Athar at QIAS2025: LLM-based Question Answering Systems for Islamic Inheritance and Classical Islamic Knowledge

Yossra Noureldien[1] **Hassan Suliman**[2] **Farah Attallah**[1]
Abdelrazig Mohamed[1] **Sara Abdalla**[3]
[1]University of Khartoum [2]African Institute for Mathematical Sciences
[3]International Islamic University Malaysia
{yossra.noureldien, farah.hassan, abdelrazig.mohamed}@uofk.edu
hassan.suliman017@gmail.com ssa_abdalla@iium.edu.my

Abstract

The intersection of Arabic linguistic complexity and specialized reasoning presents a key challenge for Islamic question-answering systems, particularly in the under-addressed area of inheritance law. This paper presents our methodology for the QIAS2025 shared task, assessing LLM capabilities in Islamic knowledge through two subtasks: Inheritance Reasoning (ᶜilm al-mawārīth) and General Islamic Assessment. A zero-shot, prompt-based approach with DeepSeek-R1 (deepseek-reasoner) addresses the former, while a three-stage RAG pipeline handles the latter. Our approaches achieved competitive results, with an accuracy of 0.704 for inheritance reasoning (10th place/15 teams) and 0.9272 for general Islamic assessment (2nd place/10 teams), demonstrating the efficacy of tailored model strategies for religious QA. These insights pave the way for more culturally and linguistically adaptive AI systems in Islamic scholarly applications.

1 Introduction

Arabic Islamic question-answering (QA) systems face dual challenges of linguistic complexity and specialized domain knowledge requirements. Inheritance law (ᶜilm al-mawārīth) and classical Islamic scholarship remain computationally underexplored, despite growing demand for accessible religious knowledge through digital platforms.

Historically, Islamic QA relied on symbolic systems such as rule-based expert systems and ontology-driven frameworks (Alshahad and Abutiheen, 2015; Zouaoui and Rezeg, 2021) or traditional information retrieval, effective in structured domains like inheritance law, but limited in handling linguistic variation and complex reasoning.

Modern large language models (LLMs) (e.g., GPT series (Radford et al., 2018)) and Arabic-centric models (e.g., ALLaM (Bari et al., 2024)) offer greater flexibility and cultural alignment, yet their evaluation in specialized domains like inheritance and multi-disciplinary Islamic studies remains scarce, motivating the need for dedicated benchmarks.

The QIAS2025 shared task (Bouchekif et al., 2025a) establishes a benchmark for evaluating LLMs across two domains: SubTask 1, *Islamic Inheritance Reasoning*, uses multiple-choice questions (MCQs) to test rule application, proportional reduction (*ᶜawl*), exclusion (*ḥajb*), and precise share allocation; and SubTask 2, *Islamic Studies Assessment*, comprises MCQs derived from 23 classical Islamic texts spanning Qur'anic studies, ḥadīth, fiqh, uṣūl al-fiqh, and sīrah

This paper presents our approach to the QIAS2025 shared task, addressing the two subtasks:

- SubTask 1: Zero-shot DeepSeek-R1 pipeline for inheritance reasoning, with output-constrained prompting and regex-based label extraction.
- SubTask 2: Three-stage hybrid RAG pipeline for general Islamic assessment, combining dense and BM25 retrieval with LLM reranking.
- Results: Competitive leaderboard rankings, 10th/15 for inheritance reasoning and 2nd/10 for general Islamic assessment.

2 Related Work

Recent advancements in transformer-based architectures and fine-tuning methodologies have significantly shaped Arabic Islamic question-answering systems. The field has seen significant progress through shared tasks such as Qur'an QA 2022 (Malhas et al., 2022), with notable contributions including Basem et al. (2025) expanding the Qur'an QA dataset to 1,895 question-answer pairs, achieving MAP@10 of 0.36 and 75% success in zero-

Proceedings of The Third Arabic Natural Language Processing Conference, pages 914–922

answer detection, and Abdallah et al. (2024) introducing ArabicaQA with over 89,000 questions for comprehensive Arabic QA benchmarking.

Domain-specific approaches have emerged for Islamic knowledge processing. For instance, Adel et al. (2023) developed AraQA for authentic religious texts with careful dataset curation to reduce misleading answers, while Alan et al. (2025) proposed MufassirQAS, a RAG-based system outperforming ChatGPT through vector databases and fact-checking mechanisms. Additionally, Qamar et al. (2024) developed a large-scale dataset with 73,000+ QA pairs for Tafsir and Ahadith, revealing limitations in automatic evaluation metrics and emphasizing the need for human expert assessment in religious QA contexts. Sibaee et al. (2025) have also addressed Arabic language model assessment challenges, with comprehensive studies revealing significant performance variations across cultural and specialized domains.

Despite these advances, significant gaps remain particularly in computational approaches to Islamic inheritance law (ᶜilm al-mawārīth), which requires precise numerical calculations. While Alshammary et al. (2024) demonstrated promising results with their RFPG RAG model, most prior work focuses on extractive QA or general Islamic content. Most recently, Bouchekif et al. (2025b) conducted a large-scale evaluation of seven LLMs on Islamic inheritance, finding strong results for reasoning-oriented models but major errors in open-source Arabic ones.

Our participation in QIAS2025 explores both specialized inheritance reasoning and broader Islamic knowledge assessment through domain-specific MCQs. By tackling these distinct challenges, our work contributes novel empirical insights into the capabilities and limitations of LLMs in religious question answering.

3 Data

The QIAS2025 shared task provided two datasets from distinct domains, summarized in Table 1.

For SubTask 1, the dataset comprises Arabic multiple-choice questions in Islamic inheritance (ᶜIlm al-Mawārīth), each with six options (A–F). It includes 20,000 training, 1,000 development, and 1,000 test examples, plus 3,165 IslamWeb fatwas as extra data. For SubTask 2, the dataset consists of multiple-choice questions with four options (A–D) drawn from classical Islamic texts spanning fiqh, ḥadīth, tafsīr, and other disciplines. The development set has 700 questions from 21 books, and the test set has 1,000 questions from 23 books (including two unseen in the development set).

Task	Train	Dev	Test	Extra Data
SubTask 1	20,000	1,000	1,000	3,165 fatwas
SubTask 2	–	700	1,000	23 classical texts

Table 1: Dataset statistics and additional resources for the QIAS2025 subtasks.

4 System Overview

Our proposed solution addressed the QIAS2025 shared task through two distinct pipelines, each tailored to the requirements of its respective subtask.

For Subtask 1, we tested several reasoning-capable models via in-context prompting and selected DeepSeek-R1 for its strong Arabic reasoning and cost-effective API. For Subtask 2, we normalized the provided corpus of 23 classical books spanning HTML and DOCX formats, enabling the construction of a unified hybrid index. We experimented with several retrieval strategies, including retrieving surrounding passages and hybrid fusion, and found that applying a LLM reranker yielded the best approach. Across both subtasks, the design emphasizes robustness to varied encodings, domain specificity, and consistent answer formatting.

4.1 SubTask 1: Islamic Inheritance Reasoning

To handle the complex reasoning required in Islamic inheritance (ᶜIlm al-Mawārīth), we employed a zero-shot, prompt-based approach with the `deepseek-reasoner` model (DeepSeek-R1-0528) via API (DeepSeek-AI et al., 2025). No fine-tuning was performed; instead, the model was directly evaluated in zero-shot mode, leveraging its Arabic reasoning capability. The domain-specific Arabic prompt is illustrated in Figure 1, and the English translation is provided in Appendix A.

Prompt Design. The prompt included the question, six answer options, and a strict instruction to output only the correct choice in the format `<answer> X </answer>`, producing deterministic, machine-readable results. This format eliminated ambiguity and avoided the mixing of Arabic text with answer labels. We did not experiment with alternative prompt formats, as our primary ob-

System Prompt:

أنت عالم متخصص في علم الفرائض المواريث.
تجاوب باستخدام الفقه الإسلامي والحساب.

User Prompt:

اجب عن السؤال التالي بشكل مباشر دون
شرح ثم ضع حرف الخيار الصحيح فقط داخل الوسم مثل
<answer> B </answer>

السؤال:
QUESTION

الخيارات:
A) OPTION_A B) OPTION_B C) OPTION_C
D) OPTION_D E) OPTION_E F) OPTION_F

اكتب الجواب فقط بهذه الصيغة:
<answer> X </answer>

Figure 1: Zero-shot prompt used in SubTask 1

jective was to suppress free-form "thinking" outputs and enforce consistent, extractable answers.

Pipeline Execution. Following prompt construction, each instance was submitted to the DeepSeek API using fixed decoding parameters. Model responses were then parsed using a regular expression to extract the predicted label. All model outputs and extracted answers were logged per instance to ensure reproducibility and support error analysis. The pipeline operated in a CPU-only environment via the paid API tier, ensuring stable latency and no token constraints.

4.2 Subtask 2: Islamic Studies Assessment

For this task, a RAG pipeline was adopted to manage the semantic diversity of questions, heterogeneous text formats, and the need for source-grounded reasoning. The pipeline was inspired by methodologies from the RAG-Challenge-2 repository[1], and the overall workflow is shown in Figure 2. Translation of Arabic text is available in Appendix A.

Corpus Ingestion and Indexing. After normalizing the corpus into plain text for consistency across formats, each book was segmented into semantically coherent passages using LangChain's `RecursiveCharacterTextSplitter`, configured with a chunk size of 500 characters and a 50 character overlap to preserve contextual continuity. This overlap mitigates semantic fragmentation across chunk boundaries, a technique commonly used in multilingual and Arabic NLP. Each chunk was embedded using OpenAI's

[1] https://github.com/IlyaRice/RAG-Challenge-2

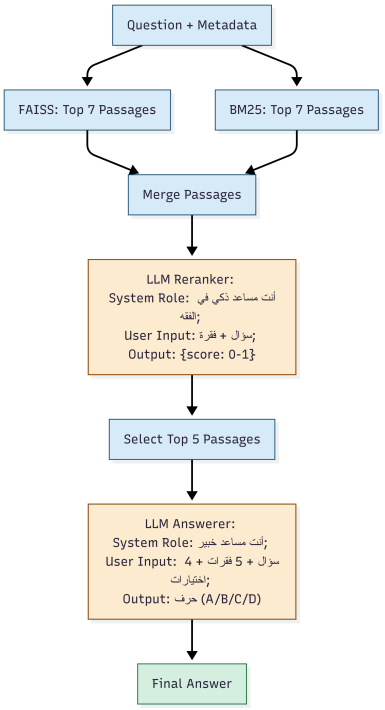

Figure 2: Pipeline used in SubTask 2

`text-embedding-3-large` model, producing dense semantic vectors. These were indexed using FAISS's `IndexFlatIP` for dense similarity search (Johnson et al., 2021). In parallel, sparse representations were computed using the Okapi BM25 algorithm (Robertson and Zaragoza, 2009) to support lexical-level retrieval. Each chunk was also stored with metadata such as book title to support traceability and analysis.

Hybrid Retrieval and Reranking. To leverage both semantic and lexical retrieval, we adopted a hybrid strategy without score fusion. Instead of α-weighted interpolation, we performed parallel top-k retrieval: the top 7 passages were retrieved independently from FAISS and BM25, producing a 14-passage candidate set that maintained both semantic relevance and lexical precision. Our methodology prioritized demonstrating the hybrid approach's optimal performance for Islamic inheritance QA, with individual retrieval component analysis considered beyond the current work's scope and suitable for future comparative studies. A lightweight reranking stage using GPT-4o-mini was then applied to semantically compare the question with each of the 14 retrieved passages and select the 5 most relevant ones for the final answer generation stage.

Answer Generation. In the final stage, the top 5 passages, selected by the reranker, were used as contextual input for answer generation with GPT-4o. These passages were injected into a constrained multiple-choice question prompt, which explicitly instructed the model to return only a single answer choice, formatted within an `<answer>` tag. This strict output format minimized generation variability. Importantly, no model fine-tuning was performed at any stage. Both the reranking and answer generation components operated in zero-shot inference mode, relying solely on carefully crafted prompts and high-quality context to guide the model's reasoning.

5 Results

5.1 Evaluation and Performance

Accuracy was used as the primary evaluation metric across both subtasks, defined as:

$$Accuracy = \frac{Correct\_predictions}{Total\_samples}$$

This metric directly reflects the proportion of correctly answered questions, making it appropriate for multiple-choice QA tasks. Our evaluation was carried out on both the development and test sets, with results summarized in Table 2.

Task	System	Devset	Testset
Subtask 1	DeepSeek API with Direct Prompting	0.885	0.704
Subtask 2	RAG with Hybrid Retrieval and LLM Reranker	0.914	0.927

Table 2: Accuracy of Development and Test Sets

Leaderboard rankings were determined using the test set. In **Subtask 1**, our system ranked 10th out of 15 teams, with the top score reaching 0.972. In **Subtask 2**, we placed 2nd out of 10 teams, with the best score recorded at 0.937.

Beyond leaderboard ranks, we provide statistical analysis using Wilson confidence intervals, which offer superior boundary handling for proportion estimates. In Subtask 1 (6 options; chance = 16.7%), our system achieved 70.4% accuracy on N=1000 with CI [67.5%, 73.2%]. In Subtask 2 (4 options; chance = 25%), we applied majority-rule deduplication yielding N=729 samples, with accuracy of 92.32% and CI [90.16%, 94.04%]. Both confidence intervals demonstrate substantial separation from chance levels, indicating robust performance well beyond random selection. The Wilson interval methodology ensures reliable statistical inference even near boundary conditions, while the substantial sample sizes support the stability of our accuracy estimates, though formal hypothesis testing could further strengthen these findings.

5.2 Results Analysis

For **Subtask 1**, our evaluation highlights three main trends:

- **General Performance Gap:** Accuracy dropped from 88.5% on the development set to 70.4% on the test set (–18.1%). Test questions were longer (140 vs. 97 characters, +43.8%) and answer options were much longer (653 vs. 117, 5.57×), as shown in Figure 3, suggesting a possible domain shift with added lexical variety and detail that increased reasoning difficulty.

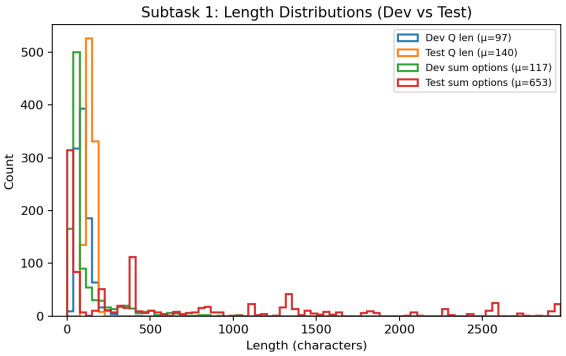

Figure 3: Subtask 1: Dev/Test sets length distributions (questions & options). Test is longer on average.

- **Level Sensitivity:** Questions were labeled Beginner or Advanced. On the development set, accuracy was 90.0% for Advanced and 87.0% for Beginner. On the test set, it dropped to 70.6% and 70.2% respectively. The similar decline across both levels indicates the drop was driven by overall complexity rather than by a specific difficulty category.

- **Error Patterns:** Accuracy varied by heir category. For example, questions mentioning أخت شقيقة/لأم/لأب (sisters) had an accuracy of 0.644, while those mentioning زوجة/زوج (spouse) reached 0.794. These results are based on *inclusive* category, meaning each question is counted under every heir it in-

volves. Appendix B contains a qualitative error example.

For **Subtask 2**, performance was consistent across development and test sets. Analysis focuses on three aspects:

- **Error Causes:** Two main issues contributed to mistakes: (i) relevant passages were not retrieved; and (ii) even when correct passages were retrieved, the LLM sometimes failed to select the right option. Examples of errors are available in Appendix B.
- **Performance by level:** Questions were labeled Beginner, Intermediate or Advanced. Accuracy was 96.7% on Beginner questions, 95.3% on Intermediate and 78.7% on Advanced. This shows that the system handled easier questions well but dropped on more challenging ones.
- **Performance by Source:** Results varied by book. Accuracy was lowest on فتح المغيث (63.6%) but reached (100%) on sources such as الفقه المنهجي and الرحيق المختوم. This indicates that differences in style, terminology, and content across sources significantly affected retrieval and answer selection. Figure 4 illustrates the top 5 lowest and highest sources by accuracy.

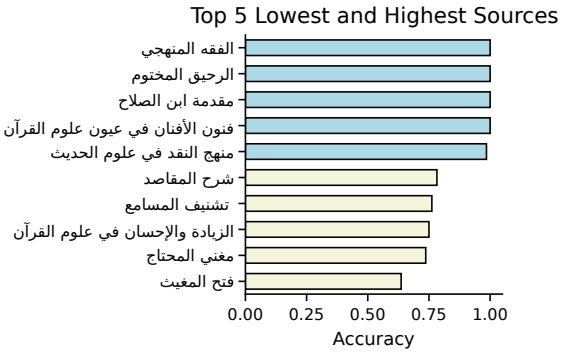

Figure 4: Top 5 lowest and highest sources by accuracy.

The findings highlights that Subtask 1 requires stronger complex reasoning, whereas Subtask 2 would benefit from enhanced retrieval and LLM comprehension to achieve more reliable answer selection.

6 Conclusion

This paper presented our systems for the QIAS2025 shared task on Islamic Inheritance Reasoning and Classical Islamic Knowledge. We implemented a direct prompting approach for Subtask 1, achieving 70.4% accuracy, and a hybrid RAG pipeline combining FAISS and BM25 retrieval with GPT-4o-mini reranking for Subtask 2, achieving 92.72%. The analysis revealed that inheritance reasoning demands careful handling of longer and more complex scenarios, while Subtask 2 highlighted retrieval performance variation across diverse classical sources. While our systems demonstrated excellent performance, broader deployment requires addressing critical ethical and performance challenges.

Ethical Considerations

While the QA systems presented in this paper demonstrates excellent accuracy performance, the broader deployment of such systems requires careful attention to inherent challenges. These challenges include hallucination and misinformation risks (Khalila et al., 2025), algorithmic bias affecting diverse religious communities (Gupta and Giannoccaro, 2024), privacy concerns with sensitive spiritual data (Liu et al., 2025), and questions about authenticity in AI-mediated spiritual experiences (Alkhouri, 2024). Successful implementation requires inclusive algorithm design (Habib, 2025), transparent accountability measures (Sarker, 2024), and human oversight to ensure responsible and effective deployment in religious contexts. Such measures are essential for ensuring responsible AI deployment that respects the diversity and sensitivity inherent in religious contexts.

Acknowledgments

We thank the anonymous reviewers for their helpful comments.

References

Abdelrahman Abdallah, Mahmoud Kasem, Mahmoud Abdalla, Mohamed Mahmoud, Mohamed Elkasaby, Yasser Elbendary, and Adam Jatowt. 2024. ArabicaQA: A Comprehensive Dataset for Arabic Question Answering. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2049–2059, Washington DC USA. ACM.

Yousef Adel, Mostafa Dorrah, Ahmed Ashraf, Abdallah ElSaadany, Mahmoud Mohamed, Mariam Wael, and

Ghada Khoriba. 2023. AraQA: An Arabic Generative Question-Answering Model for Authentic Religious Text. In *2023 11th International Japan-Africa Conference on Electronics, Communications, and Computations (JAC-ECC)*, pages 235–239, Alexandria, Egypt. IEEE.

Ahmet Yusuf Alan, Enis Karaarslan, and Ömer Aydin. 2025. A RAG-based Question Answering System Proposal for Understanding Islam: MufassirQAS LLM. *arXiv preprint*. ArXiv:2401.15378 [cs].

Khader I. Alkhouri. 2024. The Role of Artificial Intelligence in the Study of the Psychology of Religion. *Religions*, 15(3):290.

H. F. Alshahad and Z. A. Abutiheen. 2015. Computation of inheritance share in islamic law by an expert system using decision tables. *Quarterly Adjudicated Journal for Natural and Engineering Research and Studies*, 1:105–114.

Mitha Alshammary, Md Nahiyan Uddin, and Latifur Khan. 2024. RFPG: Question-Answering from Low-Resource Language (Arabic) Texts using Factually Aware RAG. In *2024 IEEE 10th International Conference on Collaboration and Internet Computing (CIC)*, pages 107–116, Washington, DC, USA. IEEE.

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, 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. *Preprint*, arXiv:2407.15390.

Mohamed Basem, Islam Oshallah, Baraa Hikal, Ali Hamdi, and Ammar Mohamed. 2025. Optimized Quran Passage Retrieval Using an Expanded QA Dataset and Fine-Tuned Language Models. In Faisal Saeed, Fathey Mohammed, Errais Mohammed, Shadi Basurra, and Mohammed Al-Sarem, editors, *Advances on Intelligent Computing and Data Science II*, volume 255, pages 244–254. Springer Nature Switzerland, Cham. Series Title: Lecture Notes on Data Engineering and Communications Technologies.

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*. 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.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Brij Gupta and Ivan Giannoccaro, editors. 2024. *Challenges in Large Language Model Development and AI Ethics:*. Advances in Computational Intelligence and Robotics. IGI Global.

Zainal Habib. 2025. Ethics of Artificial Intelligence in Maqāṣid Al-Sharīa's Perspective. *KARSA Journal of Social and Islamic Culture*, 33(1):105–134.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.

Zahra Khalila, Arbi Haza Nasution, Winda Monika, Aytug Onan, Yohei Murakami, Yasir Bin Ismail Radi, and Noor Mohammad Osmani. 2025. Investigating Retrieval-Augmented Generation in Quranic Studies: A Study of 13 Open-Source Large Language Models. *International Journal of Advanced Computer Science and Applications*, 16(2).

Feng Liu, Jiaqi Jiang, Yating Lu, Zhanyi Huang, and Jiuming Jiang. 2025. The ethical security of large language models: A systematic review. *Frontiers of Engineering Management*, 12(1):128–140.

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.

Faiza Qamar, Seemab Latif, and Asad Shah. 2024. Techniques, datasets, evaluation metrics and future directions of a question answering system. *Knowledge and Information Systems*, 66(4):2235–2268.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. Preprint.

Stephen E. Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Iqbal H. Sarker. 2024. LLM potentiality and awareness: a position paper from the perspective of trustworthy and responsible AI modeling. *Discover Artificial Intelligence*, 4(1):40.

Serry Sibaee, Omer Nacar, Adel Ammar, Yasser Al-Habashi, Abdulrahman Al-Batati, and Wadii Boulila. 2025. From Guidelines to Practice: A New Paradigm for Arabic Language Model Evaluation. *arXiv preprint*.

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.

Appendices

A Translations of Figures' Arabic Text

Figure 1: Zero-shot prompt used in SubTask 1

- **System Prompt:** You are a scholar specialized in inheritance law (ʿilm al-farāʾiḍ), answer using Islamic jurisprudence and arithmetic.
- **User Prompt:** Answer the following question directly without explanation, then place only the correct option letter inside the tag e.g. `<answer> B </answer>`.
- **Label:** Write the answer only in this format: `<answer> X </answer>`

Figure 2: Pipeline used in Subtask 2

- **LLM Reranker**
  System Role: You are an intelligent assistant in Islamic jurisprudence.
  User Input: Question + Passage.
  Output: score: 0-1
- **LLM Answerer**
  System Role: You are an expert assistant.
  User Input: Question + 5 Passages + 4 Options.
  Output: (A/B/C/D)

B Qualitative Errors Examples

Subtask 1 (Inheritance Reasoning):

- Multiple-choice inheritance question in which **Gold = D**, **Pred = A** (Figure B1).

Subtask 2 (Islamic Studies Assessment):

- Answer-absent: the correct answer is missing from hybrid retrieval and thus absent from the reranked context; the model guessed incorrectly. **Gold = B**, **Pred = A** (Figure B2).
- Evidence-present: The correct option is supported in the reranked context, but the model chose a different option. **Gold = C**, **Pred = A** (Figure B3).

Question ID: 2173_nl7p9x5o_16
أجب عن السؤال التالي بشكل مباشر
دون شرح، ثم ضع حرف الخيار الصحيح
فقط داخل الوسم مثل
`<answer>` B `</answer>`

السؤال:
مات وترك: أخ لأم (2) و عم الأب لأب (3)
و ابن أخ شقيق (2) و أب أب الأب و أم
الأب و ابن ابن و أب، كم عدد الأسهم
بعد التصحيح التي يحصل عليها ابن ابن؟

الخيارات:
A) 0 سهم
B) 4 سهم
C) 6 سهم
D) 5 أسهم
E) 3 سهم
F) 7 سهم

اكتب الجواب فقط بهذه الصيغة:
`<answer>` X `</answer>`

Model Response:
`<answer>` A `</answer>`
Extracted: A

Figure B1: Example Error from Subtask 1

Question: بعد رد أبي بكر الصديق جوار ابن الدغنة، تعرض لابتلاء من أحد سفهاء قريش، وفي هذا الموقف الصعب، كرر قولًا يعكس صبره العظيم وإيمانه العميق. ما هو هذا القول الذي كرره ثلاث مرات؟

Options:
A) "إحسبي الله ونعم الوكيل"
B) "إيا رب، ما أحلمك"
C) "اللهم اجرني في مصيبتي!"
D) "الحمد لله على كل حال!"

Gold: B **Pred:** A

Reranked top-5 passages:

- P1 (book=sira, score=0.9):
  رددت إليك جوارك فقال له يا ابن أخي لعله آذاك أحد من قومي قال لا ولكني أرضى بجوار الله ولا أريد أن أستجير بغيره قال فانطلق الى المسجد فأرد علي جواري علانية كما أجرتك علانية قال فانطلقا فخرجا حتى أتيا المسجد فقال الوليد هذا عثمان قد جاء يرد علي جواري قال صدق قد وجدته وفيا كريم الجوار ولكني قد أحببت أن لا استجير بغير الله فقد رددت عليه جواره ثم انصرف عثمان وليبد ابن ربيعة بن مالك بن جعفر بن كلاب في مجلس من قريش يتشدهم فجلس معهم عثمان فقال لبيد ** ألا كل شيء ما خلا الله باطل **
- P2 (book=sira, score=0.9):
  عَلَيْهِ وَسَلَّمَ: "نَعَمْ، أَنَا الَّذِي أَقُولُ ذَلِكَ". قَالَ: فَلَقَدْ رَأَيْتُ رَجُلًا مِنْهُمْ أَخَذَ بِمَجْمَعِ رِدَائِهِ. قَالَ: فَقَامَ أَبُو بَكْرٍ رَضِيَ اللَّهُ عَنْهُ دُونَهُ، وَهُوَ يَبْكِي وَيَقُولُ: أَتَقْتُلُونَ رَجُلًا أَنْ يَقُولَ رَبِّيَ اللَّهُ؟ ثُمَّ انْصَرَفُوا عَنْهُ، فَإِنَّ ذَلِكَ لَأَشَدُّ مَا رَأَيْتُ قُرَيْشًا نَالُوا مِنْهُ قط.
- P3 (book=sira, score=0.9):
  أَهْلِ الشِّرْكِ، وَأَصْحَابِي وَأَهْلُ دِينِي يَلْقَوْنَ مِنَ الْبَلَاءِ وَالْأَذَى فِي اللَّهِ مَا لَا يُصِيبُنِي، لَنَقْصٌ كَبِيرٌ فِي نَفْسِي. فَمَشَى إِلَى الْوَلِيدِ بْنِ الْمُغِيرَةِ، فَقَالَ لَهُ: يَا أَبَا عَبْدِ شَمْسٍ، وَفَّتْ ذِمَّتُكَ، قَدْ رَدَدْتُ إِلَيْكَ جِوَارَكَ، فَقَالَ لَهُ: (لِمَ) [1] يَا بن أَخِي؟ لَعَلَّهُ آذَاكَ أَحَدٌ مِنْ قَوْمِي، قَالَ:: لَا، وَلَكِنِّي أَرْضَى بِجِوَارِ اللَّهِ، وَلَا أُرِيدُ أَنْ أَسْتَجِيرَ بِغَيْرِهِ؟ قَالَ: فَانْطَلِقْ إِلَى الْمَسْجِدِ، فَارْدُدْ
- P4 (book=rwd, score=0.3):
  قَالَ أَبُو قُحَافَةَ لِأَبِي بَكْرٍ يَا بُنَيَّ إِنِّي أَرَاكَ تَتَّبِعُ رِقَابًا ضِعَافًا، فَلَوْ أَنَّكَ إِذْ فَعَلْتَ مَا فَعَلْتَ أَعْنَقْتَ رِجَالًا جُلَّدًا يَمْنَعُونَ ذُمَارَكَ، وَيَقُومُونَ دُونَكَ؟ قَالَ فَقَالَ أَبُو بَكْرٍ رَضِيَ اللهُ عَنْهُ يَا أَبَتِ إِنِّي إِنَّمَا أُرِيدُ مَا أُرِيدُ لِلَّهِ عَزَّ وَجَلَّ، قَالَ فَيَحْدَثُ أَنَّهُ مَا نَزَلَ هَؤُلَاءِ الْآيَاتُ إِلَّا فِيهِ وَفِيمَا قَالَ لَهُ أَبُوهُ {فَأَمَّا مَنْ أَعْطَى وَاتَّقَى وَصَدَّقَ بِالْحُسْنَى} [اللَّيْلِ 5، 6] . إِلَى قَوْلِهِ تَعَالَى: {وَمَا
- P5 (book=sira, score=0.3):
  قال ابن اسحاق وحدثني هشام بن عروة عن أبيه قال كان ورقة بن نوفل يمر به وهو يعذب بذلك وهو يقول أحد أحد فيقول أحد أحد والله يا بلال ثم يقبل على أمية بن خلف ومن يصنع ذلك به من بني جمح فيقول أحلف بالله لئن قتلتموه على هذا لأتخذنه حنانا حتى مر به أبو بكر الصديق ابن أبي قحافة رضي الله عنه يوما وهم يصنعون ذلك به وكانت دار أبي بكر في بني جمح فقال لأمية بن خلف ألا تتقي الله في هذا المسكين حتى متى قال أنت الذي أفسدته فأنقذه مما ترى فقال أبو بكر أفعل عندي خلاف أسود أجلد منه وأقوى على دينك أعطيكه به قال قد

Figure B2: Example Error from Subtask 2 (Answer-absent)

Question: لماذا يُعتبر "الصحو الذي يعقب السكر" حالة أتم وأكمل من حالة "السكر" نفسها؟

Options:

A) لأن صاحب الصحو يتجنب المكروهات تمامًا، بينما صاحب السكر يقع فيها.

B) لأن صاحب السكر يكون في حالة وعي وإدراك تام، بينما صاحب الصحو يكون غائبًا عن التمييز.

C) لأن صاحب السكر قد يقع على المكروه دون أن يدري، بينما صاحب الصحو يختار المكروه عن تمييز، مما يدل على تمكن أكبر.

D) لأن حالة السكر مؤقتة بينما حالة الصحو دائمة.

Gold: C **Pred:** A

Reranked top-5 passages:

- P1 (book=mdarij, score=1.0):
  قَالَ: الصَّحْوُ: فَوْقَ السُّكْرِ، وَهُوَ يُذَاهِبُ مَقَامَ الْبَسْطِ، وَالصَّحْوُ: مَقَامٌ صَاعِدٌ عَنْ الِانْبِطَارِ، مُغْنٍ عَنِ الطَّلَبِ، طَاهِرٌ مِنَ الْحَرَجِ، فَإِنَّ السُّكْرَ إِنَّمَا هُوَ فِي الْحَقِّ، وَالصَّحْوَ إِنَّمَا هُوَ بِالْحَقِّ، كُلُّ مَا كَانَ فِي عَيْنِ الْحَقِّ لَمْ يَخْلُ مِنْ خَيْرَةٍ، لَا خَيْرَةَ الشُّبْهَةِ، بَلْ خَيْرَةَ مُشَاهَدَةِ نُورِ الْعِزَّةِ، وَمَا كَانَ بِالْحَقِّ لَمْ يَخْلُ مِنْ صِحَّةٍ. وَلَمْ تَجِفْ عَلَيْهِ نَقِيصَةٌ، وَلَمْ تَتَعَاوَرْهُ عِلَّةٌ.
- P2 (book=mdarij, score=1.0):
  الِاتِّصَالِ، وَأَيْضًا فَالسُّكْرُ فَنَاءٌ، وَالصَّحْوُ بَقَاءٌ.
  وَأَيْضًا فَالسُّكْرُ غَيْبَةٌ وَالصَّحْوُ حُضُورٌ، وَأَيْضًا فَالسُّكْرُ عَلَيَّةٌ وَالصَّحْوُ تَمَكُّنٌ، وَأَيْضًا فَالسُّكْرُ كَالْغَيْمِ وَالصَّحْوُ كَالْيَقَظَةِ.
  وَبَعْضُهُمْ يُفَضِّلُ مَقَامَ السُّكْرِ عَلَى مَقَامِ الصَّحْوِ وَيَقُولُ: لَوْلَا الْبَقِيَّةُ الَّتِي بَقِيَتْ فِيهِ لَمَا صَحَا، وَيُنْشِدُ مُتَمَثِّلًا:
  وَمَهْمَا بَقِي لِلصَّحْوِ فِيكَ بَقِيَّةٌ ... يَجِدْ نَحْوَكَ اللَّاحِي سَبِيلًا إِلَى الْعَذْلِ
- P3 (book=risala, score=0.9):
  ومن ذَلِك الصحو والسكر فالصحو رجوع إِلَى الإحساس بَعْد الغيبة والسكر غيبة بوارد قوي والسكر زيادة عَلَى الغيبة من وجه، وَذَلِك أن صاحب السكر قَدْ يَكُون مبسوطا إِذَا لَمْ يكن مستوفيا فِي سكره، وَقَدْ يسقط أخطار الأشياء عَن قلبه فِي حال سكره، وتلك حال المتساكر الَّذِي لَمْ يستوفه الوارد فيكون للإحساس فِيهِ مساغ وَقَدْ يقوى سكره حَتَّى يَزِيد عَلَى الغيبة، فربما يَكُون صاحب السكر أشد غيبة من صاحب الغيبة إِذَا قوى سكره، وربما يَكُون صاحب الغيبة أتم فِي الغيبة من صاحب السكر إِذَا كَانَ متساكرا
- P4 (book=mdarij, score=0.9):
  وَهَذَا غَلَطٌ مَحْضٌ، لِمَا ذَكَرْنَا. نَعَمِ السُّكْرُ فَوْقَ الصَّحْوِ الْفَارِغِ، وَالسُّكْرَانُ بِالْمَحَبَّةِ خَيْرٌ مِنَ الصَّاحِي مِثْهَا، وَالصَّاحِي بِهَا خَيْرٌ مِنَ السُّكْرَانِ فِيهَا.
  قَوْلُهُ: " وَهُوَ يُذَاهِبُ مَقَامَ الْبَسْطِ " وَجْهُ الْمُنَاسَبَةِ بَيْنَهُمَا: أَنَّ الِانْبِسَاطَ لَا يَكُونُ إِلَّا مَعَ الصَّحْوِ، وَإِلَّا فَالسُّكْرُ لَا يَحْتَمِلُ الِانْبِسَاطَ.
- P5 (book=risala, score=0.9):
  واعلم أن الصحو عَلَى حسب السكر، فمن كَانَ سكره بحق كَانَ صحوه بحق، ومن كَانَ سكره بحفظ مشوبا كَانَ صحوه بحفظ صحيح مصحوبا، ومن كَانَ محقا فِي حاله كَانَ محفوظا فِي سكره، والسكر والصحو يشيران إِلَى طرف من التفرقة، وإذا ظهر من سلطان الحقيقة علم أن صفة العبد التبور والقهر.
  وَفِي معناه أنشدوا:
  إِذَا طلع الصباح لنجم راح ... تساوى فِيهِ سكران وصاح

Figure B3: Example Error from Subtask 2 (Evidence-present)