involves answering classical Islamic multiple-choice questions (MCQs) in Arabic. Each input consists of a question stem and four possible answers (labeled A–D), with a single correct option. For example, a typical input might present a jurisprudential question derived from classical Islamic texts and require the system to output the correct choice label. This task requires deep semantic understanding, domain-specific expertise—particularly within Islamic contexts—and a keen ability to discern subtle linguistic nuances in the Arabic language.

3.2 Dataset

The dataset employed in this task comprises 1,400 Arabic multiple-choice questions (MCQs), evenly divided into 700 for validation and 700 for testing. These questions are meticulously curated from authoritative classical Islamic texts and cover a range of domains, including Fiqh (Islamic jurisprudence), Sīrah (the prophetic biography), Ulūm al-Qur'ān (Qur'anic sciences), and Ulūm al-Hadith (Hadith studies). To assess the system's reasoning capabilities, the questions are categorized into three levels of difficulty—Beginner, Intermediate, and Advanced—each reflecting a progressively deeper level of conceptual and analytical complexity (Bouchekif et al., 2025a). Additionally, well-known Islamic e-books such as Ar-Raḥīq al-Makhtūm (The Sealed Nectar) and Al-Itqān fī Ulūm al-Qur'ān (The Perfect Guide to the Sciences of the Qur'an) are provided as supplementary resources, serving as foundational references for the task. Figure 1 presents a sample multiple-choice question (MCQ) from the QIAS 2025 Shared Task

Example:
ما هو القول القديم للشافعي في صوم أيام التشريق؟

A) **لا يجوز صومها مطلقاً.**
B) **يجوز صومها للمتمتع إذا عدم الهدي عن الأيام الثلاثة الواجبة في الحج.**
C) **يجوز صومها لمن لم يجد الهدي فقط.**
D). **يجوز صومها للمسافر فقط**

Figure 1: A sample MCQ from QIAS 2025 Subtask 2.

(Subtask 2: Islamic Assessment), which evaluates language models' understanding of classical Islamic knowledge.

3.3 Track Participation

We participated in the QIAS 2025 Shared Task (Subtask 2: Islamic Assessment), part of the ArabicNLP 2025 conference held in conjunction with EMNLP 2025. This subtask centers on evaluating large language models (LLMs) in the domain of classical Islamic knowledge through multiple-choice questions. As one of the first benchmarks specifically designed for Arabic MCQs in religious and jurisprudential contexts, it provides a structured and rigorous framework for assessing deep semantic understanding and domain-specific reasoning in Islamic studies.

4 System Overview

Our pipeline, illustrated in (Figure 2), consists of three main stages designed for robust Arabic Islamic multiple-choice question answering:

1. **Prompt Sampling and Few-Shot CoT Prompting:** In the first stage, we leverage few-shot chain-of-thought (CoT) prompting techniques. Five carefully selected demonstration examples from the QIAS validation MCQs are embedded into a standardized Arabic prompt template.

2. **Majority Ensemble:** In the second stage, we employ a majority voting ensemble using the top three performing models selected from GPT-4o, Qwen-Plus, Gemini 2.5, and DeepSeek. For each instance, we collect the predictions from these three models and determine the final output based on majority agreement—specifically, a label is selected only

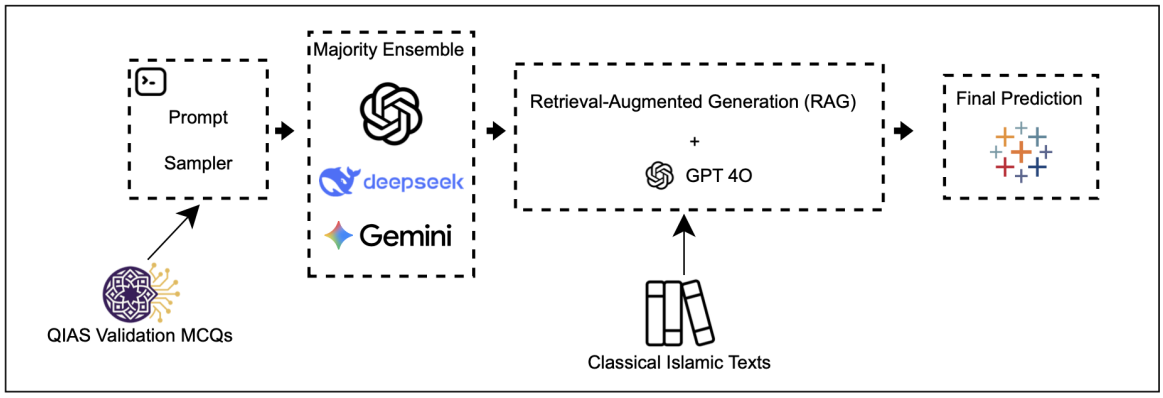

Figure 2: Overview of our ensemble-RAG pipeline combining LLMs and classical Islamic texts for answering QIAS 2025 MCQs.

if it is endorsed by at least two of the three models.

3. **Retrieval-augmented re-prompting (R^2P):** For cases where the ensemble stage results in uncertainty (i.e., no option reaches the required majority), we apply a retrieval-augmented re-prompting strategy. This approach involves:

   - Dense-only retrieval over classical Islamic texts: Arabic-LaBSE (768-d, mean-pooled, L2-normalized; inner-product) + FAISS IndexFlatIP on chunks 180–220 tokens (overlap 40–50) **retrieving the top-10** relevant passages.
   - Re-ranking these retrieved passages using a hybrid BM25 and cross-encoder scorer to select the top 3 most relevant passages.
   - Re-prompting the GPT-4o model with these carefully selected passages to produce a refined final prediction.

5 Experimental Setup

Data splits. We used the official dataset provided by the organizers, comprising 700 validation items and 700 test items, without additional splitting.

Hyper-parameters. To promote diversity while maintaining coherence in generation, we adopt the following settings: temperature = 0.2, top-p = 0.95, and a maximum of 512 output tokens.

Models considered. We evaluate the following models: GPT-4o, Gemini 2.5-Flash, Qwen-Plus, and DeepSeek-V3 [1]. Only the top three performers are included in the ensemble.

Evaluation metrics. We evaluate performance solely based on accuracy, measured as the percentage of questions for which the model's prediction exactly matches the correct answer, using the official **Task2_MCQ_Test_gold_labels** provided by the organizers.

6 Results

The performance of the evaluated models under different learning scenarios (zero-shot, 3-shot, and 5-shot) is summarized in Table 1. GPT-4o consistently demonstrated strong results across all settings, with a slight improvement observed in the 5-shot scenario. Gemini 2.5 exhibited a substantial performance increase when moving from zero-shot to few-shot learning conditions. This notable improvement can be attributed mainly to Gemini's initial difficulty in strictly adhering to task instructions in the zero-shot setting. Despite clear directives—such as prompts explicitly stating, "`Final Answer (letter only [A, B, C, D]) DO NOT output your thinking process or any other text except [A, B, C, D]:`"—Gemini often generated excessively detailed outputs, frequently exceeding the maximum token limit, leading to incomplete or empty responses. However, providing few-shot examples substantially improved Gemini's ability to comply with the task requirements, resulting in competitive accuracy. DeepSeek and Qwen-Plus also showed consistent improvement with the increase in examples provided, though

[1] All models were accessed via their official APIs between 15 - 20 July 2025.

their overall performance lagged slightly behind GPT-4o and Gemini, particularly in the few-shot scenarios. Our proposed system achieved an accuracy of 0.90 in the 5-shot setting, surpassing all individual models tested. This highlights the effectiveness of integrating few-shot prompting, model ensembling, and retrieval-augmented re-prompting. By combining the complementary strengths of multiple models and addressing uncertainty through targeted retrieval and refined prompting, our system demonstrates greater accuracy and robustness than any single model operating independently.

Model	Zero-shot	3-shot	5-shot
GPT-4o	0.85	0.85	0.86
Gemini 2.5	0.59	0.87	0.87
DeepSeek-V3	0.79	0.80	0.84
Qwen-Plus	0.77	0.77	0.78
Our Model			0.898

Table 1: Performance comparison of four models under zero-shot, 3-shot, and 5-shot settings. Scores are approximated to two decimal places.

7 Conclusion

In this study, we effectively combined few-shot chain-of-thought prompting, a majority-vote ensemble strategy, and retrieval-augmented re-prompting to address the challenging task of answering classical Arabic Islamic multiple-choice questions (MCQs). Our proposed system achieved superior performance, demonstrating the effectiveness of integrating ensemble strategies with retrieval methods for domain-specific knowledge tasks.

Acknowledgments

We would like to thank the organizers of the QIAS 2025 Shared Task for providing valuable resources and guidance throughout the competition. Special thanks to anonymous reviewers for their constructive comments, which helped improve the quality of this work.

References

Zabir Al Nazi, Md Rajib Hossain, and Faisal Al Mamun. 2025. Evaluation of open and closed-source llms for low-resource language with zero-shot, few-shot, and chain-of-thought prompting. *Natural Language Processing Journal*, 10:100124.

Wafa Bazzi and Mervat Gaith. 2025. The wonders of rag: Streamlining knowledge with advanced techniques systematic literature review report. Technical report.

Abdessalam Bouchekif, Samer Rashwani, Mohammed Ghaly, Mutaz Al-Khatib, Emad Mohamed, Wajdi Zaghouani, Heba Sbahi, Shahd Gaben, and Aiman Erbad. 2025a. Qias 2025: Overview of the shared task on islamic inheritance reasoning and knowledge assessment. In *Proceedings of The Second Arabic Natural Language Processing Conference, Arabic-NLP 2025, Suzhou, China, November 5–9, 2025*. Association for Computational Linguistics.

Abdessalam Bouchekif, Samer Rashwani, Heba Sbahi, Shahd Gaben, Mutaz Al-Khatib, and Mohammed Ghaly. 2025b. Assessing large language models on islamic legal reasoning: Evidence from inheritance law evaluation. In *Proceedings of The Second Arabic Natural Language Processing Conference (Arabic-NLP 2025)*, Suzhou, China. Association for Computational Linguistics.

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. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Hamed Meshkin, Joel Zirkle, Ghazal Arabidarrehdor, Anik Chaturbedi, Shilpa Chakravartula, John Mann, Bradlee Thrasher, and Zhihua Li. 2024. Harnessing large language models' zero-shot and few-shot learning capabilities for regulatory research. *Briefings in Bioinformatics*, 25(5):bbae354.

Ebtehal H. Omoush and Rawan Ghnemat. 2025. Advancing arabic medical question answering systems with rag and llms integration. In *Proceedings of the International Conference on New Trends in Computing Sciences (ICTCS)*, pages 511–516. IEEE.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *In Proceedings of the Advances in Neural Information Processing Systems*, 35:24824–24837.

A Appendix

Prompt template:

"You are an Islamic knowledge expert

tasked with solving multiple-choice questions. Think step-by-step, carefully justify your reasoning, and then select the correct answer clearly.Here are some examples to follow:

مثال ١: السؤال: ما هو القول القديم للشافعي في صوم أيام التشريق؟
الخيارات:
A) لا يجوز صومها مطلقاً.
B) يجوز صومها للمتمتع إذا عدم الهدي عن الأيام الثلاثة الواجبة في الحج. يجوز صومها لمن لم يجد الهدي فقط. يجوز صومها للمسافر فقط.
C) خطوات التفكير:
أيام التشريق هي الأيام الثلاثة التي تلي يوم النحر (يوم عيد الأضحى) في موسم الحج.
القول القديم للشافعي يشير إلى رأيه الأول قبل أن يرجح عليه قوله الجديد، وقد يختلف عنه في بعض المسائل.
القول الجديد والمعتمد عند الشافعية هو تحريم صوم أيام التشريق مطلقاً، حتى للمتمتع إذا لم يجد الهدي، وهذا هو القول الذي عليه المتأخرون من أصحاب الشافعي.
في كتب الفقه على المذهب الشافعي، يُذكر أن من أقواله القديمة جواز صيام أيام التشريق في حالات مخصوصة.
من تلك الحالات: أن يكون الحاج متمتعًا (أي جمع بين العمرة والحج في سفرة واحدة مع فاصل)، ولم يجد الهدي، فيلزمه صيام ثلاثة أيام في الحج، ويسمح له بصيامها في أيام التشريق.
هذا القول موثق ومعروف في كتب المذهب عند الحديث عن اختلاف القولين القديم والجديد للشافعي.
الإجابة النهائية: B

⋮
... the other 4 examples...
⋮

Now answer the question below by selecting the appropriate answer:
Question: {question}
Choices: {formatted_options}
Final Answer (letter only [A, B, C, D])
DO NOT output your thinking process:"