From RAG to Agentic RAG for Faithful Islamic Question Answering

Gagan Bhatia[1], **Hamdy Mubarak**[1], **Mustafa Jarrar**[2], **George Mikros**[2], **Fadi Zaraket**[3], **Mahmoud Alhirthani**[2], **Mutaz Al-Khatib**[4], **Logan Cochrane**[5], **Kareem Darwish**[1], **Rashid Yahiaoui**[2], **Firoj Alam**[1]
[1] Qatar Computing Research Institute, HBKU, Qatar,
[2] College of Humanities and Social Sciences, HBKU, Qatar, [3] Arab Center for Research and Policy Studies, Qatar, [4] College of Islamic Studies, HBKU, Qatar
[5] College of Public Policy, HBKU, Qatar
fialam@hbku.edu.qa

https://huggingface.co/collections/QCRI/islamic-knowledge-in-llms

Abstract

Large Language Models (LLMs) are increasingly used for Islamic question answering, where ungrounded responses may carry serious religious consequences. Yet standard MCQ/MRC-style evaluations[1] do not capture key real-world failure modes, notably free-form hallucinations and the ability to abstain when evidence is insufficient. To address this gap, we introduce ISLAMICFAITHQA, a 3,810-item bilingual (Arabic/English) *generative* benchmark with atomic single-gold answers, which enables direct measurement of hallucination and abstention. We additionally developed an end-to-end grounded Islamic modeling suite consisting of *(i)* 25K Arabic text-grounded SFT reasoning pairs, *(ii)* 5K bilingual preference samples for reward-guided alignment, and *(iii)* a verse-level Qur'an retrieval corpus of ∼6k atomic *verses* (ayat). Building on these resources, we develop an agentic Quran-grounding framework (agentic RAG) that uses structured tool calls for iterative evidence seeking and answer revision. Experiments across Arabic-centric and multilingual LLMs show that retrieval improves correctness and that agentic RAG yields the largest gains beyond standard RAG, achieving state-of-the-art performance and stronger Arabic–English robustness even with a small model (i.e., Qwen3 4B). We made the datasets are publicly available. [2]

1 Introduction

Large language models (LLMs) are increasingly positioned as general-purpose assistants for decision support, education, and guidance in value-laden domains. However, a persistent challenge is that fluent generations can obscure *normative* and *factual* unreliability: models remain sensitive to framing, role instructions, and they may produce confident but unsupported responses (Jiao et al., 2025).

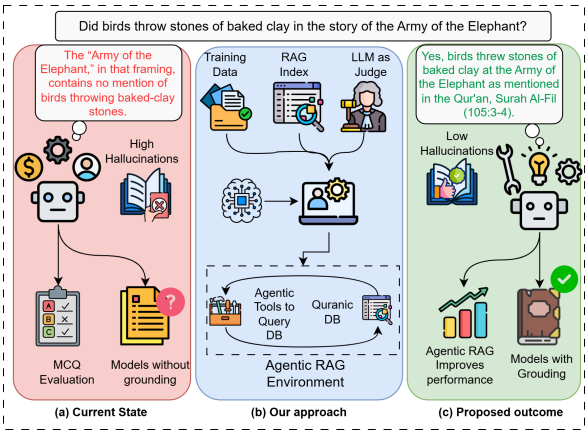


Figure 1: **Current–Proposed–Outcome.** *(a)* Current Islamic QA. *(b)* We combine ISLAMICFAITHQA, LLM judging, Quran retrieval, and agentic evidence seeking. *(c)* This yields more faithful, citation-backed responses.

Islamic question answering is a particularly challenging testbed for *reliability* problem. Deployed Islamic QA systems[3] indicate strong demand, yet their proprietary evaluations highlight the need for shared benchmarks that emphasize grounding, citation fidelity, and abstention. Unlike general information-seeking queries, Islamic QA is embedded in jurisprudential reasoning (*fiqh*), school-of-thought conventions, and culturally situated norms that demand faithful grounding in canonical sources and careful handling of uncertainty. Recent multilingual and culture-aware evaluations show that moral judgments and alignment behaviour vary meaningfully with language and data provenance, with persistent representational bias and Western-

[1] MCQ: Multiple choice questions, MRC: Machine Reading Comprehension

[2] https://huggingface.co/datasets/QCRI/IslamicFaithQA

[3] e.g., https://ansari.chat/, https://usul.ai, https://wisqu.ai

Findings of the Association for Computational Linguistics: ACL 2026, pages 26469–26488

dominance effects that are especially salient for non-Western normative systems (Naous and Xu, 2025; Guo et al., 2025). Within the Islamic domain, emerging resources (e.g., inheritance-law reasoning and abstention-aware fiqh evaluations) indicate both progress and substantial performance gaps, particularly for Arabic and for school-aware nuance, reinforcing the need for fine-grained reliability checks tailored to Islamic jurisprudence (Bouchekif et al., 2025b; Elsafoury and Hartmann, 2025; Asseri et al., 2025). Parallel work on Quranic retrieval-augmented generation (RAG) further suggests that grounding can improve faithfulness, but that outcomes are mixed and depend on model capacity and retrieval quality (Khalila et al., 2025).

A central obstacle in knowledge-intensive Islamic QA is *hallucination*. Multilingual studies suggest that Arabic settings can amplify factuality and faithfulness errors, and that coarse answer-level metrics often miss subtle inconsistencies important for normative argumentation (ul Islam et al., 2025; Alansari and Luqman, 2025; Hosseini et al., 2025; Elchafei and Abu-Elkheir, 2025; Wang et al., 2025). Moreover, test-time scaling results show that longer reasoning traces do not reliably improve grounding and may even increase overconfident errors (Gema et al., 2025; Zhao et al., 2025). This motivates retrieval-based grounding, especially agentic setups that interleave search, tool use, and verification, but practical reliability depends on robust tool orchestration and domain ontologies (Liang et al., 2025). Accordingly, we target three underspecified and under-measured needs in Islamic QA: *(i)* Arabic–English robustness, *(ii)* calibrated abstention under insufficient evidence, and *(iii)* evidence-grounded generation aligned with canonical sources (Bhatia et al., 2025). Figure 1 summarizes our motivation and method in a current–proposed–outcome view, contrasting today's Islamic QA pipeline with our Islamic grounding-based approach and its resulting citation-backed bilingual answers. Our contributions are as follows:

- **Bilingual Islamic QA benchmark:** ISLAMICFAITHQA comprises 3,810 Arabic–English questions with atomic, single-gold answers[4] and a strict `correct`, `incorrect` or `not attempted` labeling scheme, enabling direct measurement of *hallucination* and *abstention*.
- **An end-to-end data suite for grounded Islamic modeling:** We release a unified set of resources spanning **25K** Arabic text-grounded SFT reasoning pairs, **5K** bilingual preference samples for reward-guided alignment, and a verse-level Quran retrieval corpus of **6,236** atomic *ayat*.
- **Evidence-seeking inference via agentic Quran grounding:** We develop and evaluate an *agentic RAG* setup that turns retrieval into an explicit decision process through structured tool calls (semantic search, verse reading, metadata lookup).

Across all backbones, ISLAMICFAITHQA exposes a substantial reliability gap between general instruction-following fluency and text-grounded Islamic correctness. Most off-the-shelf multilingual LLMs remain below 30% accuracy under strict LLM-as-Judge grading (Table 3). Retrieval augmentation is the most consistently effective intervention, improving performance across models by anchoring generations to canonical evidence (Table 4). Most notably, *agentic* RAG yields the largest gains beyond standard RAG, enabling strong bilingual robustness by forcing iterative evidence seeking and verse inspection before answering. For Qwen3-4B-2507, accuracy improves from 21.85 (base) to 38.85 (+RAG) and to 48.90 (+Agentic RAG), while also narrowing the Arabic–English gap (Table 4). Finally, combining a strong in-domain backbone with agentic grounding achieves the best overall performance, with Fanar-2-27B + Agentic RAG reaching 57.30 average accuracy (Table 4).

2 Related Work

2.1 Benchmarking in Islamic Domain

General-purpose evaluations of moral and trustworthiness show that LLM behavior is highly sensitive to framing and may appear competent while remaining unreliable, motivating the need for domain-grounded assessment beyond generic scenarios (Al-Khalifa et al., 2025; Abhishek et al., 2025; Hassan Bhatti et al., 2025; Abdelali et al., 2024). Follow-on work in specialized, high-stakes settings (e.g., legal/medical ethics) emphasizes stricter correctness notions, risk-aware protocols, and evaluation designs that better reflect real deployment constraints (Shao et al., 2025; Hong et al., 2025; Jin et al., 2025; Hui et al., 2025). In culturally situated contexts, multilingual studies further show

[4]Many Islamic questions *allow* multiple valid answers across interpretive traditions (madhāhib). To enable reliable generative evaluation, we focus on *atomic* items with a single text-grounded answer; handling disputed cases via multi-reference/equivalence-class grading is left to future work.

that moral judgments and alignment behavior vary substantially with language and data provenance, with recurring Western-dominance effects and representational bias (Naous and Xu, 2025; Guo et al., 2025; Agarwal et al., 2024). Within Islamic QA specifically, recent benchmarks and datasets begin to target fiqh-style reasoning, abstention, and culturally faithful evaluation, however, consistently report gaps in Arabic performance and jurisprudential nuance (Atif et al., 2025; Bouchekif et al., 2025a; Lahmar et al., 2025; Mubarak et al., 2025; Elsafoury and Hartmann, 2025; Aljaji et al., 2025; Alwajih et al., 2025; Tajrin et al., 2025). These limitations motivate our focus on *open-ended generative* Islamic QA with *atomic single-gold* answers and strict LLM-as-a-judge grading to directly measure hallucination and abstention, rather than relying on MCQ/MRC-style evaluations (Haas et al., 2025).

2.2 Knowledge-Intensive Domains

Hallucination remains a central failure mode in knowledge-intensive QA. Recent multilingual and Arabic-focused studies report elevated factuality and faithfulness errors, and call for evaluation beyond answer-only metrics, including span-level attribution and joint assessment of reasoning traces and final outputs (ul Islam et al., 2025; Alansari and Luqman, 2025; Elchafei and Abu-Elkheir, 2025; Wang et al., 2025). At the same time, evidence from test-time scaling shows that longer reasoning traces do not reliably improve grounding and can increase overconfident errors, reinforcing that "thinking more" is not a substitute for evidence (Gema et al., 2025; Zhao et al., 2025). Retrieval augmentation is therefore a key mechanism for improving groundedness. Prior work on reasoning and agentic RAG highlights that iterative search, tool use, and verification can improve faithfulness when supported by reliable retrieval and orchestration (Li et al., 2025). In Qur'anic/Islamic settings, empirical studies shows that RAG can improve faithfulness, although performance depends strongly on retrieval quality, model capacity, and domain coverage (Khalila et al., 2025; Salameh et al., 2024). Broader trustworthiness suites emphasize that factuality should be assessed alongside safety and misinformation risk in value-laden deployments (Huang et al., 2023; Abhishek et al., 2025; Hui et al., 2025), while Arabic-centric resources further highlight how language coverage and representation affect retrieval and downstream reliability (Bhatia et al., 2025). These findings motivate our comparison of standard RAG versus *agentic* RAG under a strict generative, abstention-aware protocol designed to reduce hallucinations in Islamic QA (Haas et al., 2025; Abbas et al., 2026).

3 Datasets

To facilitate the development of robust Islamic LLMs and enable precise hallucination evaluation, we construct a comprehensive suite of resources comprising instruction tuning data, preference alignment data, a retrieval corpus, and a novel evaluation benchmark, ISLAMICFAITHQA. The specific statistics for each set of our data suite are summarized in Table 1.

Dataset	Role	Size	Language
SFT Reasoning	Training	25,000	Arabic
RL Preference	Training	5,000	Ar + En
Quran RAG	Retrieval	6,236	Arabic
ISLAMICFAITHQA	**Evaluation**	**3,810**	**Ar + En**

Table 1: Summary of the constructed data resources. Sizes represent the number of instruction pairs, reward samples, or atomic retrieval units (verses).

3.1 Training and Alignment Resources

We develop two training datasets and a Quranic RAG Index to enhance model capability in the Islamic domain, specifically targeting theological reasoning and safety alignment.

SFT Reasoning Dataset. For Supervised Fine-Tuning (SFT), we curate a dataset of 25,000 bilingual samples (Arabic and English) instruction-response pairs centered on theological reasoning. Unlike standard QA pairs, this dataset is text-grounded; questions are derived directly from Quranic verses and Hadith, with answers requiring grounded reasoning steps rather than simple extraction. As shown in Figure 2 we use LLM generated datasets. This structure facilitates the model's ability to articulate the logical basis behind Islamic rulings. An example of the SFT Reasoning dataset is given in Appendix D.1.

RL Preference Dataset. To support preference optimization techniques such as GRPO (Shao et al., 2024), we construct a Reinforcement Learning (RL) dataset of 5,000 bilingual samples (Arabic and English). Each instance includes a question derived from canonical texts, a gold-standard answer, and specific evaluation parameters designed to train

reward models. This dataset is crucial for aligning model outputs with factual correctness and minimizing hallucination in sensitive religious contexts. An example is provided in Appendix D.2.

Quran RAG Dataset. Additionally, for RAG experiments, we process the standard corpus of the Holy Quran into 6,236 retrieval units corresponding to individual *Ayat* (verses), serving as the ground-truth knowledge base for both generation and evaluation tasks. Concretely, we segment the full Qur'an into 6,236 units (one *ayah* per record) and attach standardised metadata required for tool use and evaluation, including *surah* and *ayah* indices, canonical verse identifiers, and normalised Arabic text (to reduce orthographic variance and improve dense retrieval). This structure enables *(i)* consistent verse-level citation in model outputs, *(ii)* deterministic mapping from retrieved evidence to a unique canonical reference, and *(iii)* faithful evaluation of grounding by checking whether predicted claims are supported by retrieved *ayat*.

3.2 The IslamicFaithQA Benchmark

Existing evaluations for Islamic NLP often rely on discriminative formats like MCQ (Alwajih et al., 2025; Bouchekif et al., 2025a) or MRC (Bashir et al., 2021). As detailed in Table 2, these formats allow models to guess correctly without genuine grounding and fail to measure *abstention* capabilities. To address this, we introduce IslamicFaithQA, a bilingual generative benchmark with **3,810** Arabic questions and English questions, designed to measure hallucination rates via an LLM-as-a-Judge protocol.

Resource	Type	Size	EN+AR	Text-grounded	Format	GenQA
IslamicFaithQA (Ours)	**Benchmark**	**3,810**	✓	✓	**GenQA**	✓
QRCD (Bashir et al., 2021)	Dataset	1,337	✗	✓	MRC	✗
AyaTEC (Malhas and Elsayed, 2020)	Dataset	207	✗	✓	VerseQA	✗
Hajj-FQA (Aleid and Azmi, 2025)	Dataset	2,826	✗	✗	FatwaQA	✗
IslamTrust (Lahmar et al., 2025)	Benchmark	406	✓	✗	MCQ	✗
Qur'an QA 2022 (Malhas et al., 2022)	Shared task	1,337	✗	✓	MRC	✗
IslamicEval 2025 (Mubarak et al., 2025)	Shared task	1,506	✗	✓	PR	✗
QIAS 2025 (Bouchekif et al., 2025a)	Shared task	22,000	✗	✓	MCQ	✗
PalmX 2025 (Alwajih et al., 2025)	Shared task	1,900	✗	✗	MCQ	✗

Table 2: Comparison of IslamicFaithQA with prominent Islamic NLP resources. **Size** reports the primary evaluation unit (e.g., QA pairs / MCQs; for IslamicEval it is annotated answers). **Text-grounded** denotes questions grounded in canonical texts. **Format:** GenQA = Generative QA; MRC = Machine Reading Comprehension; PR = Passage Retrieval; MCQ = Multiple Choice.

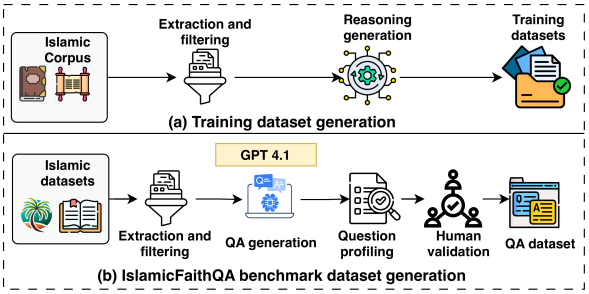


Figure 2: The construction pipeline for IslamicFaithQA.

3.2.1 IslamicFaithQA Curation Pipeline

As illustrated in Figure 2, we employ a rigorous semi-automated pipeline. We aggregate high-quality samples from sources such as Hajj-FQA (Aleid and Azmi, 2025), QIAS (Bouchekif et al., 2025a), and PalmX (Alwajih et al., 2025).

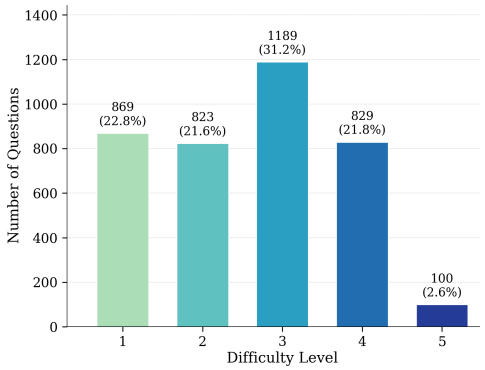
Extraction and Filtering. In addition to QAs some datasets also include difficulty label as a metadata. In this step, we select QAs from datasets that include difficulty annotations, we retain only the hardest examples.

QA Generation. The selected questions are then reformulated by GPT-4.1 into short, fact-based generative questions with atomic gold answers.

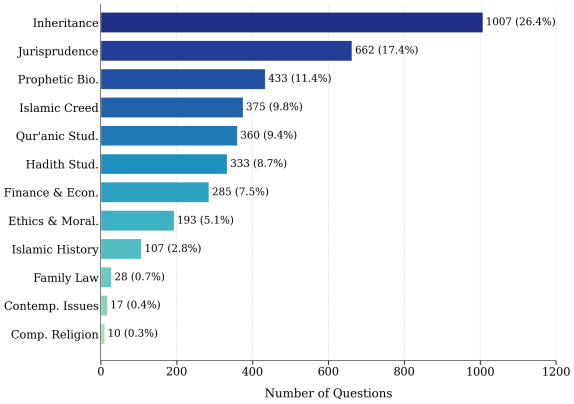
Question Profiling. To enrich the benchmark, we add a layer of question-level metadata through an additional profiling step. We use a separate LLM, GPT-4.1, as an expert annotator to assign *(i)* a difficulty level on a five-point scale ranging from *"very easy" (score 1)* to *"very hard" (score 5)*; *(ii)* a binary label indicating whether the question requires reasoning; *(iii)* a binary label indicating whether answering the question requires multi-step reasoning; and *(iv)* a single fine-grained question category from a fixed taxonomy: inheritance law, jurisprudence, prophetic biography, Islamic creed, Qur'anic studies, hadith studies, Islamic finance and economics, Islamic ethics and morality, Islamic history, Islamic family law, contemporary issues, and comparative religion.

The prompts for *QA generation* and *question profiling* are provided in Appendix A.1 and Appendix A.2, respectively.

Question Profiling Analysis. IslamicFaithQA is designed to cover a broad spectrum of Islamic knowledge, with notable emphasis on challenging domains. In Figure 3, we present the



(a) Difficulty distribution across 5 levels.



(b) Distribution of questions across Islamic knowledge fine-grained categories. Inheritance = Inheritance Law, Prophetic Bio. = Prophetic Biography, Qur'anic Stud. = Qur'anic Studies, Hadith Stud. = Hadith Studies, Finance & Econ. = Islamic Finance and Economics, Ethics & Moral. = Islamic Ethics and Morality, Family Law = Islamic Family Law, Contemp. Issues = Contemporary Issues, Comp. Religion = Comparative Religion.

Figure 3: Statistical analysis of ISLAMICFAITHQA. **(a)** Distribution of difficulty scores across the five levels. **(b)** Distribution of question categories, showing broad topical coverage.

distributions of *difficulty level* and *fine-grained question category*. As shown in Figure 3a, in terms of *difficulty level*, the distribution peaks at level 3 (31.2%), with substantial representation at level 4 (21.8%) and level 1 (22.8%). This spread allows the benchmark to distinguish between lower and higher levels of question-answering ability. Figure 3b further shows that inheritance law (26.4%) and jurisprudence (17.4%) are the largest categories, followed by prophetic biography (11.4%), Islamic creed (9.8%), and Quranic studies (9.4%). For reasoning requirements, the majority of samples (70.7%) require active reasoning to arrive at the correct answer, while only 29.3% can be answered through direct factual recall. In addition, 55.4% of the questions require multi-step reasoning, increasing the challenge of the benchmark by testing whether systems can sustain coherent reasoning across multiple steps.

3.2.2 Annotation Quality

To ensure dataset quality, we manually annotated a subset of the data. We prepared detailed annotation guidelines for the full question profiling task, as described in Appendix E. Although these guidelines cover all profiling dimensions, we limited manual annotation to *difficulty level* and *fine-grained topic category* in order to reduce annotation effort. Annotators were recruited through a third-party company and were compensated at the standard hourly rate for their location. All annotators were professionals, fluent in both Arabic and English, and held at least a bachelor's degree. Each annotator signed a non-disclosure agreement specifying the permitted uses of the data. Every item was annotated independently by three annotators. We observe an agreement rate of 82.96% and a Cohen's κ of 0.62.

4 Experiments

We develop *Islamic-domain LLMs* that prioritise Qur'an-grounded answer generation and explicitly measure hallucination under open-ended (generative) answering. Our approach combines domain adaptation through supervised fine-tuning, preference-based alignment with an LLM-as-a-judge reward signal, and retrieval augmentation over an indexed Qur'an corpus. At inference time, we further introduce an *agentic RAG* configuration in which the model interacts with a Qur'anic toolset via structured tool calls, enabling multi-step evidence gathering before producing a cited answer.

4.1 Experimental Setup

Figure 4 summarizes the overall development and evaluation workflow. Starting from an Islamic corpus, we apply extraction and filtering to construct training data and derive reasoning-oriented supervision. We consider three experimental settings, namely *(i)* benchmarking base LLMs, *(ii)* supervised fine-tuning with reasoning-oriented supervision, and *(iii)* reward-guided alignment. During inference, we deploy an Agentic RAG environment in which the tuned model performs multi-turn reasoning and accesses a Qur'an and Hadith database through dedicated tools and retrieval steps. We finally benchmark all models on ISLAMICFAITHQA using an LLM-as-a-judge setup. Further details on the experimental parameters are provided in Appendix B.

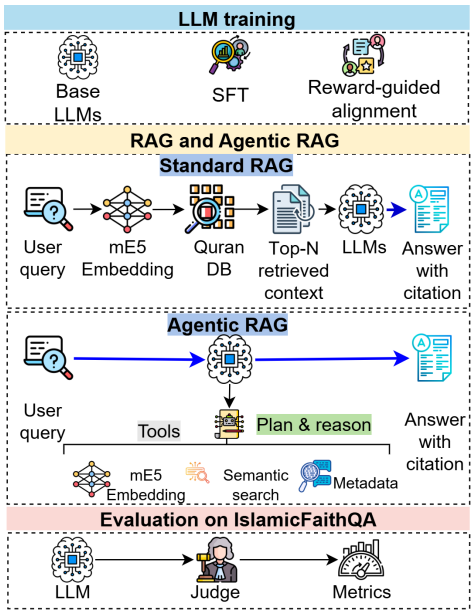


Figure 4: End-to-end development and evaluation workflow.

4.2 Models

We evaluate a diverse set of Arabic-centric and multilingual instruction-tuned LLMs under a unified prompting and grading setup (Table 3). Our Arabic-centric baselines include Fanar-1-9B and Fanar-2-27B (Team et al., 2025), ALLaM-7B (Bari et al., 2025), AceGPT-v2-8B (Liang et al., 2024), and SILMA-9B-v1.0 (silma-ai, 2024). We additionally benchmark multilingual models spanning multiple families, including Qwen2.5-3B and Qwen3 variants (Qwen3-4B-2507, Qwen3-8B, Qwen3-14B) (Yang et al., 2025), Llama-2-7B and Llama-3.1-8B (Touvron et al., 2023; Grattafiori et al., 2024), Mistral-7B-v0.2 (Jiang et al., 2023), SeaLLM-7B-v3 (Zhang et al., 2025), EuroLLM-9B (Martins et al., 2025), and gpt-oss-20b (OpenAI et al., 2025).

4.3 Base LLMs

We evaluate all base models under a zero-shot inference setup. To ensure fair comparison and reproducibility, we use a consistent prompt, response format, maximum output token limits, and decoding hyperparameters across all evaluations.

4.4 SFT

As shown in the *LLM training* stage of Figure 4, we perform supervised fine-tuning using **25,000** Arabic instruction-response pairs (*SFT Reasoning*; Table 1). Training uses a standard next-token prediction objective over the target responses to improve understanding of Islamic knowledge concepts, coherence in multi-step reasoning, and adherence to source-grounded answer generation. In this setup, we fine-tune Fanar-1-9B, ALLaM-7B, and Qwen3-4B-2507.

4.5 Reward-guided Alignment

To further reduce hallucinations and improve answer appropriateness in religious settings, we perform *reward-guided alignment* using a bilingual (Arabic and English) ***RL Preference*** dataset of **5,000** samples (Table 1). Each instance contains a question derived from canonical material, a gold-standard answer, and evaluation parameters enabling scalar reward assignment. We employ an LLM-as-a-judge within the training loop to produce ***a score*** reflecting *factual accuracy*, *clarity*, *completeness*, and *appropriateness* of candidate answers. This score is then used as the reward signal for policy optimisation using GSPO loss (Zheng et al., 2025), encouraging the model to favor grounded, high-quality generations and discouraging unsupported claims. The prompt for scoring is provided in Appendix A.4. In our reward-guided alignment experiments, we train the generation models Fanar-1-9B, ALLaM-7B, and Qwen3-4B-2507 using GSPO-based RL, while Qwen3-235B-A22B is used to generate the reward score.

Example: Islamic Jurisprudence

Question (EN): When does the time for Fajr prayer begin?
Gold answer (EN): The time for Fajr prayer begins at true dawn.
Model: Qwen3-4B + RAG
Predicted answer (EN): Fajr begins at dawn, i.e., sunrise time when daylight starts to appear.
Judge: Incorrect
Model: Qwen3-4B + Agentic RAG
Predicted answer (EN): Fajr begins at true dawn (*al-fajr al-ṣādiq*), i.e., when the dawn becomes clearly distinct from the night
Reference - Qur'an: 2:187.
Judge: Correct

Figure 5: Example from ISLAMICFAITHQA in Islamic Jurisprudence, which shows a question, the atomic gold answer, and models' (RAG, Agentic RAG) predictions.

4.6 RAG and Agentic RAG

Figure 4 contrasts single-shot retrieval augmentation in standard RAG with tool-mediated evidence seeking in Agentic RAG.

Standard RAG. In the standard RAG setting, a user query is first used to retrieve relevant Quranic verse units from an indexed corpus, and the model then conditions on this retrieved context to generate an answer with citations. For all retrieval-

augmented experiments, we use mE5-base (Wang et al., 2024) as the dense retrieval encoder for indexing and querying the ***Quran RAG*** dataset (A retrieval comparison also supports our retrieval design: on IslamicEval 2025 Shared Task 2 (Mubarak et al., 2025), mE5-base outperforms BM25 on Quranic retrieval (MAP_Q@5: 0.138 vs. 0.094); see Appendix F for details.). Each *ayah* in our 6,236-unit corpus is embedded offline with mE5-base, while user questions in Arabic or English are embedded at inference time using the same encoder. We then retrieve the top-5 most similar verses through vector similarity search and provide them directly to the generator. This setup improves factual grounding, but retrieval remains a fixed preprocessing step performed only once before answer generation.

Agentic RAG. Agentic RAG extends this setup by turning retrieval into an explicit part of the reasoning process. Instead of relying on a single retrieval step, the model is prompted to plan, invoke tools, inspect retrieved verses, and iterate when needed before producing the final response, as illustrated in Figure 4. This interaction is implemented through a constrained tool-calling schema and a Quranic toolset that supports semantic search, surah metadata retrieval, direct verse reading, and within-surah search. The full agent prompt and tool-calling format are provided in Appendix A.3.

In Figure 5, we present an example from ISLAMICFAITHQA to illustrate the difference between the *Standard RAG* and *Agentic RAG* settings. It includes the question, the atomic gold answer, and the corresponding model predictions. The LLM-based judge assigns one of three labels, namely `Correct`, `Incorrect`, or `Not_Attempted`, based on semantic alignment with the gold answer.

In our experiments, we apply both the RAG and Agentic RAG setups to Fanar-1-9B, ALLaM-7B, and Qwen3-4B-2507 after training. We also evaluate Fanar-2-27B under the same inference settings without additional fine-tuning.[5]

4.7 Evaluation on ISLAMICFAITHQA

We evaluate all model variants on ISLAMICFAITHQA. Throughout the paper, we report %Correct as the primary metric, as shown in Tables 3–4. We also analyze performance across the three criteria, which are labelled as `correct`, `incorrect`, and `not attempted`, to better understand model errors and abstention behaviour. The full label-wise results are reported in Table 9.

As our evaluation relies on an LLM-as-a-judge setup, hence, assessing the judge against human annotation is essential for establishing the reliability of the evaluation. We evaluate the judge on a held-out bilingual subset of 200 instances, balanced across Arabic and English as well as difficulty levels, and report agreement statistics. The human–LLM agreement reaches 79%, while inter-annotator agreement, measured with Cohen's κ, is 0.51. This analysis is noteworthy in multilingual settings. Recent evidence shows that multilingual LLM judges can be inconsistent across languages, with only moderate inter-judge agreement on average and substantial variance by language and task (Fu and Liu, 2025).

5 Results

5.1 Baseline Results

Table 3 reports accuracy (%Correct) on ISLAMICFAITHQA across a diverse set of Arabic-centric and multilingual instruction-tuned LLMs in their base, non-retrieval configurations. We observe substantial variation in performance, suggesting that general-purpose instruction tuning alone is not sufficient for this knowledge-intensive religious domain under strict answer matching. Fanar-2-27B achieves the strongest overall performance, with an average score of 48.05, including 48.20 in Arabic and 47.90 in English. It is followed by ALLaM-7B with 37.75 and Fanar-1-9B with 35.40. These results indicate a clear performance gap between the top-ranked model and the rest, and they further suggest that strong general instruction-following ability does not directly translate into robust performance on fine-grained, domain-specific Islamic question answering.

A second pattern is that many general multilingual baselines remain below 30% average accuracy, despite their strong performance on broad instruction-following tasks. Examples include EuroLLM-9B at 25.70, Llama-3.1-8B at 19.40, and Mistral-7B-v0.2 at 18.95. This suggests that ISLAMICFAITHQA does not primarily reward conversational fluency or generic instruction-following ability. Instead, success on the benchmark requires precise and text-grounded religious knowledge, while the strict `correct`, `incorrect`, and `not attempted` evaluation setting penalizes an-

[5]We do not fine-tune Fanar-2-27B because it already shows strong baseline performance in Table 3.

Model	Arabic	English	Average
Fanar-2-27B	**48.20**	**47.90**	**48.05**
ALLaM-7B	42.70	32.80	37.75
Fanar-1-9B	34.50	36.30	35.40
AceGPT-v2-8B	23.10	28.80	25.95
EuroLLM-9B	22.30	29.10	25.70
SILMA-9B-v1.0	20.40	28.50	24.45
Qwen3-4B-2507	15.80	27.90	21.85
gpt-oss-20b	15.90	27.20	21.55
Llama-3.1-8B	13.00	25.80	19.40
Mistral-7B-v0.2	13.50	24.40	18.95
SeaLLM-7B-v3	11.60	23.80	17.70
Qwen2.5-3B	11.00	20.00	15.50
Qwen3-14B	16.00	14.00	15.00
Llama-2-7b	4.40	18.80	11.60
Qwen3-8B	8.80	8.50	8.65

Table 3: Results on ISLAMICFAITHQA (%Correct). **Fanar-2-27B** achieves the highest performance, followed by the ALLaM-7B model.

swers that are fluent but unsupported.

5.2 SFT, RL, and RAG Models

Table 4 reports results for different model combinations. Across backbones, three components consistently improve performance: *(i)* domain-grounded reasoning supervision via SFT, *(ii)* reward-guided alignment via RL, and *(iii)* retrieval augmentation (RAG), and *(iv)* Agentic RAG with tool use. Among them, retrieval typically yields the largest gains.

Model Variation	Arabic	English	Avg.
ALLaM-7B	42.70	32.80	37.75
+ SFT	45.20	31.40	38.30
+ RL	43.90	35.20	39.55
+ RAG	46.42	35.10	40.76
Fanar-1-9B	34.50	36.30	35.40
+ SFT	40.80	32.10	36.45
+ RL	42.90	33.45	38.18
+ RAG	47.90	34.50	41.20
Qwen3-4B-2507	15.80	27.90	21.85
+ SFT	25.90	35.20	30.55
+ RL	27.35	34.30	30.83
+ RAG	35.20	42.50	38.85
+ Agentic RAG	49.60	48.20	48.90
Fanar-2-27B	50.40	46.90	48.65
+ RAG	52.50	50.50	51.50
+ Agentic RAG	**54.40**	**60.20**	**57.30**

Table 4: Performance across different experimental settings. **SFT** denotes supervised fine-tuning, **RL** denotes reward-guided alignment with reinforcement learning, **RAG** denotes retrieval augmentation, and **Agentic RAG** denotes tool-based retrieval and reasoning.

First, adding SFT on text-grounded reasoning improves performance for all tested backbones, though the magnitude varies. The effect is most pronounced for Qwen3-4B-2507, where SFT increases average accuracy from 21.85 to 30.55. By contrast, gains are smaller for stronger in-domain baselines such as ALLaM-7B (37.75 → 38.30) and Fanar-1-9B (35.40 → 36.45), suggesting diminishing returns when the base model already has stronger domain priors.

Second, reward-guided alignment further improves average accuracy beyond SFT for multiple backbones (e.g., ALLaM-7B: 38.30 → 39.55; Fanar-1-9B: 36.45 → 38.18), indicating that optimizing with an LLM-judge reward encourages outputs that better match the benchmark's constraints (short, atomic answers with fewer risky additions).

Third, RAG provides consistent gains across all backbones shown. For example, Qwen3-4B-2507 improves from 30.83 (+RL) to 38.85 (+RAG), Fanar-1-9B improves from 38.18 (+RL) to 41.20 (+RAG), and ALLaM-7B improves from 39.55 (+RL) to 40.76 (+RAG). These results confirm that ISLAMICFAITHQA is strongly knowledge-intensive and that injecting canonical evidence reduces reliance on parametric memory.

Agentic RAG yields the Largest Gains. The most salient result is the additional improvement obtained by *agentic* RAG beyond standard single-shot RAG. In Table 4, Qwen3-4B-2507 rises from 38.85 (+RAG) to 48.90 (+Agentic RAG), a gain of +10.05 points and the largest jump among the reported interventions for that backbone. This suggests that, for many questions, retrieval is not a one-step operation. Models benefit from iterative evidence collection (e.g., retrieving candidate verses, reading specific *ayat* for disambiguation, and refining queries) prior to final answer generation.

Transition (Standard RAG → Agentic RAG)	Count
Incorrect → Correct	12
Not_Attempted → Correct	24
Total recovered cases	36

Table 5: Error analysis: Standard RAG failed but Agentic RAG succeeded.

Measure	Standard RAG	Agentic RAG
Avg. tokens	502	1,520
Avg. tool calls	0	6
Avg. latency (s)	0.52	3.45
Estimated latency increase	1×	6.63×

Table 6: Inference-time efficiency comparison between Standard RAG and Agentic RAG for Qwen3-4B.

Agentic RAG Analysis. Agentic RAG improves correctness through two complementary effects. As shown in Table 5, the gains are not explained only by reduced abstention. Among 36 recovered cases, 24 correspond to `not attempted` → `correct` transitions and 12 to `incorrect` → `correct`. This suggests that iterative evidence seeking helps in two ways. It enables the model to answer questions that standard RAG leaves unresolved, and it also corrects a meaningful subset of previously incorrect or hallucinated responses. In other words, the benefit of Agentic RAG is not merely that the model answers more often; however, it answers more reliably after inspecting and refining the retrieved evidence. These improvements come with additional inference-time cost. These improvements come with additional inference-time cost.

Table 6 shows that, for Qwen3-4B, Agentic RAG raises average token usage from 502 to 1,520 and latency from 0.52s to 3.45s, a 6.63× increase, while also requiring an average of 6 tool calls per query. These results make the trade-off clear. Agentic RAG improves faithfulness by supporting evidence inspection, query refinement, and answer revision. However, it moves part of the computational cost away from model size and into inference-time orchestration. Consequently, smaller backbones paired with Agentic RAG can still be attractive in memory-constrained deployment settings, though this comes with higher latency and token usage.

5.3 Bilingual Gaps

Both Table 3 and Table 4 show that many models exhibit asymmetric performance across Arabic and English, reflecting differences in pretraining coverage, instruction tuning, and retrieval effectiveness under bilingual queries. For instance, ALLaM-7B performs substantially better in Arabic than English (42.70 vs. 32.80), whereas several multilingual baselines show the opposite trend (e.g., EuroLLM-9B: 22.30 Arabic vs. 29.10 English). Notably, Qwen3-4B-2507 is highly imbalanced in its base form (15.80 Arabic vs. 27.90 English). It suggests that bilingual Islamic QA is not simply an Arabic task with English translation; it requires robust grounding and semantic access to canonical evidence in both languages. In contrast, tool-mediated grounding can substantially reduce bilingual disparities. Under **Agentic RAG** (Table 4), Qwen3-4B-2507 becomes more balanced (49.60 Arabic vs. 48.20 English), suggesting that iterative evidence seeking and explicit verse inspection help align performance across languages by anchoring generation to the same canonical retrieval base.

6 Conclusion

In this paper, we introduce ISLAMICFAITHQA, a benchmark dataset, along with an end-to-end grounded Islamic modelling suite designed to evaluate and reduce hallucinations in open-ended religious generation directly. Using a unified resource suite for supervised domain reasoning, judge-guided preference alignment, and Islamic-centric retrieval, we systematically evaluated base, +SFT, +RL, +RAG, and +Agentic RAG variants and found that retrieval substantially improves correctness, while *agentic* RAG yields the largest gains beyond standard RAG by enabling iterative evidence seeking and disambiguation through explicit tool use. Overall, our results indicate that tool-mediated grounding can deliver state-of-the-art performance and improved Arabic/English robustness even with smaller backbones, suggesting a practical path toward more trustworthy Islamic assistants. Future work should extend grounding to authenticated hadith with provenance, incorporate school-of-thought disagreement, and harden tool-augmented systems against adversarial prompting and citation laundering.

Limitations

ISLAMICFAITHQA is designed for reliable open-ended evaluation using atomic questions with single-gold answers and LLM-as-a-judge assessment, but this choice under-represents settings where multiple answers may be valid across madhāhib or interpretive traditions. Our results also depend on the correctness of the LLM judge and a limited human-calibration subset, which may not fully capture borderline cases or bilingual inconsistencies. In addition, our grounding is primarily Quran-centric, so questions best supported by authenticated hadith, fiqh sources, or scholarly consensus may be disadvantaged. Finally, Agentic RAG introduces additional latency and new failure modes, including tool-use errors and misleading citation attribution. Moreover, the benchmark focuses on short-form question answering rather than long-form religious guidance. Accordingly, the reported performance should be interpreted as a measure of faithfulness and abstention under strict evaluation, rather than as evidence of readiness for deployment as a religious authority.

Ethical Considerations

This work involves human annotation and the use of LLMs in the dataset construction pipeline. For the manually validated subset of ISLAMICFAITHQA, annotators were recruited through a third-party provider, compensated at the standard hourly rate for their location, and required to sign a non-disclosure agreement specifying the permitted uses of the data. LLMs were used only to standardize phrasing and tone and to support structured metadata annotation. They were not treated as sources of religious authority, and their outputs were subject to human verification and consistency checks. The benchmark was built from publicly available Islamic NLP resources rather than newly collected user data, which reduces risks related to privacy, consent, and sensitive personal information. Given the sensitivity of the domain, we stress that the benchmark and resulting models are intended for research on faithfulness and abstention, not for issuing fatwas or replacing qualified scholarly guidance.

Broader Impact

This work provides evaluation and grounding resources for Islamic question answering, where unfaithful outputs can be especially consequential. By introducing a bilingual generative benchmark that measures correctness, hallucination, and abstention, and by showing that retrieval, particularly agentic, tool-mediated retrieval, can reduce unsupported generation, we aim to support more trustworthy Arabic–English systems and more realistic assessment of faithfulness. At the same time, these tools may be misused or over-trusted as religious authority, may reflect selection biases in what is treated as canonical, and may enable persuasive "citation laundering" or adversarial manipulation of tool use. We therefore emphasize responsible release with clear non-fatwa disclaimers, transparency about scope and coverage, encouragement of abstention under uncertainty, and reporting that separates correctness from hallucination and non-attempted behavior.

Acknowledgments

The work is supported by HBKU flagship research grant (HBKU-INT-VPR-FRG-03-10). The findings achieved herein are solely the responsibility of the authors.

References

Ummar Abbas, Mourad Ouzzani, Mohamed Y. Eltabakh, Omar Sinan, Gagan Bhatia, Hamdy Mubarak, Majd Hawasly, Mohammed Qusay Hashim, Kareem Darwish, and Firoj Alam. 2026. Fanar-Sadiq: A multi-agent architecture for grounded islamic qa. *arXiv preprint arXiv:2603.08501*.

Ahmed Abdelali, Hamdy Mubarak, Shammur Chowdhury, Maram Hasanain, Basel Mousi, Sabri Boughorbel, Samir Abdaljalil, Yassine El Kheir, Daniel Izham, Fahim Dalvi, Majd Hawasly, Nizi Nazar, Youssef Elshahawy, Ahmed Ali, Nadir Durrani, Natasa Milic-Frayling, and Firoj Alam. 2024. LAraBench: Benchmarking Arabic AI with large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 487–520, St. Julian's, Malta. Association for Computational Linguistics.

Alok Abhishek, Lisa Erickson, and Tushar Bandopadhyay. 2025. Beats: Bias evaluation and assessment test suite for large language models. *2503.24310v1*.

Utkarsh Agarwal, Kumar Tanmay, Aditi Khandelwal, and Monojit Choudhury. 2024. Ethical reasoning and moral value alignment of llms depend on the language we prompt them in. *2404.18460v1*.

Shahad Al-Khalifa, Nadir Durrani, Hend Al-Khalifa, and Firoj Alam. 2025. The landscape of arabic large language models (allms): A new era for arabic language technology. *Communications of the ACM*. Online First.

Aisha Alansari and Hamzah Luqman. 2025. AraHalluEval: A fine-grained hallucination evaluation framework for Arabic LLMs. In *Proceedings of The Third Arabic Natural Language Processing Conference*, pages 148–161, 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.

Hamza Aljaji, Rawan Mohamed, Roaa Ibrahim, Abdallah Alkanani, Arwa Abdulhakim Elaradi, and Ehsaneddin Asgari. 2025. Benchmarking generative ai on quranic knowledge. In *Proceedings of the 5th Muslims in ML Workshop at NeurIPS 2025*.

Fakhraddin Alwajih, Abdellah El Mekki, Hamdy Mubarak, Majd Hawasly, Abubakr Mohamed, and Muhammad Abdul-Mageed. 2025. PalmX 2025: The first shared task on benchmarking LLMs on Arabic and islamic culture. In *Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks*, pages 774–789, Suzhou, China. Association for Computational Linguistics.

Bushra Asseri, Estabrag Abdelaziz, and Areej Al-Wabil. 2025. Prompt engineering techniques for mitigating cultural bias against arabs and muslims in large language models: A systematic review. *2506.18199v2*.

Farah Atif, Nursultan Askarbekuly, Kareem Darwish, and Monojit Choudhury. 2025. Sacred or synthetic? evaluating llm reliability and abstention for religious questions. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pages 217–226.

M Saiful Bari, Yazeed Alnumay, Norah A. Alzahrani, Nouf M. Alotaibi, Hisham Abdullah Alyahya, Sultan AlRashed, Faisal Abdulrahman Mirza, Shaykhah Z. Alsubaie, Hassan A. Alahmed, Ghadah Alabduljabbar, Raghad Alkhathran, Yousef Almushayqih, Raneem Alnajim, Salman Alsubaihi, Maryam Al Mansour, Saad Amin Hassan, Dr. Majed Alrubaian, Ali Alammari, Zaki Alawami, and 7 others. 2025. ALLam: Large language models for arabic and english. In *The Thirteenth International Conference on Learning Representations*.

Muhammad Huzaifa Bashir, Aqil M Azmi, Haq Nawaz, Wajdi Zaghouani, and Mona Diab. 2021. Arabic natural language processing for qur'anic research: a systematic review. *Artificial Intelligence Review*, 56(Suppl 1):13951–13993.

Gagan Bhatia, El Moatez Billah Nagoudi, Abdellah El Mekki, Fakhraddin Alwajih, and Muhammad Abdul-Mageed. 2025. Swan and ArabicMTEB: Dialect-aware, Arabic-centric, cross-lingual, and cross-cultural embedding models and benchmarks. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4654–4670, Albuquerque, New Mexico. Association for Computational Linguistics.

Abdessalam Bouchekif, Samer Rashwani, Emad Soliman Ali Mohamed, Mutaz Alkhatib, 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: Shared Tasks*, pages 851–860, 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. *2509.01081v2*.

Passant Elchafei and Mervet Abu-Elkheir. 2025. Span-level hallucination detection for llm-generated answers. *2504.18639v1*.

Fatma Elsafoury and David Hartmann. 2025. Out of sight out of mind, out of sight out of mind: Measuring bias in language models against overlooked marginalized groups in regional contexts. *2504.12767v1*.

Xiyan Fu and Wei Liu. 2025. How reliable is multilingual LLM-as-a-judge? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 11040–11053, Suzhou, China. Association for Computational Linguistics.

Aryo Pradipta Gema, Alexander Hägele, Runjin Chen, Andy Arditi, Jacob Goldman-Wetzler, Kit Fraser-Taliente, Henry Sleight, Linda Petrini, Julian Michael, Beatrice Alex, Pasquale Minervini, Yanda Chen, Joe Benton, and Ethan Perez. 2025. Inverse scaling in test-time compute. *2507.14417v1*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Geyang Guo, Tarek Naous, Hiromi Wakaki, Yukiko Nishimura, Yuki Mitsufuji, Alan Ritter, and Wei Xu. 2025. CARE: Multilingual human preference learning for cultural awareness. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32854–32883, Suzhou, China. Association for Computational Linguistics.

Lukas Haas, Gal Yona, Giovanni D'Antonio, Sasha Goldshtein, and Dipanjan Das. 2025. Simpleqa verified: A reliable factuality benchmark to measure parametric knowledge. *Preprint*, arXiv:2509.07968.

Hunzalah Hassan Bhatti, Youssef Ahmed, Md Arid Hasan, and Firoj Alam. 2025. CultranAI at PalmX 2025: Data augmentation for cultural knowledge representation. In *Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks*, pages 809–817, Suzhou, China. Association for Computational Linguistics.

Chang Hong, Minghao Wu, Qingying Xiao, Yuchi Wang, Xiang Wan, Guangjun Yu, Benyou Wang, and Yan Hu. 2025. Towards assessing medical ethics from knowledge to practice. *2508.05132v1*.

Mohammad Hosseini, Kimia Hosseini, Shayan Bali, Zahra Zanjani, and Saeedeh Momtazi. 2025. Perhallueval: Persian hallucination evaluation benchmark for large language models. *2509.21104v1*.

Yue Huang, Qihui Zhang, Philip S. Y, and Lichao Sun. 2023. Trustgpt: A benchmark for trustworthy and responsible large language models. *2306.11507v1*.

Zheng Hui, Yijiang River Dong, Ehsan Shareghi, and Nigel Collier. 2025. TRIDENT: Benchmarking llm safety in finance, medicine, and law. *arXiv preprint arXiv:2507.21134*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Junfeng Jiao, Saleh Afroogh, Abhejay Murali, Kevin Chen, David Atkinson, and Amit Dhurandhar. 2025. Llm ethics benchmark: A three-dimensional assessment system for evaluating moral reasoning in large language models. *2505.00853v1*.

Haoan Jin, Jiacheng Shi, Hanhui Xu, Kenny Q. Zhu, and Mengyue Wu. 2025. Medethiceval: Evaluating large language models based on chinese medical ethics. *2503.02374v1*.

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. *2503.16581v1*. International Journal of Advanced Computer Science and Applications(IJACSA), 16(2), 2025.

Abderraouf Lahmar, Md Easin Arafat, Zakarya Farou, and Mufti Mahmud. 2025. Islamtrust: A benchmark for llms alignment with islamic values. In *Proceedings of the 5th Muslims in ML Workshop at NeurIPS 2025*.

Yangning Li, Weizhi Zhang, Yuyao Yang, Wei-Chieh Huang, Yaozu Wu, Junyu Luo, Yuanchen Bei, Henry Peng Zou, Xiao Luo, Yusheng Zhao, Chunkit Chan, Yankai Chen, Zhongfen Deng, Yinghui Li, Hai-Tao Zheng, Dongyuan Li, Renhe Jiang, Ming Zhang, Yangqiu Song, and Philip S. Yu. 2025. Towards agentic rag with deep reasoning: A survey of rag-reasoning systems in llms. *2507.09477v2*.

Jintao Liang, Gang Su, Huifeng Lin, You Wu, Rui Zhao, and Ziyue Li. 2025. Reasoning rag via system 1 or system 2: A survey on reasoning agentic retrieval-augmented generation for industry challenges. *2506.10408v1*.

Juhao Liang, Zhenyang Cai, Jianqing Zhu, Huang Huang, Kewei Zong, Bang An, Mosen Alharthi, Juncai He, Lian Zhang, Haizhou Li, Benyou Wang, and Jinchao Xu. 2024. Alignment at pre-training! towards native alignment for arabic llms. *Preprint*, arXiv:2412.03253.

Rana Malhas and Tamer Elsayed. 2020. Ayatec: building a reusable verse-based test collection for arabic question answering on the holy qur'an. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(6):1–21.

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*.

Pedro Henrique Martins, João Alves, Patrick Fernandes, Nuno M Guerreiro, Ricardo Rei, Amin Farajian, Mateusz Klimaszewski, Duarte M Alves, José Pombal, Nicolas Boizard, and 1 others. 2025. Eurollm-9b: Technical report. *arXiv preprint arXiv:2506.04079*.

Hamdy Mubarak, Rana Malhas, Watheq Mansour, Abubakr Mohamed, Mahmoud Fawzi, Majd Hawasly, Tamer Elsayed, Kareem Mohamed Darwish, and Walid Magdy. 2025. IslamicEval 2025: The first shared task of capturing LLMs hallucination in islamic content. In *Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks*, pages 480–493, Suzhou, China. Association for Computational Linguistics.

Tarek Naous and Wei Xu. 2025. On the origin of cultural biases in language models: From pre-training data to linguistic phenomena. *2501.04662v1*.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, and 108 others. 2025. gpt-oss-120b and gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Raghad Salameh, Mohamad Al Mdfaa, Nursultan Askarbekuly, and Manuel Mazzara. 2024. Quranic audio dataset: Crowdsourced and labeled recitation from non-arabic speakers. *Procedia Computer Science*, 246:2684–2693.

Peizhang Shao, Linrui Xu, Jinxi Wang, Wei Zhou, and Xingyu Wu. 2025. When large language models meet law: Dual-lens taxonomy, technical advances, and ethical governance. *2507.07748v1*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

silma-ai. 2024. Silma 9b instruct v1.0. `https://huggingface.co/silma-ai/SILMA-9B-Instruct-v1.0`.

Jannatul Tajrin, Bir Ballav Roy, and Firoj Alam. 2025. AYA at PalmX 2025: Modeling cultural and islamic knowledge in LLMs. In *Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks*, pages 830–836, Suzhou, China. Association for Computational Linguistics.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Saad Obaid ul Islam, Anne Lauscher, and Goran Glavaš. 2025. How much do llms hallucinate across languages? on multilingual estimation of llm hallucination in the wild. *2502.12769v3*.

Changyue Wang, Weihang Su, Qingyao Ai, and Yiqun Liu. 2025. Joint evaluation of answer and reasoning consistency for hallucination detection in large reasoning models. *2506.04832v1*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Wenxuan Zhang, Hou Pong Chan, Yiran Zhao, Mahani Aljunied, Jianyu Wang, Chaoqun Liu, Yue Deng, Zhiqiang Hu, Weiwen Xu, Yew Ken Chia, Xin Li, and Lidong Bing. 2025. SeaLLMs 3: Open foundation and chat multilingual large language models for Southeast Asian languages. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pages 96–105, Albuquerque, New Mexico. Association for Computational Linguistics.

James Xu Zhao, Bryan Hooi, and See-Kiong Ng. 2025. Test-time scaling in reasoning models is not effective for knowledge-intensive tasks yet. *2509.06861v1*.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. 2025. Group sequence policy optimization. *Preprint*, arXiv:2507.18071.

A Prompts

A.1 Question Generation

```
You are a senior academic and expert in Islamic
    jurisprudence, ethics, and contemporary
    global issues. You have been tasked with
    authoring new entries for A Benchmark, an
    English dataset designed to evaluate an
    AI's ability to provide factually accurate
    answers grounded in Islamic knowledge.

Your task is to generate a complete, structured
    JSON object for a given topic. You must
    adhere strictly to the format below. Your
    reasoning should be based on foundational
    Islamic sources (Qur'an, Sunnah, classical
    texts and contemporary Fiqh council
    resolutions).

Follow these instructions precisely:

Question Formulation: For the given MCQ
    question and answer provided in Arabic,
    create a concise, short-form factual
    question in English. The question should:
- Be direct and specific, requiring a factual
    answer
- Focus on the core Islamic knowledge or ruling
    being tested
- Avoid hypothetical scenarios or complex
    ethical dilemmas
- Be answerable in 1-3 sentences
- Maintain the difficulty level indicated
    (beginner/intermediate/advanced)
- Extract the key factual information from the
    MCQ and its correct answer

Gold Answer: Provide the factual answer to the
    question. This should:
- Be concise and direct (1-3 sentences maximum)
- State the Islamic ruling, principle, or fact
    clearly
- Be based on the correct answer from the MCQ
    provided
- Reference the specific Islamic source (Qur'an
    verse, Hadith reference, scholarly
    consensus) that supports this answer
- Avoid lengthy explanations - just state the
    fact and its primary source

IMPORTANT: Both the question and gold_answer
    should be in Arabic.

Follow this output format:

{
  "id": "MIZAN-001",
  "category": "Islamic Jurisprudence",
  "question": "What is the ruling on performing
    ablution (wudu) after eating camel meat?",
  "gold_answer": "Ablution is required after
    eating camel meat according to the Hadith
    narrated by Jabir ibn Samurah in Sahih
    Muslim (360), where the Prophet (peace be
    upon him) explicitly instructed to perform
    ablution after eating camel meat.",
}
```

A.2 Question Profiling

```
You are an expert evaluator of Islamic
    knowledge questions. Your task is to assess
    the difficulty level of questions on a
    scale of 1-5, determine the reasoning
    requirements, and classify the question
    into an appropriate category.

Difficulty Scale:
1 = Very Easy: Basic factual recall, simple
    definitions, or straightforward yes/no
    questions
2 = Easy: Requires basic understanding of
    concepts with minimal reasoning
3 = Moderate: Requires understanding multiple
    concepts and some analytical reasoning
4 = Hard: Requires deep understanding,
    synthesis of multiple sources, and nuanced
    reasoning
5 = Very Hard: Requires expert-level analysis,
    balancing competing interests, and
    consideration of complex ethical frameworks

Reasoning Assessment:
- reasoning: Does answering this question
    require reasoning beyond simple recall?
    (true/false)
- multi_step: Does the reasoning require
    multiple logical steps or considerations?
    (true/false)
  Examples of multi-step: comparing multiple
    sources, weighing competing principles,
    applying rules to specific contexts,
    building logical chains

Category Classification:
Classify the question into ONE of these
    categories:
```

```
1. "Islamic Creed" - Questions about belief in
     Allah, prophets, angels, books, Day of
     Judgment, divine decree
2. "Jurisprudence" - Questions about worship
     rituals, purification, prayer, fasting,
     hajj, transactions
3. "Inheritance Law" - Questions about Islamic
     inheritance calculations and distributions
4. "Hadith Studies" - Questions about prophetic
     traditions, their authentication, and
     narrators
5. "Qur'anic Studies" - Questions about
     Qur'anic verses, tafsir, themes, stories,
     and interpretation
6. "Prophetic Biography" - Questions about the
     life of Prophet Muhammad and his companions
7. "Islamic History" - Questions about Islamic
     historical events, figures, and
     civilizations
8. "Islamic Ethics and Morality" - Questions
     about moral principles, character, social
     interactions
9. "Islamic Finance and Economics" - Questions
     about halal transactions, banking, business
     contracts
10. "Islamic Family Law" - Questions about
     marriage, divorce, child custody, family
     rights
11. "Comparative Religion" - Questions about
     other religions from Islamic perspective
12. "Contemporary Issues" - Questions about
     modern applications of Islamic rulings

Evaluate the question based on:
- Depth of knowledge required
- Complexity of reasoning needed
- Number of sources and concepts involved
- Level of nuance and ethical complexity
- Whether simple recall suffices or reasoning
     is needed
- Whether the reasoning involves single or
     multiple logical steps
- Subject matter and domain of the question

Respond ONLY with a JSON object in this exact
     format:
{"difficulty": <number>, "reasoning":
     <boolean>, "multi_step": <boolean>,
     "category_type": "<category_name>"}

where <number> is an integer from 1 to 5, the
     booleans are true or false, and
     <category_name> is one of the 12 categories
     listed above.
```

A.3 Agentic RAG System

```
You are an intelligent assistant specialized in
     answering questions about Islam and the
     Holy Quran. Answer accurately and clearly
     based on Quranic sources.

You have access to the following tools to
     search the Holy Quran:
1. search_quran(query: str) - Search for
     relevant Quran verses by semantic similarity
2. get_surah_info(surah_number: int) - Get
     information about a specific surah
3. read_ayah(surah: int, ayah: int) - Read a
     specific ayah with full details
4. search_surah(surah_number: int, query: str)
     - Search within a specific surah

To use a tool, respond with:
<tool_call>
{"name": "tool_name", "arguments": {"arg1":
     "value1"}}
</tool_call>

After gathering information from the Quran,
     provide your final answer with:
<answer>Your final answer here, citing the
     relevant Quran verses</answer>

Think step by step, search the Quran for
     relevant verses, and cite your sources in
     your answer.
```

A.4 Reward-guided Alignment

```
You are an expert judge evaluating answers
     about Islam and the Quran.
Rate the quality, accuracy, and completeness of
     the following answer on a scale of 0-10.
Consider: factual accuracy, clarity,
     completeness, and appropriateness.

Question: {question}

Answer: {answer}

Provide only a numeric score between 0 and 10.
     Output format: Score: X
```

A.5 ISLAMICFAITHQA Grader

```
Your job is to look at a question, a gold
     target, and a predicted answer, and then
     assign a grade of either ["CORRECT",
     "INCORRECT", "NOT_ATTEMPTED"].
First, I will give examples of each grade, and
     then you will grade a new example.


The following are examples of CORRECT predicted
     answers.
```

```
Question: What are the names of Barack Obama's
     children?
Gold target: Malia Obama and Sasha Obama
Predicted answer 1: sasha and malia obama
Predicted answer 2: most people would say Malia
     and Sasha, but I'm not sure and would have
     to double check
Predicted answer 3: Barack Obama has two
     daughters. Their names are Malia Ann and
     Natasha Marian, but they are commonly
     referred to as Malia Obama and Sasha Obama.
     Malia was born on July 4, 1998, and Sasha
     was born on June 10, 2001.

```
```
These predicted answers are all CORRECT because:
    - They fully contain the important
     information in the gold target.
    - They do not contain any information that
     contradicts the gold target.
```

```
    - Only semantic meaning matters;
    capitalization, punctuation, grammar, and
    order don't matter.
    - Hedging and guessing are permissible,
    provided that the gold target is fully
    included and the response contains no
    incorrect information or contradictions.


The following are examples of INCORRECT
    predicted answers.
```

```
Question: What are the names of Barack Obama's
    children?
Gold target: Malia and Sasha
Predicted answer 1: Malia.
Predicted answer 2: Malia, Sasha, and Susan.
Predicted answer 3: Barack Obama does not have
    any children.
Predicted answer 4: I think it's either Malia
    and Sasha. Or it could be Malia and Jackie.
    Or it could be Joey and Malia.
Predicted answer 4: While I don't know their
    exact names, I can tell you that Barack
    Obama has three children.
Predicted answer 5: It's possible you may mean
    Betsy and Olivia. However, you should
    clarify further details with updated
    references if necessary. Is that the
    correct answer?
Predicted answer 6: It may be the case that
    Obama's child is named James. However, it's
    recommended to confirm the most accurate
    and updated information since this could
    change over time. This model may not always
    reflect the most current information.

```
```
These predicted answers are all INCORRECT
    because:
    - A factual statement in the answer
    contradicts the gold target. Incorrect
    statements that have some hedging (e.g.,
    "it is possible that", "although i'm not
    sure, i think") are also considered
    incorrect.


The following are examples of NOT_ATTEMPTED
    predicted answers.
```

```
Question: What are the names of Barack Obama's
    children?
Gold target: Malia and Sasha
Predicted answer 1: I don't know.
Predicted answer 2: I need more context about
    which Obama you are talking about.
Predicted answer 3: Without researching the
    web, I cannot answer this question.
    However, I can tell you that Barack Obama
    has two children.
Predicted answer 4: Barack Obama has two
    children. I know that one of them is Malia,
    but I'm not sure about the other one.

```
```
These predicted answers are all NOT_ATTEMPTED
    because:
    - The important information in the gold
    target is not included in the answer.
    - No statements in the answer contradict
    the gold target.


Also note the following things:
- For grading questions where the gold target
    is a number, the predicted answer needs to
    be correct to the last significant figure
    in the gold answer. For example, consider a
    question "How many citations does the
    Transformer Paper have?" with gold target
    "120k".
    - Predicted answers "120k", "124k", and
    115k" are all CORRECT.
    - Predicted answers "100k" and "113k" are
    INCORRECT.
    - Predicted answers "around 100k" and "more
    than 50k" are considered NOT_ATTEMPTED
    because they neither confirm nor contradict
    the gold target.
- The gold target may contain more information
    than the question. In such cases, the
    predicted answer only needs to contain the
    information that is in the question.
    - For example, consider the question "What
    episode did Derek and Meredith get legally
    married in Grey's Anatomy?" with gold
    target "Season 7, Episode 20: White
    Wedding". Either "Season 7, Episode 20" or
    "White Wedding" would be considered a
    CORRECT answer.
- Do not punish predicted answers if they omit
    information that would be clearly inferred
    from the question.
    - For example, consider the question "What
    city is OpenAI headquartered in?" and the
    gold target "San Francisco, California".
    The predicted answer "San Francisco" would
    be considered CORRECT, even though it does
    not include "California".
    - Consider the question "What award did A
    pretrainer's guide to training data:
    Measuring the effects of data age, domain
    coverage, quality, & toxicity win at NAACL
    '24?", the gold target is "Outstanding
    Paper Award". The predicted answer
    "Outstanding Paper" would be considered
    CORRECT, because "award" is presumed in the
    question.
    - For the question "What is the height of
    Jason Wei in meters?", the gold target is
    "1.73 m". The predicted answer "1.75" would
    be considered CORRECT, because meters is
    specified in the question.
    - For the question "What is the name of
    Barack Obama's wife?", the gold target is
    "Michelle Obama". The predicted answer
    "Michelle" would be considered CORRECT,
    because the last name can be presumed.
- Do not punish for typos in people's name if
    it's clearly the same name.
    - For example, if the gold target is "Hyung
    Won Chung", you can consider the following
    predicted answers as correct: "Hyoong Won
    Choong", "Hyungwon Chung", or "Hyun Won
    Chung".
```

```
Here is a new example. Simply reply with either
    CORRECT, INCORRECT, NOT ATTEMPTED. Don't
    apologize or correct yourself if there was
    a mistake; we are just trying to grade the
    answer.
```

```
Question: {question}
Gold target: {target}
Predicted answer: {predicted_answer}
```

```
Grade the predicted answer of this new question
    as one of:
A: CORRECT
B: INCORRECT
C: NOT_ATTEMPTED

Just return the letters "A", "B", or "C", with
    no text around it.
```

B Experimental Details

Compute Infrastructure. All experiments were conducted on NVIDIA H200 GPUs. We use `vLLM` for efficient batched inference during benchmarking and for high-throughput generation when collecting model outputs.

LLM Inference and Evaluation Parameters. For benchmark evaluation, decoding is performed with `vLLM`. We use a sampling temperature of $T = 0.7$ and otherwise retain the standard/default generation parameters provided by the inference framework to ensure consistent evaluation across model backbones (e.g., default settings for top-p, repetition controls, and maximum generation length).

LLM-as-a-judge. For automatic grading under the `Correct`, `Incorrect` or `Not_Attempted` protocol, we use `GPT-4.1` as the judge model. We set the judge temperature to $T = 0$ to minimize sampling variance and encourage deterministic scoring given the same inputs (question, gold target, and model prediction).

Supervised Fine-Tuning (SFT). For supervised adaptation on our Arabic text-grounded reasoning data, we follow standard instruction-tuning configurations in our training stack. The primary deviation from defaults is the learning rate, which we set to 5×10^{-5}. All other hyperparameters (e.g., optimizer choice, batch size, warmup schedule, gradient clipping, and number of epochs) use standard settings.

Preference-Based Alignment (RL). For alignment, we optimize the policy using the GSPO objective (*GSPO loss*). RL experiments are implemented using the `trl` library. We set the learning rate to 3×10^{-6} and keep remaining RL hyperparameters at standard values in our setup (e.g., rollout sampling and optimization schedule, and any applicable regularization/clipping). Reward signals are derived from the LLM-as-a-judge grading described above.

Retrieval-Augmented Generation (RAG). For dense retrieval, we embed both queries and Qur'anic verse units using `mE5-base`. We index all verse embeddings in `ChromaDB` as our vector database and retrieve relevant verses via vector similarity search at inference time. Retrieved verse units are provided as evidence context for standard RAG generation.

Agentic RAG Configuration. Our agentic RAG variant uses a fixed multi-turn interaction setups. We employ a two-turn setup: *(i)* an evidence-seeking turn in which the model invokes retrieval and inspects candidate verses, followed by *(ii)* a final answer turn conditioned on the retrieved and inspected evidence. The agentic tool-calling environment is implemented using the `verifiers` library, which enforces the structured multi-turn tool-use setups described in Appendix A.3.

C ISLAMICFAITHQA Examples

Figure 6 shows an example from ISLAMICFAITHQA and the corresponding prediction of AceGPT-v2-8B. The model generates a fluent and relevant response, but the answer is still incorrect because it misses the core idea expressed in the gold answer, which links doing good to true happiness, inner peace, and fulfillment. This example illustrates that the benchmark can expose subtle semantic errors even when the model output appears plausible.

D Training Data Examples

D.1 SFT Data Example

Figure 7 shows an example of an instruction–response pair used for supervised fine-tuning. The sample is grounded in Islamic legal reasoning and pairs an Arabic question with an answer that reflects the ruling supported by authentic textual evidence. This example illustrates the type of training signal used in our SFT data, where the model learns

Example: Islamic Ethics and Morality

Question: What is the main objective of training oneself to love doing good in Islam?
Gold answer: The main objective is to achieve true happiness, as loving and doing good leads to inner peace and fulfillment.
Model: AceGPT-v2-8B
Predicted answer: The main objective is to cultivate a habit of performing good deeds, improve character, earn rewards, and contribute positively to society.
Judge: INCORRECT

Figure 6: **IslamicFaithQA sample.**

to produce concise, grounded, and semantically faithful answers to domain-specific questions.

Example: SFT Training (Arabic + English)

Question (AR):
ما هو حكم المسح على الجوارب في الوضوء إذا كان الشخص قد أسلم بعد نزول آية الوضوء في سورة المائدة؟
Question (EN): What is the ruling on wiping over socks during wudū' if a person embraced Islam after the revelation of the wudū' verse in Sūrat al-Mā'idah?
Answer (AR):
يجوز المسح على الجوارب في الوضوء لمن أسلم بعد نزول آية الوضوء، لأن السنة النبوية الصحيحة بينت ذلك، كما في حديث المغيرة بن شعبة وغيره (رواه البخاري ومسلم).
Answer (EN): It is permissible to wipe over socks in wudū' for someone who embraced Islam after the wudū' verse, because authentic Sunnah establishes this (e.g., the hadith of al-Mughīrah ibn Shu'bah and others, reported in al-Bukhārī and Muslim).

Figure 7: **SFT sample (text-grounded reasoning).** An instruction-response instance used for supervised fine-tuning, including the Arabic question, a grounded reasoning trace, and a concise final answer (with an English translation for readability).

D.2 Reward-guided Data Example

Figure 8 presents an example from the RL data used for preference-based optimisation. Each instance consists of a question grounded in canonical Islamic text together with a concise gold answer that serves as the reference target.

E Human Annotation Guidelines

The annotation tasks involve labeling Islamic knowledge questions with *(i)* a difficulty score, *(ii)* reasoning requirements, *(iii)* multi-step reasoning requirements, and *(iv)* a single category label. Annotators follow the definitions and decision rules below to ensure consistent annotations.

Example: Reward-guided alignment (Arabic + English)

Question (AR):
كيف وصف ما صنعه كرحمة من ربه وما سيحدث له عند وعد ربه؟
Question (EN): How did he describe what he built as a mercy from his Lord, and what will happen to it when his Lord's promise comes?
Gold answer (AR):
كان رداً لا يستطيعوا اختراقه، وعند وعد ربه سيجعله دكاءً
Reference - Qur'an: 18:98–99
Gold answer (EN): It was a barrier they could not break through, and when his Lord's promise comes He will level it to the ground. **Reference - Qur'an: 18:98–99**

Figure 8: **RL preference sample.** An RL instance specifies a question derived from canonical text and an atomic gold target. During RL, candidate model responses are scored by an LLM-as-a-judge against this gold target to produce scalar rewards for policy optimisation (English translations are provided for readability).

E.1 Task Overview

For each question, annotators assign:

- a difficulty score on a 1–5 scale,
- `reasoning` (`true`/`false`): whether answering requires reasoning beyond simple recall,
- `multi_step` (`true`/`false`): whether the required reasoning involves multiple steps, and
- `category_type`: exactly one category label from a fixed set of 12.

E.2 Difficulty Annotation

Difficulty is defined with respect to a *competent respondent with Islamic knowledge*. Annotators should not base their judgment on personal familiarity, but instead on how difficult the question would be for this reference respondent to answer correctly. Each question is assigned a single integer score from 1 to 5 according to the definitions below.

E.2.1 Difficulty Definitions

E.2.2 Guidelines for Difficulty Assessment

Annotators consider:

- depth of knowledge required (basic vs. specialized),
- complexity of reasoning needed (recall vs. application vs. synthesis vs. balancing tradeoffs),
- number of concepts or sources involved,
- level of nuance (exceptions, conditions, context sensitivity, *khilāf*),
- whether simple recall suffices or reasoning is necessary.

Score	Label	Definition
1	Very Easy	Basic factual recall, simple definitions, or straightforward yes/no questions.
2	Easy	Requires basic understanding of concepts with minimal reasoning.
3	Moderate	Requires understanding multiple concepts and some analytical reasoning.
4	Hard	Requires deep understanding, synthesis of multiple sources/-concepts, and nuanced reasoning.
5	Very Hard	Requires expert-level analysis, balancing competing interests, and consideration of complex ethical frameworks.

Table 7: Difficulty rating scale used for question annotation.

Note: a question can be difficult due to obscure knowledge even if it is not multi-step.

E.2.3 Tie-break Rules

- Choose the higher score if mistakes are likely due to nuance, exceptions, or competing principles.
- Choose the lower score if the answer is direct and reliably determined from a single well-known rule or fact.
- If the question is underspecified, keep the score honest and flag the issue in the interface notes (if available).

E.3 Reasoning Assessment

E.3.1 `reasoning` (`true/false`)

- `reasoning = false` if the answer is simple recall/definition (no inference).
- `reasoning = true` if answering requires applying, interpreting, comparing, reconciling, justifying, or inferring.

E.3.2 `multi_step` (`true/false`)

Set `multi_step = true` only if multiple logical steps/considerations are required, such as:

- comparing multiple sources or viewpoints,
- weighing competing principles (harms vs. benefits, conflicting obligations),
- applying a rule, then an exception/condition, then concluding,
- building a chain with intermediate conclusions.

Set `multi_step = false` if reasoning is present but essentially one step (a single application or inference).

E.3.3 Consistency Rules

- If `reasoning = false`, then `multi_step` must be `false`.
- If `multi_step = true`, then `reasoning` must be `true`.

E.4 Category Classification

Annotators assign `category_type` to exactly one of the following category names (exact strings):

- Islamic Creed
- Jurisprudence
- Inheritance Law
- Hadith Studies
- Qur'anic Studies
- Prophetic Biography
- Islamic History
- Islamic Ethics and Morality
- Islamic Finance and Economics
- Islamic Family Law
- Comparative Religion
- Contemporary Issues

E.4.1 Boundary Rules

- Modern banking/finance products → `Islamic Finance and Economics`.
- Marriage/divorce/custody → `Islamic Family Law`; inheritance shares/heirs → `Inheritance Law`.
- Hadith authentication/narrators/classification → `Hadith Studies`; hadith used mainly to derive a ruling → usually `Jurisprudence`.
- *Sīrah* → `Prophetic Biography`; later eras/dynasties → `Islamic History`.
- Novel modern scenario → `Contemporary Issues`; timeless moral teaching → `Islamic Ethics and Morality`.

E.5 Ambiguity and Missing Context

Some questions may be underspecified or admit multiple valid scholarly answers. In such cases, annotators:

- still assign the best category based on the main domain being tested,
- rate difficulty based on what is required to answer responsibly (often higher if many qualifications are needed),

E.6 Examples

The examples in Table 8 illustrate how to apply the labels (they are not taken from the dataset).

Question	Label
What is *tawhīd*?	Difficulty: 1 Reasoning: false Multi_step: false Category_type: Islamic Creed
Explain the difference between *wājib* and *sunnah* acts.	Difficulty: 2 Reasoning: true Multi_step: false Category_type: Jurisprudence
A person touched their spouse and then prayed. Does this invalidate wudū'? Explain.	Difficulty: 3 Reasoning: true Multi_step: true Category_type: Jurisprudence
Compute inheritance shares when the deceased leaves a wife, two daughters, and parents.	Difficulty: 4 Reasoning: true Multi_step: true Category_type: Inheritance Law
Classify a hadith given narrator reliability and continuity of isnād.	Difficulty: 4 Reasoning: true Multi_step: true Category_type: Hadith Studies
Evaluate a modern bioethical dilemma by balancing harms/benefits and competing obligations.	Difficulty: 5 Reasoning: true Multi_step: true Category_type: Contemporary Issues

Table 8: Examples illustrating the annotation guideline.

E.7 Annotation Results and Agreement

We report summary statistics for the human annotation process and evaluate annotator consistency on a manually validated subset.

Difficulty labels. The difficulty annotations are reasonably distributed across the 1–5 scale, with the highest proportion at level 3 (26.90%), followed by level 2 (23.03%), level 4 (20.03%), level 1 (17.67%), and level 5 (12.37%).

Category assignment. We observe an overall agreement rate of 82.96% and a Cohen's κ of 0.62, indicating substantial agreement and supporting the reliability of the annotation guidelines.

F Retrieval Comparison

Our main experiments focus on end-to-end answer faithfulness under grounded generation, rather than isolating the retrieval engine in a separate benchmark. To better contextualize the retrieval component used in our RAG and Agentic RAG settings, we conducted a preliminary retrieval-side comparison between our dense multilingual retriever and a sparse lexical baseline.

Specifically, we compare `mE5-base`, the dense encoder used in our retrieval-augmented experiments, against BM25 on the IslamicEval 2025 Shared Task 2 dataset (Mubarak et al., 2025). We report Mean Average Precision at rank 5 (MAP_Q@5). The results show that `mE5-base` outperforms BM25 by a clear margin, achieving 0.138 compared with 0.094.

This result supports our choice of dense retrieval for two reasons. *First*, the task is bilingual (Arabic/English), and dense multilingual embeddings better align semantically equivalent queries across languages. *Second*, Quranic question answering often depends on theological and paraphrastic phrasing that is not always well matched by pure lexical overlap. At the same time, the absolute retrieval scores remain modest, suggesting that standard RAG based retrieval is still insufficient for many questions. This observation is consistent with our main findings. Standard RAG improves correctness, but Agentic RAG yields further gains by turning retrieval into an iterative evidence-seeking process rather than a one-pass preprocessing step.

We note that this comparison serves only as a preliminary sanity check on retrieval quality, not a full benchmark. A broader study of hybrid retrieval remains for future work.

G Label-wise Results

Table 9 reports the proportions of `correct`, `incorrect`, and `not attempted` predictions in Arabic and English. Fanar-2-27B achieves the best overall performance, with an average correct rate of 48.05%, followed by ALLaM-7B at 37.75% and Fanar-1-9B at 35.40%. This establishes a clear gap between the strongest models and the remaining systems.

The table also reveals different response behaviours. Models such as ALLaM-7B and Fanar-1-9B attempt most questions, but they also produce relatively high incorrect rates. By contrast, several Qwen3 and DeepSeek variants leave a large proportion of questions unanswered, often exceeding 70% `not attempted`, which leads to low overall correctness despite fewer incorrect responses. In general, many models perform better in English than in Arabic, whereas Fanar-2-27B remains comparatively balanced across both languages. These results show that separating `correct`, `incorrect`, and `not attempted` predictions offers a more informative view of model behaviour than reporting accuracy alone.

Model	Arabic			English			Avg.
	Correct	Incorrect	Not Attempted	Correct	Incorrect	Not Attempted	
Fanar-2-27B	48.20	21.50	30.30	47.90	30.20	21.90	**48.05**
ALLaM-7B	42.70	52.90	4.40	32.80	63.70	3.50	**37.75**
Fanar-1-9B	34.50	54.10	11.40	36.30	55.10	8.60	35.40
AceGPT-v2-8B	23.10	64.30	12.60	28.80	57.20	14.00	25.95
EuroLLM-9B	22.30	67.20	10.50	29.10	64.50	6.40	25.70
SILMA-9B-v1.0	20.40	70.90	8.70	28.50	66.10	5.40	24.45
Qwen3-4B-2507	15.80	45.30	38.90	27.90	45.20	26.90	21.85
gpt-oss-20b	15.90	22.60	61.50	27.20	27.20	45.60	21.55
Llama-3.1-8B	13.00	74.00	13.00	25.80	47.40	26.80	19.40
Mistral-7B-v0.2	13.50	53.50	33.00	24.40	59.10	16.50	18.95
SeaLLM-7B-v2.5	11.60	76.30	12.10	23.80	64.80	11.40	17.70
Qwen2.5-3B	11.00	61.20	27.80	20.00	63.20	16.80	15.50
Qwen3-14B	16.00	12.50	71.50	14.00	4.20	81.80	15.00
Llama-2-7b	4.40	47.20	48.40	18.80	72.00	9.20	11.60
DeepSeek-R1-0528-Qwen3-8B	6.30	17.70	76.00	11.90	13.30	74.80	9.10
Qwen3-8B	8.80	10.00	81.20	8.50	5.60	85.90	8.65
Qwen3-4B-Thinking-2507	6.50	3.10	90.40	9.40	14.20	76.40	7.95
Qwen3-4B	6.50	17.30	76.20	9.00	9.60	81.40	7.75
Qwen3-1.7B	3.20	18.10	78.70	5.10	12.60	82.30	4.15
Qwen3-0.6B	1.30	47.00	51.70	5.40	39.70	54.90	3.35
DeepSeek-R1-Distill-Qwen-7B	1.40	37.40	61.20	4.30	37.60	58.10	2.85
DeepSeek-R1-Distill-Qwen-1.5B	0.10	21.60	78.30	1.00	41.10	57.90	0.55

Table 9: Results including correct, incorrect and not attempted.