Advances in AI Systems on Islamic Knowledge Capabilities: A Critical Survey

Gagan Bhatia[1], Hamdy Mubarak[1], Majd Hawasly[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.

Contributing authors: fialam@hbku.edu.qa;

Abstract

AI systems are increasingly mediating how Islamic communities access, study, and apply Islamic sources; still, research on Islamic-knowledge capabilities remains fragmented across NLP, information retrieval, speech, multimodal learning, educational technology, and recent LLM alignment work. This survey presents a critical systematic review of 160 papers from the past decade that incorporate Islamic knowledge in Machine Learning/AI. We propose a layered taxonomy that separates an epistemic view of Islamic knowledge (authority-bearing foundations and established disciplines) from an instrumental AI task layer (data and corpora, retrieval and grounding, understanding, reasoning support, evaluation and governance, and multimodal methods), while treating normative concerns as cross-cutting constraints. Using this framework, we synthesize trends in datasets, benchmarks, and system architectures, highlighting the shift toward retrieval-grounded LLM pipelines, verification and deferral mechanisms, and emerging multimodal recitation and manuscript-processing systems. We also consolidate evaluation practices for trustworthiness, including provenance and faithfulness, disagreement-aware and school of thought-sensitive framing, calibrated abstention under underspecified queries, and safety and bias assessment for Islamic contexts. Finally, we identify deployment-critical gaps and engineering priorities for building auditable, pluralism-aware, and risk-sensitive Islamic-knowledge AI. Project webpage: https://gagan3012.github.io/islamic-knowledge-survey/

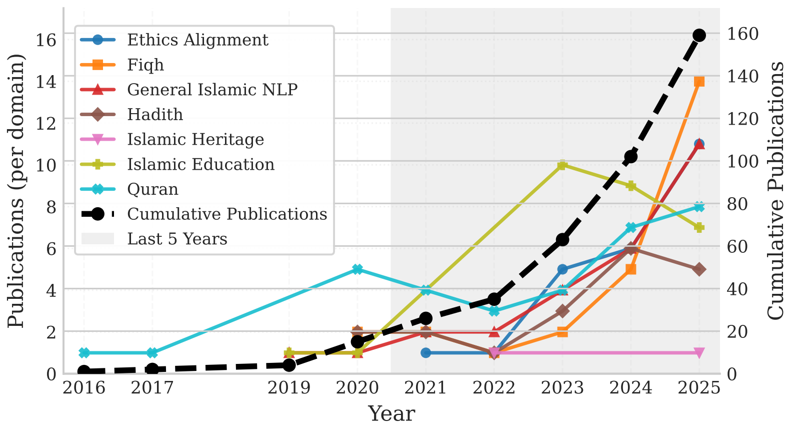


Fig. 1: Publication trends over time. The plot illustrates the research output volume, highlighting a significant surge (+104% growth) in the last three years.

Keywords: Islamic knowledge, Large language models, Retrieval-augmented generation, Trustworthy AI, Evaluation benchmarks, Cultural alignment

1 Introduction

Islam-related content is growing at an unprecedented pace to address the needs of over 2 billion people worldwide (Harvey 2020; Mahdi 2026). Artificial Intelligence (AI) systems are playing an increasingly central role in shaping how this content is accessed, understood, and applied—whether through question-answering, summarization, tutoring tools, recitation apps, or conversational assistants. The recent transition from task-specific NLP techniques to powerful foundation models and large language models (LLMs) has unlocked remarkable opportunities, while also presenting new and complex challenges. Though these systems can broaden access to Islamic sources and scholarship, integrating *"Islamic knowledge"* into AI is challenging due to diverse sources and differing interpretations (Atif et al. 2025). Further, such system can also hallucinate citations and fabricate attributions (Mubarak et al. 2025; Alansari and Luqman 2025; Hosseini et al. 2025), collapse nuanced juristic disagreement into a single "answer," (Atif et al. 2025; Bouchekif et al. 2025b) and reproduce cultural or religious stereotypes (Plaza-del Arco et al. 2024).

Research on AI and Islamic knowledge has increased over the past decade. Contributions appear in Natural Language Processing (NLP) and Information Retrieval (IR) for Quran and ḥadīth processing (Yasser Shohoud 2023; Asma Abdul-Qader Abdullah Al-Ani 2024; Haval H. Ameen 2024), in resource construction and knowledge representation (e.g., linguistic annotation, embeddings, ontologies) (Akra et al. 2025; M. Othman 2020; Antoun et al. 2020; Bhatia et al. 2025; Patel et al. 2023), and more recently, in LLM evaluation and alignment research that examines values, bias, safety, and culturally grounded behavior (Alwajih et al. 2025c; Guo et al. 2025; Meadows et al. 2024; Plaza-del Arco et al. 2024; Hui et al. 2025). As a result, models frequently exhibit what Yu et al. (2025) refer to as *cultural flattening*: a tendency to overgeneralize or misrepresent religious values (Guo et al. 2025; Meadows et al. 2024; Keleg 2025).

Despite the rapid growth in NLP, IR, speech, and LLM research (as shown in Figure 1), AI research on Islamic knowledge has followed multiple paths. Researchers have introduced diverse datasets, models, and evaluation methods. Recent studies have uncovered concerns about bias, as current LLMs inherit the value systems and cultural bias of their training data, which are largely Western-centric and skewed towards high-resource languages.(Sun et al. 2025; Naous et al. 2023; Naous and Xu 2025; Zhong et al. 2024), and biased AI outputs may reinforce stereotypes about Muslims and Arabs. Moreover, risks could emerge from hallucinations and errors, e.g. in Islamic inheritance or financial rulings, harming rights (Bouchekif et al. 2025a).

The diversity of these works offers a rich foundation, but also highlights the need for a unified view that connects and synthesizes contributions, clarifies underlying assumptions, and enables more systematic cross-comparison. This survey aims to delve into these efforts, catalog and summarize the resources developed over the past decade, examine strategies for aligning LLMs with religious values, and propose taxonomies that organize this space. In this survey, we use *Islamic knowledge* in an operational sense to refer to AI system capabilities on Islamic knowledge-related tasks, such as categorizing religious content, understanding and answering questions, and generating or assessing responses grounded in Islamic sources. Following PRISMA (Page et al. 2021) framework (see Section 2), which helps map the rapidly evolving research efforts while minimizing selection bias, we systematically review over 160 papers published in the last ten years. Our goal is not only to synthesize methods and findings, but also to provide a conceptual map of how AI research operationalizes Islamic knowledge, how it evaluates correctness and trustworthiness, and where it risks flattening diverse traditions, communities, and worldviews.

This survey consolidates Islamic-knowledge capabilities in modern AI systems. We: *(1)* propose a taxonomy that is grounded by epistemic dimensions; *(2)* map the available datasets into five streams (pretraining corpora, knowledge bases/ontologies, Qur'an/Hadith resources incl. audio/vision, jurisprudence/reasoning benchmarks, and ethical/cultural evaluations); *(3)* review evaluation methodologies and shared tasks; *(4)* analyze deployment-critical failure modes (e.g., source hallucination, attribution errors, juristic disagreement); *(5)* synthesize alignment approaches (retrieval grounding, authority-aware filtering, abstention, and governance-oriented evaluation); and *(6)* outline open research gaps and engineering priorities for trustworthy religion-aware systems. We frame Islamic alignment as a case study in culturally grounded, domain-specific alignment in a low-resource yet high-stakes setting.

1.1 Use Cases

In practice, AI systems related to Islamic-knowledge is deployed in a small number of recurring settings that stress different reliability constraints: *(i)* scripture-grounded study and reference assistants for Quran/Hadith QA, where the dominant risks are provenance breakage and source fabrication rather than fluency (Mubarak et al. 2024;

Prior survey	**Year**	**Primary scope**	**Main research areas**	**Prov./Disagr. Abst.**	**MM**	**Norm./ Plur.**	**Methods**
Azmi et al. (2019)	2019	Hadith-focused	IR/ML/DL (pre-LLM)	˜	–	˜	Narrative (no PRISMA)
Bashir et al. (2023)	2023	Qur'an-focused	IR/ML/DL; limited Transformers; pre-LLM/RAG	˜	✓	–	Inclusion/exclusion + flow
Alnefaie et al. (2023)	2023	Islamic QA (Qur'an /Hadith /Fatwa)	Retrieval + PLMs; task framing	˜	–	˜	Survey + evaluation criteria
Ahmad et al. (2025)	2025	Qur'anic education	AI-in-education (systematic review)	–	–	˜	Systematic review
Hakim and Anggraini (2023)	2023	Teaching Islamic studies	AI-in-education (systematic review)	–	–	˜	Review (PRISMA referenced)
Alhammad et al. (2025)	2025	Islamic education	Pedagogy + learning outcomes (review)	–	–	˜	Review
Mashaabi et al. (2024)	2024	Arabic LLMs (general)	Transformers/LLMs; resources	–	–	˜	Method + openness/resources
Rhel and Roussinov (2025)	2025	Arabic LLMs (general)	PLMs/LLMs; prompting; evaluation trends	–	–	–	Review-style
Abouzied et al. (2025)	2025	Arabic LLMs landscape	LLMs + benchmarks; harm/bias themes	˜	–	˜	Landscape (no PRISMA)
Alzubaidi et al. (2025)	2025	Arabic LLM benchmarks	LLM evaluation; benchmark taxonomy	˜	–	˜	Systematic review
Asseri et al. (2025)	2025	Bias (Arabs/Muslims)	Prompting/pipelines; bias measurement	–	–	✓	PRISMA
Ours	**2026**	**Whole Islamic stack**	**IR, Transformers, LLM/RAG/agents; end-to-end evaluation**	✓	✓	✓	**PRISMA-ScR + reproducibility artifacts**

Table 1: Comparison of prior surveys vs. coverage dimensions relevant to Islamic-knowledge AI. Legend: ✓= explicit focus; ˜= partial/implicit; – = not covered. MM = multimodal; Norm./Plur. = normative/pluralism considerations. MM = multimodal; Norm./Plur. = normative/pluralism considerations; PLM = Pre-trained Language Models; RAG = Retrieval-Augmented Generation.

Alansari and Luqman 2025); *(ii) fiqh* [1]/fatwa [2] decision support, including structured reasoning tasks and school-aware disagreement handling, where systems must qualify answers under legitimate *ikhtilāf* [3] and defer when queries are underspecified (Atif et al. 2025; Bouchekif et al. 2025a); *(iii)* practice-support workflows (e.g., Hajj/Umrah guidance) that are operationally high-stakes and therefore demand conservative abstention and traceable sourcing (Aleid and Azmi 2025); *(iv)* multimodal recitation coaching and tajwīd [4] feedback, where end-to-end performance depends on robust speech pipelines (El Kheir et al. 2025; Mazid and Ahmad 2025); and *(v)* culturally grounded safety and trust layers that evaluate bias, harmful framings, and governance criteria specific to Muslim users and contexts (Lahmar et al. 2025; Alwajih et al. 2025b; Fawzi et al. 2024, 2025).

1.2 How This Survey Differs from Prior Reviews

Prior reviews in this space have progressed along parallel tracks. Surveys of Quranic NLP/IR cover foundational tasks such as text processing, retrieval, and annotation (Bashir et al. 2023). Hadith-focused surveys emphasize authenticity-centric processing of narrations and metadata, largely using classical NLP/IR methods (Azmi

[1] Islamic jurisprudence; the human understanding of divine law.

[2] A formal ruling or interpretation on a point of Islamic law given by a qualified legal scholar.

[3] Scholarly intellectual difference of opinion among Islamic jurists

[4] The set of rules governing the correct pronunciation and recitation of the Qur'an.

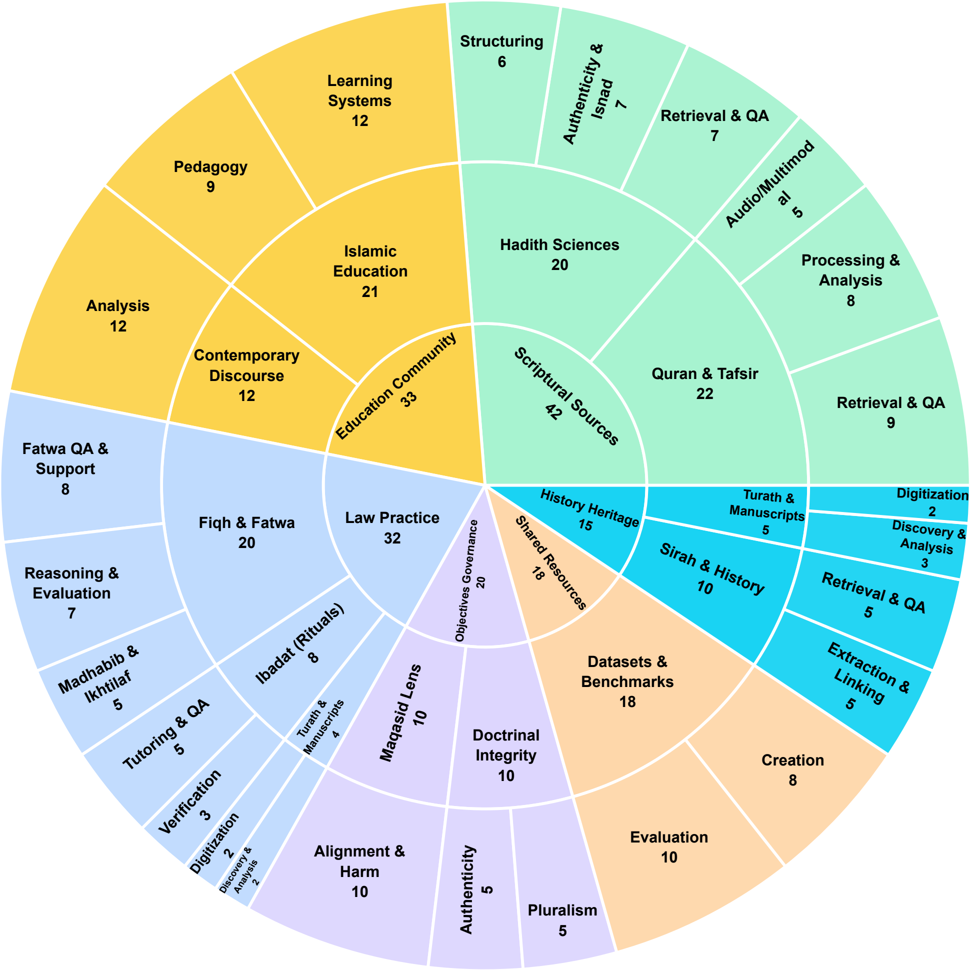


Fig. 2: A sunburst chart depicting the hierarchical distribution of the surveyed literature. The inner ring represents major domains while the outer rings detail specific sub-applications and methodologies. Numbers denote the count of surveyed papers tagged as covering that topic. Consequently, overlap across branches is expected: some cross-cutting themes (e.g., *Fatwa QA*, *Retrieval & QA*, *Digitization*) may appear under multiple parent domains, and repeated labels indicate the same theme manifested in different contexts rather than duplicate papers.

et al. 2019). More recently, surveys of Arabic LLMs have mapped major model families, training data, and general evaluation benchmarks (Abouzied et al. 2025; Alzubaidi et al. 2025), while separate systematic reviews have examined bias and cultural harms toward Arabs and Muslims and proposed mitigation strategies (Asseri et al. 2025).

Despite this valuable coverage, Islamic-knowledge capabilities are often spread across specialized literatures. Scriptural QA, *fiqh* reasoning, educational tools, multimodal recitation/Optical Character Recognition (OCR) pipelines, and culturally grounded safety evaluation are typically studied in different venues and under incompatible assumptions. As a result, it is difficult to compare contributions across subfields, and LLM-era deployment requirements are not consistently captured in task taxonomies or evaluation designs, especially provenance and faithfulness constraints, calibrated abstention/deferral, and pluralism-aware answering under legitimate disagreement.

This survey addresses the gap by synthesizing a decade of work across NLP, IR, speech and multimodal learning, educational technology, and LLM alignment. We follow a **PRISMA-ScR** (Page et al. 2021) guided methodology and propose a multi-dimensional taxonomy that links Islamic knowledge domains and task families to cross-cutting normative dimensions (e.g., authority, provenance, disagreement handling, and governance). Table 1 summarizes how our survey differs from prior surveys, including end-to-end coverage AI systems for Islamic knowledge, in-depth analysis of Retrieval-augmented Generation (RAG)/agentic approaches, multimodal Islamic AI, and evaluation designs for faithfulness, pluralism, and abstention.

2 Scope of the Survey

To ensure a systematic, transparent, and reproducible review of the literature, we adhered to the **PRISMA-ScR** framework as presented in Figure 3. This approach allows us to map the rapidly evolving research efforts while minimizing selection bias. Our systematic selection process followed a structured workflow encompassing *identification*, *screening*, and *inclusion*.

2.1 Research Questions

This systematic review is guided by three research questions (RQs) that reflect both the technical and sociotechnical requirements of AI systems for Islamic knowledge:
RQ1 (Domains and tasks) Which Islamic knowledge domains and application tasks have been operationalized in ML/AI systems over the past decade, and how are these efforts distributed across subfields (e.g., NLP, IR, speech/multimodal, educational technology, and LLM alignment)?
RQ2 (Resources and measurement) What datasets, benchmarks, and knowledge resources are available for Islamic-knowledge capabilities, what do they measure, and what assumptions do they encode about evidence, provenance, authority, and interpretive diversity?
RQ3 (Evaluation and trustworthiness) How do studies evaluate trustworthiness in Islamic-knowledge settings, especially source faithfulness and provenance, doctrinal/legal correctness, pluralism-aware answering under legitimate disagreement, calibrated abstention/deferral, and safety/bias and what evaluation gaps remain for deployment?

These research questions determine *(i)* the inclusion criteria during full-text assessment, *(ii)* the structured extraction schema used for each paper (domain, task family, method stack, resource contributions, and evaluation design), and *(iii)* the comparative synthesis strategy used in subsequent sections.

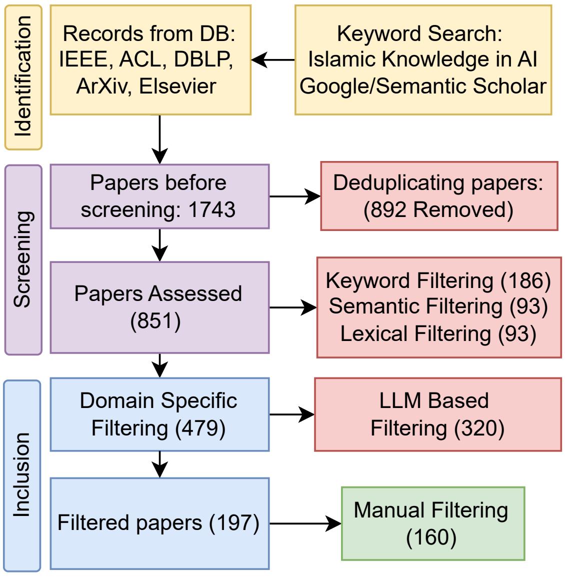

Identification
Records from DB: IEEE, ACL, DBLP, ArXiv, Elsevier
Keyword Search: Islamic Knowledge in AI Google/Semantic Scholar
Screening
Papers before screening: 1743
Deduplicating papers: (892 Removed)
Papers Assessed (851)
Keyword Filtering (186)
Semantic Filtering (93)
Lexical Filtering (93)
Inclusion
Domain Specific Filtering (479)
LLM Based Filtering (320)
Filtered papers (197)
Manual Filtering (160)

Fig. 3: PRISMA flow diagram detailing the identification, screening, and inclusion process of the surveyed literature.

2.2 PRISMA-ScR Screening Flow

2.2.1 Identification.

To maximize coverage, we searched Google Scholar and Semantic Scholar and obtained papers from IEEE Xplore, ACM Digital Library, Elsevier, DBLP, the ACL Anthology and arXiv. The search strategy executed a predefined set of queries across four pillars: *(i)* core Islamic NLP (e.g., "Fatwa generation LLM", "Fiqh LLMs"); *(ii)* Arabic language modelling (e.g., "Arabic NLP transformer"); *(iii)* evaluation resources (e.g., "Islamic NLP dataset"); and *(iv)* cultural alignment (e.g., "Islamic ethics AI"). This process yielded an initial corpus of 1,743 papers published between 1994 and 2025. Following retrieval, we performed deduplication using fuzzy matching algorithms applied to paper titles, author lists, and abstracts. This step removed 892 duplicate entries, resulting in 851 unique papers for screening. For exact search terms please see Appendix A.2.

2.2.2 Screening.

We applied a multi-stage filtering pipeline to the 851 unique papers to ensure domain relevance:

1. **Keyword Filtering:** We filtered papers based on the presence of essential domain-specific keywords, removing 186 non-relevant entries.
2. **Semantic Filtering:** To capture relevance beyond exact keyword matches, we utilized the Qwen3 embedding models (Yang et al. 2025). We generated embeddings

for titles and abstracts and retained only those papers satisfying a cosine similarity threshold of ≥ 0.7 against our core topic descriptions (see Appendix A.2) (excluding 93 papers).

3. **Lexical Filtering:** We applied lexical rules to identify and remove non-archival documents, pre-prints without metadata, or incomplete entries, excluding another 93 papers.

This rigorous screening phase reduced the set to 479 papers, which were selected for further assessment.

2.2.3 Inclusion.

In the final inclusion phase, we utilized GPT-4.1 (OpenAI 2023) to analyze the full text of the remaining manuscripts (prompt can be found in Appendix A.1.1). The model was prompted to evaluate the substantive relevance of each paper's content to the intersection of AI systems related to Islamic knowledge, filtering out papers that mentioned religious terms sparsely. This step refined the selection to 197 papers.

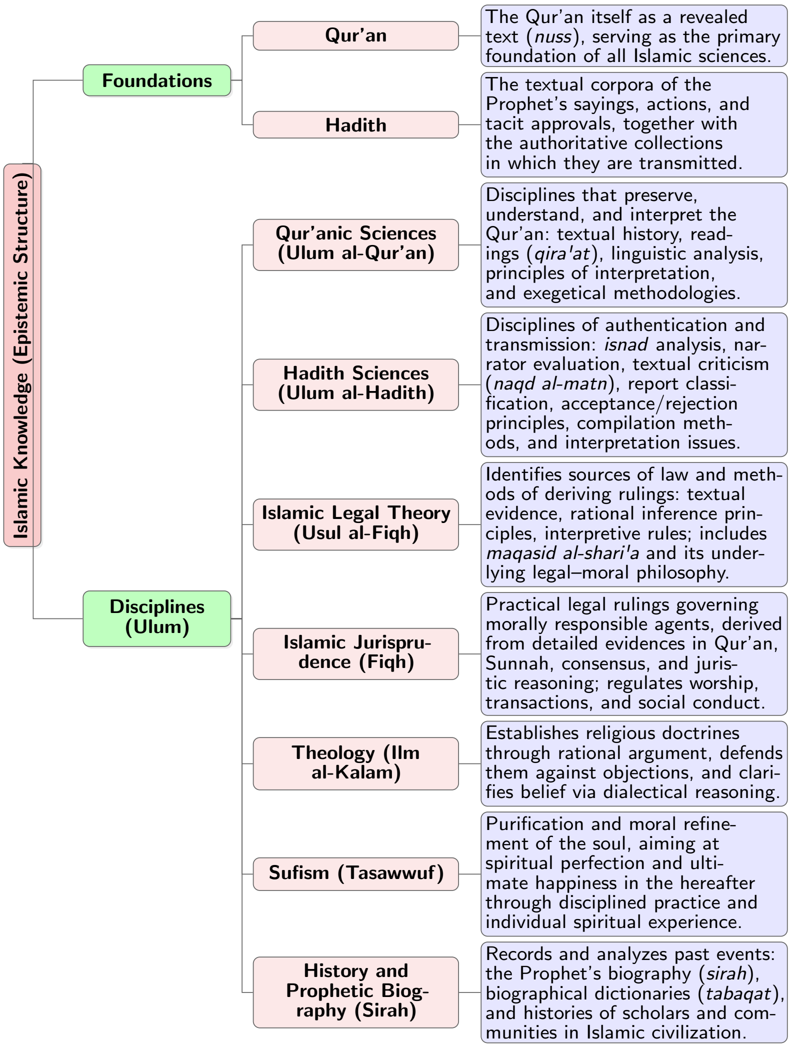
Finally, a ***manual review*** was conducted by two independent authors to verify the LLM's selections and ensure strict alignment with our research questions. Disagreements were resolved through discussion and consensus. This process resulted in a final corpus of **160 papers**, which form the basis of the subsequent analysis.

3 Taxonomy

To answer RQ1-RQ3 with systematic cross-comparison, we introduce a taxonomy to serve as the paper's unifying analytical framework. The taxonomy provides a consistent coding scheme. For RQ1, it maps Islamic knowledge domains using an epistemic organization grounded in traditional authorities and disciplines (Figure 4). For RQ2, it organizes the datasets, benchmarks, and resources that enable AI capabilities and specifies what they measure. For RQ3, it compares how trustworthiness is evaluated and governed (e.g., provenance/faithfulness, disagreement-aware framing, abstention/deferral, bias/safety, and governance), aligned with the evaluation and governance layer in Figure 5. Many requirements cut across applications, including authenticity, authority, pluralism, and risk sensitivity. We therefore use a layered, multi-dimensional structure: the *epistemic layer* organizes domains, the *AI task layer* organizes computational methods, while normative concerns act as cross-cutting lenses.

3.1 Taxonomy Development

Islamic-knowledge AI spans a broad spectrum of tasks, ranging from scriptural processing and QA to jurisprudential reasoning, heritage digitization, multimodal recitation/OCR, and LLM alignment. Synthesizing the decade of research therefore requires a shared organizing framework that is comparable across AI subfields while remaining sensitive to religious requirements such as authority, provenance, and interpretive plurality. We address this need by proposing a layered taxonomy that links *(i)* an *epistemic structure* of Islamic knowledge (foundations and disciplines; Figure 4) (Alatas 2013; Bellino 2014; Versteegh 2014; Rippin 2022; Brown 2008; Schmidtke 2016; Elston 2022),


Islamic Knowledge (Epistemic Structure)
Foundations
Qur'an
Hadith
Disciplines (Ulum)
Qur'anic Sciences (Ulum al-Qur'an)
Hadith Sciences (Ulum al-Hadith)
Islamic Legal Theory (Usul al-Fiqh)
Islamic Jurisprudence (Fiqh)
Theology (Ilm al-Kalam)
Sufism (Tasawwuf)
History and Prophetic Biography (Sirah)

The Qur'an itself as a revealed text (*nuss*), serving as the primary foundation of all Islamic sciences.

The textual corpora of the Prophet's sayings, actions, and tacit approvals, together with the authoritative collections in which they are transmitted.

Disciplines that preserve, understand, and interpret the Qur'an: textual history, readings (*qira'at*), linguistic analysis, principles of interpretation, and exegetical methodologies.

Disciplines of authentication and transmission: *isnad* analysis, narrator evaluation, textual criticism (*naqd al-matn*), report classification, acceptance/rejection principles, compilation methods, and interpretation issues.

Identifies sources of law and methods of deriving rulings: textual evidence, rational inference principles, interpretive rules; includes *maqasid al-shari'a* and its underlying legal–moral philosophy.

Practical legal rulings governing morally responsible agents, derived from detailed evidences in Qur'an, Sunnah, consensus, and juristic reasoning; regulates worship, transactions, and social conduct.

Establishes religious doctrines through rational argument, defends them against objections, and clarifies belief via dialectical reasoning.

Purification and moral refinement of the soul, aiming at spiritual perfection and ultimate happiness in the hereafter through disciplined practice and individual spiritual experience.

Records and analyzes past events: the Prophet's biography (*sirah*), biographical dictionaries (*tabaqat*), and histories of scholars and communities in Islamic civilization.

Fig. 4: Epistemic taxonomy of Islamic knowledge: authority-bearing foundations (revelation and the Turath corpus) and the historically established disciplines derived from them.

and *(ii)* an *instrumental AI task layer* (methods applicable across disciplines; Figure 5). The taxonomy was developed iteratively using embedding-based clustering to surface themes, topic modeling and LLM-assisted labeling to name the themes, and manual consolidation into a stable layered hierarchy used for systematic coding and synthesis.

3.1.1 Semantic Embedding and Clustering

We generated vector representations for the titles and abstracts of all the 160 included papers using Qwen3 Embedding models (Yang et al. 2025), capturing their semantic

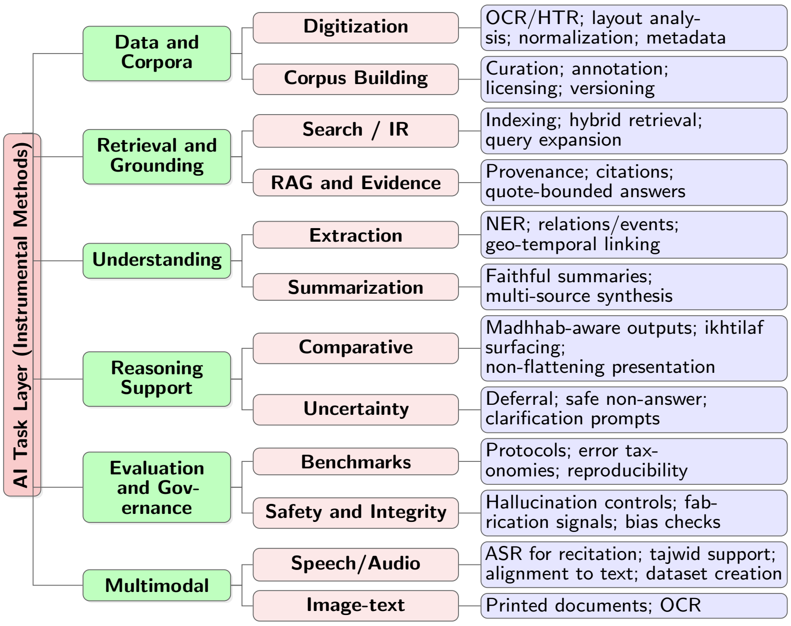

AI Task Layer (Instrumental Methods)
Data and Corpora
Digitization
OCR/HTR; layout analysis; normalization; metadata
Corpus Building
Curation; annotation; licensing; versioning
Retrieval and Grounding
Search / IR
Indexing; hybrid retrieval; query expansion
RAG and Evidence
Provenance; citations; quote-bounded answers
Understanding
Extraction
NER; relations/events; geo-temporal linking
Summarization
Faithful summaries; multi-source synthesis
Reasoning Support
Comparative
Madhhab-aware outputs; ikhtilaf surfacing; non-flattening presentation
Uncertainty
Deferral; safe non-answer; clarification prompts
Evaluation and Governance
Benchmarks
Protocols; error taxonomies; reproducibility
Safety and Integrity
Hallucination controls; fabrication signals; bias checks
Multimodal
Speech/Audio
ASR for recitation; tajwid support; alignment to text; dataset creation
Image-text
Printed documents; OCR

Fig. 5: Instrumental AI task taxonomy: computational functions applicable across multiple Islamic disciplines. This layer is not an epistemic classification; it enumerates tools (retrieval, grounding, extraction, evaluation) that may be mapped many-to-many onto the epistemic disciplines in Figure 4.

nuances. We then applied **K-Means clustering** to partition the literature into distinct thematic groups. Silhouette analysis indicated that $k = 11$ clusters achieved the optimal balance of granularity and cohesion.

3.1.2 Topic Modeling and Interpretation

To interpret these clusters and construct the layered taxonomy shown in Figures 4 and 5, we employed a hybrid methodology combining Latent Dirichlet Allocation (LDA) and the Gemini 3 Pro Large Language Model (Comanici et al. 2025). LDA was first used to extract dominant keywords and probabilistic topics from the clusters. We then utilized Gemini to refine these topics, generate semantically plausible labels, and organize them into the hierarchical taxonomy.

3.1.3 Manual Revision

While the unsupervised clustering described in Section 3.1.1 effectively surfaced the latent thematic distribution of the literature, it lacked the epistemic structure necessary for a domain-faithful framework and the separation between *knowledge classification* and *instrumental method*. To address this, we manually consolidated the empirical clusters into two coordinated taxonomies: an epistemic taxonomy of Islamic knowledge (foundations and disciplines; Figure 4) and an AI task taxonomy (instrumental methods;

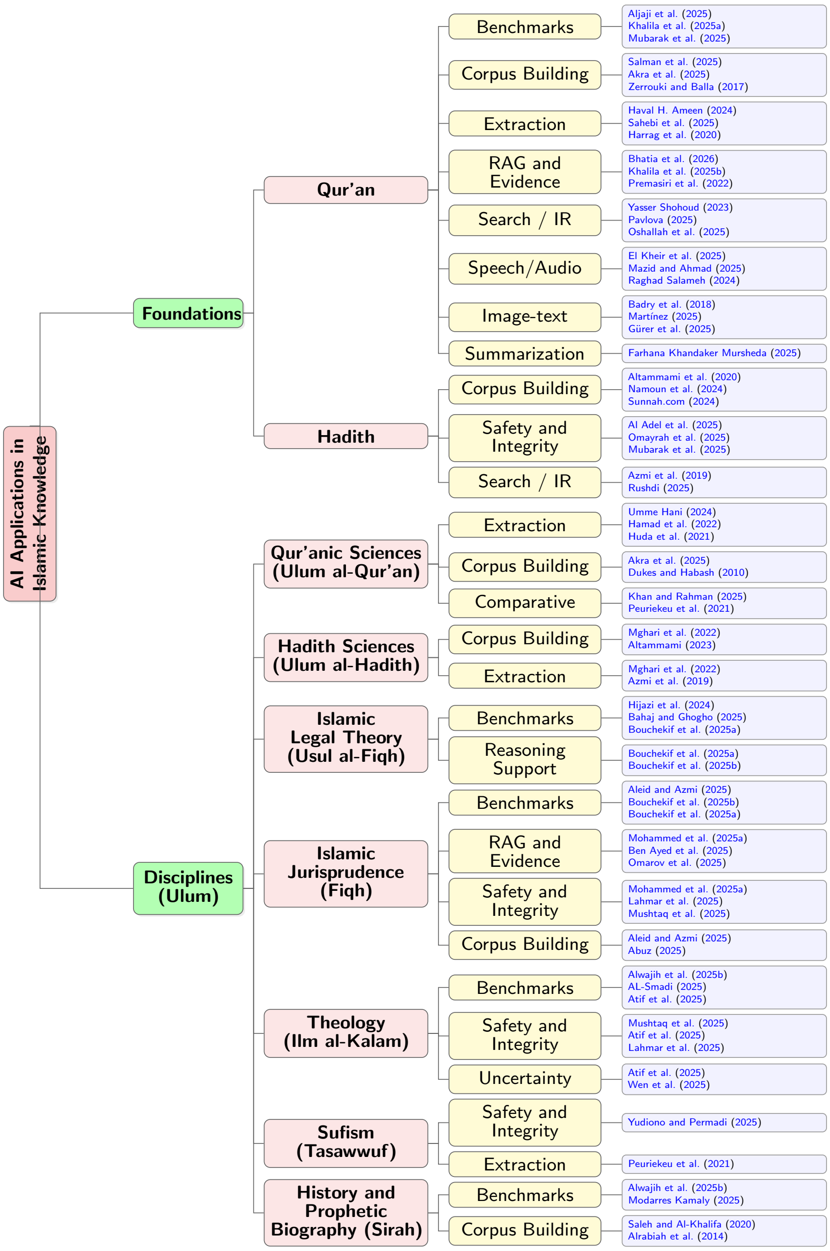


Fig. 6: Mapping the epistemic domains in Figure 4 (Foundations and Disciplines) to the AI task categories in Figure 5. This figure makes explicit the many-to-many relationship between *what* knowledge is being modelled and *how* it is handled computationally for different domains, highlighting recurring trust-critical requirements such as provenance-preserving grounding, safety and integrity controls, and rigorous benchmarking.

Figure 5). This separation clarifies that Quran and Sunnah are treated as authority-bearing sources (not merely content), and that computational functions (retrieval, grounding, extraction, evaluation) map many-to-many onto disciplines.

3.2 Organizing Layers

3.2.1 Epistemic Layer (Sources and Scholarly Tradition)

Figure 4 provides an organization of Islamic knowledge that we use to code *what* type of religious material a system engages and which scholarly methods govern its interpretation. This layer is designed to prevent a common failure mode in Islamic-knowledge AI, treating all "Islamic text" as homogeneous content, by distinguishing *(i)* sources that carry direct normative authority and *(ii)* disciplines that regulate how those sources are read, reconciled, and applied.

Foundations

We distinguish the two primary revealed textual foundations (*nuss*) upon which the Islamic sciences are built:
(1) Qur'an. The Qur'an is the revealed text and the primary foundation of all Islamic disciplines. In computational settings, this foundation motivates high-integrity representations of the Qur'anic text and its transmission, including verse-boundary integrity, script/orthography normalization, stable alignment across translations, and safeguards against fabricated or misattributed citations.

(2) Ḥadīth. Ḥadīth denotes the textual corpora of the Prophet's sayings, actions, and tacit approvals, together with the authoritative collections through which they are transmitted. In AI systems, this foundation motivates collection-aware retrieval, normalization across editions, precise attribution, and careful separation between transmitted report text and later interpretation or commentary.

Disciplines (ʿUlum – Islamic Sciences and Methodologies)

Building on the foundational texts, the epistemic layer enumerates the major disciplines that arose to preserve, authenticate, interpret, and systematize Islamic knowledge.
(1) Qur'anic Sciences (ʿ*Ulūm al-Qur'an*). This set of disciplines developed around the Qur'an in order to preserve, understand, and interpret it, including the history of the Qur'anic text, qirā'āt, linguistic analysis, principles of interpretation, and exegetical methodologies. Computational work here often concerns Qur'an-centric retrieval and explanation, alignment with tafsīr literature, meaning disambiguation across contexts, and evidence-linked generation rather than free-form paraphrase.

(2) Ḥadīth Sciences (ʿ*Ulūm al-Ḥadīth*). Distinct from Ḥadīth as a foundational corpus, Ḥadīth sciences comprise the methodological apparatus for authentication and transmission: isnād analysis, narrator evaluation, textual criticism (naqd al-matn), classification of reports, principles of acceptance and rejection, methods of compilation, and issues related to interpreting reports. AI work in this domain includes chain modeling, narrator/entity linking, network analysis, classification of authenticity indicators, and interfaces that surface uncertainty and plural assessments rather than presenting a single definitive label.

(3) Islamic Legal Theory (Uṣūl al-Fiqh). This discipline identifies the sources of law and the methods by which legal rulings are derived from them. It examines textual evidences, rational principles of inference, and interpretive rules, and includes the objectives of the Sharī'a (maqāṣid al-sharī'a) and their underlying legal and moral philosophy. Computationally, it motivates representations of evidential hierarchy and structured inference, argument–evidence schemas, and evaluations that test whether systems respect conditions of application and the logic of derivation.

(4) Islamic Jurisprudence (Fiqh). Fiqh is the science of practical legal rulings governing the actions of morally responsible agents, derived from detailed evidences in the Qur'an, the Sunnah, consensus, and juristic reasoning. It regulates worship, transactions, and social conduct. In AI systems, this motivates comparative ruling retrieval, disagreement-aware handling across schools of thought, context-sensitive questioning when inputs are underspecified, and safe deferral when a ruling cannot be responsibly produced.

(5) Theology (*ʿIlm al-Kalām*). Kalām establishes religious doctrines through rational argument, defends them against objections, and clarifies matters of belief via dialectical reasoning. In computational settings, this raises doctrinal-integrity requirements: avoiding fabricated attributions, grounding claims in recognized sources, and careful framing where theological schools legitimately diverge.

(6) Sufism (Taṣawwuf). Taṣawwuf is devoted to purification and moral refinement of the soul, aiming at spiritual perfection and ultimate happiness in the hereafter through disciplined practice and individual spiritual experience. For AI applications, it motivates genre-sensitive summarization and pedagogy, norm- and tone-aware guidance, and safeguards against overconfident personalization when texts are prescriptive or spiritually directive.

(7) History and Prophetic Biography (Sīrah). This discipline records and analyzes past events, including the Prophet's biography (sīrah), biographical dictionaries (ṭabaqāt), and histories of scholars and communities in Islamic civilization. Computational work commonly involves event extraction, geo-temporal linking, timeline QA, and provenance-aware summarization across sources that may disagree, making uncertainty handling central.

3.2.2 AI Task Layer (Instrumental Methods).

Figure 5 organizes *how* AI systems operate in a discipline-agnostic way. This layer is explicitly *instrumental* rather than epistemic, and enumerates reusable computational capabilities that apply across Quran, hadith, fiqh, Turath, history, and education. We use it to answer RQ2 (resources/benchmarks and what they measure) and RQ3 (evaluation, safety, and governance), while enabling cross-domain comparisons (e.g., "RAG and evidence" for tafsir vs. fatwa QA).

Data and Corpora.

This **category** covers the data and corpus infrastructure that enables reliable downstream modeling and evaluation.

(1) Digitisation includes OCR/HTR for manuscripts and printed heritage, layout analysis, normalization, and metadata extraction, often dealing with script variability, marginalia, complex page structures, and imperfect scans.

(2) Corpus Building covers curation and annotation practices (including schema design), licensing constraints, dataset documentation, and versioning critical for provenance, reproducibility, and governance in religious settings where source authenticity and permitted use matter.

Retrieval and Grounding.

This category organises methods that connect user queries or model outputs to verifiable sources.

(1) Search / IR includes indexing, ranking, hybrid retrieval (lexical + embedding), and query expansion. In Islamic-knowledge systems, IR must often cope with spelling/orthography variation, multiple editions, and cross-lingual access (Arabic with translations).

(2) RAG and Evidence captures retrieval-augmented generation and evidence linking: provenance tracking, citation generation, quote-bounded answering, and constraints that reduce hallucinated attributions. This category is central to trustworthiness, since many religious applications require users to inspect the underlying source rather than accepting free-form synthesis.

Understanding.

This **category** captures content interpretation tasks that structure text for downstream use.

(1) Extraction includes named entity recognition, relation and event extraction, and geo-temporal linking—particularly salient for *sirah*/history and for *isnad* narrator graphs.

(2) Summarisation emphasises faithful summaries and multi-source synthesis. Here, "faithful" means preserving meaning without introducing new claims, and "multi-source" requires surfacing disagreements rather than collapsing them into a single narrative when the sources diverge.

Reasoning Support.

This **category** captures methods that assist users with normatively sensitive inference without overstating certainty.

(1) Comparative covers *madhhab*-aware outputs, explicit surfacing of *ikhtilaf*, and non-flattening presentation. The design intent is to avoid a single "one true answer" when legitimate plurality exists, and to present alternative views with their supporting evidence.

(2) Uncertainty includes abstention/deferral, safe non-answers, and clarification prompts. This category operationalizes risk sensitivity by ensuring that when inputs are underspecified or grounding is unreliable, the system defers, requests missing context, or directs users to qualified authority.

Evaluation and Governance.

This **category** formalizes how systems are tested and controlled.

(1) Benchmarks covers evaluation protocols, error taxonomies, and reproducibility practices. Benchmark design should reflect normative constraints (e.g., penalizing fabricated citations more than generic factual errors) and, where relevant, support disagreement-aware scoring.

(2) Safety and Integrity covers hallucination controls, fabrication signals, bias checks, and related integrity mechanisms. This includes both static evaluations (test suites) and dynamic red-teaming style assessments tailored to religious harms (misattribution, offensive stereotyping, over-confident legal claims).

Multimodal.

This **category** covers non-text modalities and cross-modal alignment.

(1) Speech/Audio includes ASR for recitation, *tajwid* support, audio-text alignment, and dataset creation for Quranic recitation and related speech tasks. These systems are highly error-sensitive, since small phonetic differences can change meaning.

(2) Document Image Processing includes image enhancement, segmentation, and page understanding for heritage materials. It supports digitization at scale and enables later NLP/IR steps by producing structured representations from complex manuscript layouts.

Across both epistemic and task layers, we treat *normative dimensions* (doctrinal integrity/authenticity, disagreement handling, objectives/harms/governance) as cross-cutting constraints that shape the acceptable implementations and evaluation criteria for each subcategory.

3.2.3 Mapping Epistemic Domains to AI Task Families

In addition to the multi-layered taxonomies, we introduce a connecting taxonomy (Figure 6) that explicitly maps epistemic domains to the instrumental AI task families represented in the literature. This mapping is not a third independent classification; rather, it is a compact bridge that makes the many-to-many relationship between domains and methods legible for cross-comparison. Concretely, each epistemic **category** (e.g., revelation, Turath, fiqh, hadith sciences) is paired with the AI task families that are used to operationalize work in that domain (e.g., search/IR, RAG and evidence, extraction, summarization, speech/audio, benchmarks, safety and integrity). The result functions as an index for synthesis. It allows us to compare, across domains, which capabilities are emphasized (e.g., retrieval and grounding for scripture and fiqh), which are under-developed (e.g., multimodal and digitization for heritage), and where trustworthiness mechanisms recur (benchmarking, provenance, and safety/integrity).

This bridge is used throughout the survey to align RQ1's domain coding with RQ2's resources/benchmarks and RQ3's evaluation and governance concerns.

3.3 Normative Dimensions and Interpretive Diversity

Although our figures foreground epistemic categories and instrumental methods for clarity, normative concerns are central to our synthesis. We interpret results through three cross-cutting normative *dimensions*: **(i) doctrinal integrity and authenticity** (e.g., correct attribution and protection against fabricated or misquoted sources), **(ii) normative correctness and disagreement handling** (e.g., school-aware framing, *ikhtilaf*/*madhahib*-aware presentation, and calibrated abstention/deferral under uncertainty), and **(iii) objectives, harms, and governance** (*maqasid*-informed considerations of alignment, bias, safety, and the risks of flattening interpretive diversity). These dimensions are not mutually exclusive, and a single system (e.g., a multi-source QA assistant) may implicate multiple dimensions simultaneously.

3.4 Operationalizing the Taxonomy

This taxonomy acts as the survey's structural backbone, determining what we extract from the literature: *epistemic taxonomy* (Figure 4), *AI task family and focus area* (Figure 5), and the *domain-to-task mapping* (Figure 6) that links epistemic **categories** to the instrumental task families used to operationalize them. This standardisation enables comparative analysis across ten years of work. We structure the subsequent discussion around three drivers of progress: *(i)* dataset curation, *(ii)* modelling methodologies, and *(iii)* evaluation frameworks.

4 Data and Resources

The evaluation and enhancement of models in the domain of Islamic knowledge rely heavily on high-quality curated datasets. Unlike general open-domain tasks, Islamic knowledge requires strict adherence to theological accuracy, source faithfulness, and nuanced cultural understanding. Recent contributions in 2024 and 2025 have shifted from simple unstructured corpora to complex reasoning benchmarks and hallucination detection benchmarks. We categorize these resources into five primary streams: *(i)* Classical Pre-training Corpora, *(ii)* Structured knowledge bases and ontologies, *(iii)* Quranic and Hadith resources, *(iv)* Jurisprudence and reasoning benchmarks, and *(v)* Ethical and cultural benchmarks. Table 2 summarizes the key datasets surveyed.

4.1 Classical Pre-training Corpora

Training models to understand classical Arabic (*Fusha*) requires specialized corpora distinct from modern web text. The *OpenITI* corpus (Sibaee et al. 2025) is the gold standard for pre-modern texts, containing nearly 1 billion tokens of verified scholarly editions spanning law, theology, and history. Complementing this is the *Shamela Corpus* (Saleh and Al-Khalifa 2020), a cleaned version of the massive *Shamela* library, essential for Fiqh and Hadith commentary. For linguistic precision, the *KSUCCA* (Alrabiah et al. 2014) provides 50 million words strictly curated from the 7th–11th centuries

Dataset	Description	Type	Size	License	Lang	Ref.
Classical Pre-training Corpora						
OpenITI	The largest machine-readable corpus of pre-modern Islamicate texts.	Corpus	∼1B tokens	CC-BY-4.0	Ar/Per	Sibaee et al. (2025)
Shamela (Cleaned)	Text version of Shamela library; covers Fiqh, Tafsir, and History.	Corpus	∼1B Tokens	Public	Ar	Saleh and Al-Khalifa (2020)
KSUCCA	King Saud Univ. Corpus of Classical Arabic (7th–11th Century CE).	Corpus	50M Tokens	Research	Ar (Cls)	Alrabiah et al. (2014)
Tashkeela	Fully vocalized classical texts for training diacritic-aware models.	Corpus	75M Tokens	Public	Ar	Zerrouki and Balla (2017)
Noor Corpus	Massive diverse library of Islamic PDFs and OCR'd texts.	Corpus	>100k Books	Mixed	Ar	NoorSoft (2024)
Knowledge Bases & Ontologies						
QuranMorph Corpus	Morphologically annotated Quranic corpus with POS, Lemmatization, etc.	KB	77k Tokens	CC-BY-4.0	Ar/En	Akra et al. (2025)
Quranic Arabic Corpus	Morphological and syntactic ontology mapping concepts in the Quran.	KB/Onto	77k nodes	Research	Ar/En	Dukes and Habash (2010)
Quran Corpus POS	Annotated Quranic text for POS tagging.	Core	250 Sent.	Research	Ar	Huda et al. (2021)
IslamicPCQA KB	Curated KB of 1M+ Islamic documents for retrieval.	KB	1M+ Docs	Research	Fa/Ar	Asl and Bidgoli (2025)
Quran Ontology	graph linking verses to Hadith and Tafsir topics	KB	300+ Concepts	CC-BY	Ar/En	Abuz (2025)
Sunnah.com API	Structured Hadith collections with grading (Sahih/Hasan) metadata.	API/KB	6 Major Books	Open	Ar/En	Sunnah.com (2024)
Quranic & Hadith Resources						
LK Hadith Corpus	Ara-En parallel corpus of authen. Hadith	Corpus	39,038 Hadiths	Research	Ar/En	Altammami et al. (2020)
Qur'an QA 2022	Shared-task MRC dataset for Qur'anic QA	MRC/QA	1,093 QA	CC BY-NC-ND 4.0	Ar	Malhas et al. (2022)
Qur'an QA 2023	Shared task with Passage Retrieval	PR+MRC	251 QA [5]	CC BY-NC-ND 4.0	Ar	Malhas et al. (2023)
Sanadset 650K	Hadith dataset with narrator-chain (Sanad)	Corpus	650,986 Records	CC-BY-4.0	Ar	Mghari et al. (2022)
Quran_Hadith	Labeled related/non-related Quran–Quran and Quran–Hadith pairs.	Core	∼33.9k Records	Apache 2.0	Ar/En	Altammami (2023)
IslamicEval 2025	Hallucination detection-Quran/Hadith.	Hallu	1,506 Questions	Apache 2.0	Ar	Mubarak et al. (2025)
Iqra'Eval 2025	Quranic mispronunciation diagnosis.	Audio	82+ Hours	Research	Ar	El Kheir et al. (2025)
Quranic Audio	Crowdsourced non-Arabic recitations.	Audio	7k Clips	Apache 2.0	Multi	Raghad Salameh (2024)
TajweedAI	Audio dataset for Qalqalah detection.	Audio	Expert set	Research	Ar	Mazid and Ahmad (2025)
Quranic Surah RAG	Descriptive metadata for 114 Surahs.	RAG	114 Entries	CC-BY-NC	Ar/En	Khalila et al. (2025a)
Quran Semantic	Mapping verses to 30+ Tafsirs.	Search	30+ Tafsirs	Research	Ar	Yasser Shohoud (2023)
QURAN-MD	Unified verse-level text, translation, transliteration, and 32 reciters.	Multi	6,236 Verses	Research	Ar/En	Salman et al. (2025)
QTID	Image dataset, Quranic text for OCR & visual analysis	Vision	100k+ Images	Research	Ar	Badry et al. (2018)
Jurisprudence (Fiqh) & Reasoning						
Inheritance-Bench	Islamic Inheritance Law calculation.	Reason	1,000 MCQs	MIT	Ar/En	Bouchekif et al. (2025b)
QIAS 2025	Inheritance & general knowledge.	Reason	22,000 MCQs	Research	Ar	Bouchekif et al. (2025a)
FiqhQA	QA by 4 Sunni schools; abstention eval.	QA	960 QAs	Research	Ar/En	Atif et al. (2025)
Hajj-FQA	Specialized QA for Hajj rituals/fatwas.	QA	886 Hajj-fatwas	Research	Ar	Aleid and Azmi (2025)
IslamicPCQA	QA with 1M+ doc knowledge base.	QA/RAG	12k Pairs	Research	Fa	Asl and Bidgoli (2025)
MizanQA	Moroccan Islamic family law QA.	QA	1,700 Qs	Research	Ar	Bahaj and Ghogho (2025)
Islamic-QA	Egyptian Arabic QA pairs with topical source field.	QA	7.47k QAs	Apache 2.0	Ar	Youssef (2025)
Cultural & Ethical Alignment						
PalmX 2025	General Arabic & Islamic Culture.	Culture	6.4k MCQs	Shared Task	Ar	Alwajih et al. (2025b)
BengaliMoralBench	Moral reasoning-Bengali Islamic culture	Ethics	3k Scenarios	CC-BY-NC-ND	Bn	Ridoy et al. (2025)
ADMD	Arabic Depth Mini Dataset (Islamic Qs).	Eval	490 Qs	Public	Ar	Sibaee et al. (2025)
ADAB	App reviews corpus annotated for politeness and religious etiquette.	Style	Curated	Research	Ar	Faruk et al. (2025)
IslamTrust	Alignment benchmark with consensus-based Islamic ethical principles.	Ethics	Multi	Research	Ar/En	Lahmar et al. (2025)
IslamicFaithQA	Bilingual (Ar/En) generative benchmark.	Hallu	3,810	Research	Ar/En	Bhatia et al. (2026)

Table 2: Overview of resources related to Islamic AI systems, categorized by classical pre-training corpora, knowledge bases for retrieval, and benchmarks. **QA**=Question Answering, **Reason**=Reasoning/Math, **Hallu**=Hallucination Detection, **Audio**=Speech/Recitation, **KB**=Knowledge Base.

CE, while *Tashkeela* (Zerrouki and Balla 2017) offers 75 million fully vocalized words for training diacritic-aware models.

4.2 Knowledge Bases and Ontologies

To mitigate hallucination, models increasingly rely on structured knowledge bases (KBs) rather than parametric memory. The *QuranMorph* dataset (Akra et al. 2025) provides morphological annotations for every word in the Quran and links it with a lemma in Qabas (Jarrar and Hammouda 2024), which is a lexicographic data graph linking 2 million tokens corpora and 110 Arabic lexicons including the Arabic Ontology (Jarrar 2021). The *Quranic Arabic Corpus* (Dukes and Habash 2010) provides a node-based ontology mapping syntactic dependencies and semantic concepts within the Quran. Similarly, Asl and Bidgoli (2025); Mohammed et al. (2025b) constructed a retrieval index of over 1 million Islamic documents to support the *IslamicPCQA* system. For Prophetic traditions, structured datasets derived from *Sunnah.com* (Sunnah.com 2024) provide parallel Arabic-English Hadith with metadata on authenticity (*Sahih*, *Hasan*), enabling RAG systems to filter evidence based on theological strength.

4.3 Quranic and Hadith Resources

Foundational resources focus on the Holy Quran and Prophetic traditions. Textual resources include Yasser Shohoud (2023)'s semantic search dataset mapping verses to 30+ exegesis (Tafsirs), and Khalila et al. (2025a)'s descriptive dataset of the 114 Surahs for RAG. The *IslamicEval 2025* task (Mubarak et al. 2025) provided benchmarks for detecting fabricated or misquoted verses and Hadiths in LLM outputs. Other textual resources include the benchmark introduced by Aljaji et al. (2025) which stratifies questions by difficulty and evaluates source identification. In the audio domain, the *Iqra'Eval 2025* shared task (El Kheir et al. 2025) provides 82+ hours of data for Mispronunciation Detection and Diagnosis (MDD). Also in the audio domain, *TajweedAI* (Mazid and Ahmad 2025) provides expert-annotated data for detecting Qalqalah (echoing) errors, complementing the *Iqra'Eval* shared task. In the image domain, Salman et al. (2025) introduced *Quran-MD*, a unified multimodal resource aligning verses with translations, transliterations, and audio from 32 distinct reciters. For visual processing, *QTID* (Badry et al. 2018) offers over 100,000 images of Quranic text to advance OCR for religious scripts. In addition, Mhnaa et al. (2025) proposed datasets for recognizing Islamic religious signs in Arabic Sign Language (ArSL).

4.4 Jurisprudence and Reasoning Benchmarks

Evaluating jurisprudence capabilities has moved beyond memorized knowledge to reasoning tasks such as determining inheritance shares. Bouchekif et al. (2025b) introduced *Inheritance-Bench*, containing 1,000 MCQs requiring mathematical logic to determine inheritance shares, which *QIAS 2025* (Bouchekif et al. 2025a) expands to 22,000 questions covering general Islamic knowledge. To address doctrinal diversity, *FiqhQA* (Atif et al. 2025) categorizes questions by the four Sunni schools and tests for "abstention", the model's ability to refuse answering when it lacks sufficient knowledge.

4.5 Ethical and Cultural Benchmarks

Recent work emphasizes aligning LLMs with the moral nuances of the Islamic world. *PalmX 2025* (Alwajih et al. 2025b) benchmarks general Islamic culture across 22 Arab countries, while *Palm* instruction dataset (Alwajih et al. 2025a) ensures linguistic inclusivity of dialects. For ethical reasoning, Ridoy et al. (2025) proposed *BengaliMoralBench*, assessing alignment with religious activities and family norms in Bengali Islamic culture. *IslamTrust* (Lahmar et al. 2025) evaluates models against consensus-based Islamic ethical principles, revealing significant gaps in current multilingual LLMs. For generative style, the *ADAB* framework (Faruk et al. 2025) utilizes a dataset of app reviews annotated for religious etiquette (*Adab*) to train polite response generators. Additionally, Modarres Kamaly (2025) highlights the need for inclusion by benchmarking low-resource Indo-Iranian languages spoken by some Muslim communities.

4.6 Summary: Choosing the Dataset

We survey **37** resources grouped into five complementary streams that support end-to-end Islamic AI development:

- **Classical Pretraining Corpora** (**5** datasets) provide large-scale *Turath* text for continued pretraining and domain adaptation to classical Arabic and scholarly genres.
- **Knowledge Bases & Ontologies** (**6** datasets) offer structured representations (e.g., morphology, concepts, verse–topic graphs, hadith metadata) that enable RAG, provenance tracking, and faithfulness constraints.
- **Quranic & Hadith Resources** (**13** datasets) span text, audio, and vision, supporting tafsir-linked search, verse description retrieval, hallucination/fabrication detection, recitation assessment (MDD/*tajwıd* diagnosis), and OCR/visual understanding of scripts.
- **Jurisprudence (Fiqh) & Reasoning Benchmarks** (**7** datasets) test legal and mathematical reasoning (e.g., inheritance), specialized ritual QA (e.g., Hajj), school-aware sensitivity, and calibrated abstention under insufficient evidence.
- **Ethical & Cultural Benchmarks** (**6** datasets) target culturally grounded evaluation and alignment, including Islamic cultural knowledge, moral reasoning, etiquette-preserving generation (*adab*), and principle-based trustworthiness.

Across streams, these resources support a pipeline from **domain pretraining** and **grounded retrieval** to **robust evaluation** of factuality, reasoning, and alignment in Islamic settings.

5 Methods

Methods for operationalizing Islamic knowledge in AI have evolved through three overlapping eras: *(i)* classical IR/ML pipelines (BM25-style retrieval, SVM/CRF classifiers, ontologies), *(ii)* neural encoder systems culminating in BERT-style Arabic models, and *(iii)* modern LLM-based systems that pair instruction-tuned generators

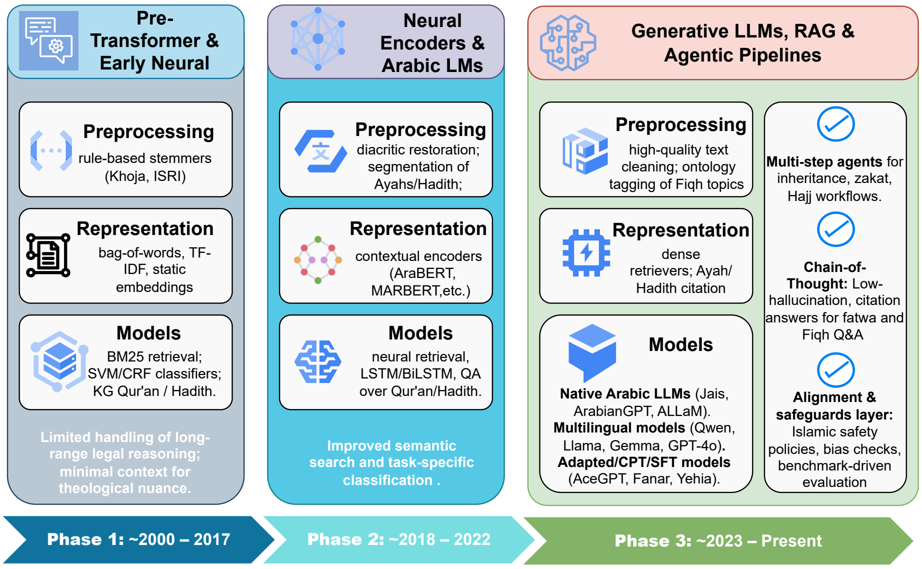

Pre-Transformer & Early Neural
Preprocessing
rule-based stemmers (Khoja, ISRI)
Representation
bag-of-words, TF-IDF, static embeddings
Models
BM25 retrieval; SVM/CRF classifiers; KG Qur'an / Hadith.
Limited handling of long-range legal reasoning; minimal context for theological nuance.
Neural Encoders & Arabic LMs
Preprocessing
diacritic restoration; segmentation of Ayahs/Hadith;
Representation
contextual encoders (AraBERT, MARBERT,etc.)
Models
neural retrieval, LSTM/BiLSTM, QA over Qur'an/Hadith.
Improved semantic search and task-specific classification .
Generative LLMs, RAG & Agentic Pipelines
Preprocessing
high-quality text cleaning; ontology tagging of Fiqh topics
Representation
dense retrievers; Ayah/ Hadith citation
Models
Native Arabic LLMs (Jais, ArabianGPT, ALLaM).
Multilingual models (Qwen, Llama, Gemma, GPT-4o).
Adapted/CPT/SFT models (AceGPT, Fanar, Yehia).
Multi-step agents for inheritance, zakat, Hajj workflows.
Chain-of-Thought: Low-hallucination, citation answers for fatwa and Fiqh Q&A
Alignment & safeguards layer: Islamic safety policies, bias checks, benchmark-driven evaluation
Phase 1: ~2000 – 2017
Phase 2: ~2018 – 2022
Phase 3: ~2023 – Present

Fig. 7: Evolution of AI methods for Islamic knowledge. The timeline summarizes three method stacks in parallel: *(i)* pre-Transformer IR/ML pipelines, *(ii)* neural encoders and Arabic language models, and *(iii)* generative LLM, RAG, and agentic pipelines, highlighting how each era tackles different problems in preprocessing, representation, and modeling, with earlier lexical and symbolic techniques being subsumed and reused within modern retrieval and alignment components.

with RAG, tools, and alignment safeguards. Figure 7 summarizes these method stacks in parallel, highlighting how each era tackles different problems. The LLM era does not replace earlier techniques so much as subsume them, for example, lexical retrieval and symbolic resources reappear inside RAG and verification pipelines. Guided by this map, Sections 5.1–5.5 review in turn: pre-transformer and early neural approaches, foundation-model paradigms, RAG approaches, agentic pipelines, and alignment and bias-mitigation strategies.

5.1 Pre-Transformer & Early Neural

Early Quranic NLP relied on rule-based stemmers (e.g., Khoja, ISRI) and static embeddings like AraVec (Soliman et al. 2017), which enabled semantic search but lacked context for Uthmani orthography and theological nuances (Bashir et al. 2021). Subsequent experiments with LSTMs improved diacritic restoration (Kanaan et al. 2013), yet remained constrained by sequential processing limits when modeling long-range dependencies in complex legal texts (Bashir et al. 2021).

5.2 Generative LLMs & Foundation Models

The transition to Transformer-based architectures established three distinct paradigms for Arabic LLMs (Al-Khalifa et al. 2025; Mashaabi et al. 2024; Rhel and Roussinov 2025):

- **Native Models:** Trained from scratch on Arabic-centric data for high linguistic fidelity, including Jais (Sengupta et al. 2023), ArabianGPT (Koubaa et al. 2024), and ALLaM (Bari et al. 2025).
- **Multilingual Foundations:** Models where Arabic is a component of a massive mix, including open-weights models like Qwen3 (Yang et al. 2025), Llama 3 (Grattafiori et al. 2024), and Gemma 3 (Team et al. 2025b), alongside proprietary models like GPT-4 (OpenAI et al. 2024).
- **Adapted Models:** Systems employing continued pre-training (CPT) or supervised fine-tuning (SFT) on multilingual checkpoints to enhance Arabic proficiency. Notable examples include AceGPT (Huang et al. 2024), SILMA (Team 2024), and Fanar (Team et al. 2025a).
- **Multimodal Models:** Multimodal methods remain less mature than text-centric stacks, but are emerging along two practical directions: ***(i)*** **audio-language pipelines** for recitation assessment, where ASR-based alignment is paired with targeted classifiers for tajwīd and mispronunciation diagnosis (e.g., Qalqalah detection) (El Kheir et al. 2025; Mazid and Ahmad 2025; Salameh et al. 2024); and ***(ii)*** **vision-language pipelines** for religious text and symbols, combining OCR and visual encoders with Arabic LMs to support Quran-script recognition, verse-level multimodal alignment, and accessibility-oriented recognition (e.g., Islamic sign understanding) (Salman et al. 2025; Mhnaa et al. 2025). In parallel, recent studies evaluate and adapt text-to-image and calligraphy-focused systems for heritage fidelity and legibility, indicating a shift from purely aesthetic generation toward grounded multimodal understanding (Ashour and Rashdan 2025; Gürer et al. 2025; Sumayli and Alkaoud 2025; Elsharif et al. 2025).

5.3 LLM and RAG

RAG has emerged as the standard architecture to address the zero-tolerance policy for hallucination.

Architectures & Faithfulness. Pipelines typically couple generative readers with dense retrievers (e.g., MARBERT) to cite specific Ayahs and filter weak Hadith, significantly outperforming multilingual baselines (Ben Ayed et al. 2025; Khalila et al. 2025a; Bhatia et al. 2025). Systems like RFPG utilize hybrid search to capture exact legal terminology (Bahaj and Ghogho 2025). While multi-stage retrieval enhances performance (Pavlova 2025), recent findings suggest increased reasoning time does not automatically reduce hallucinations, necessitating strict policies verified against benchmarks (Zhao et al. 2025; Gema et al. 2025; Alansari and Luqman 2025).

5.4 Agentic Pipelines

For complex tasks like inheritance share calculation or Hajj rituals, single-pass generation is insufficient. Agentic systems (e.g., PuxAI, Hajj-FQA) decompose problems into sub-tasks to validate evidence and execute symbolic calculations (Phuc and Thin 2025; Aleid and Azmi 2025; Asl and Bidgoli 2025). Chain-of-Thought ensembles further manage uncertainty by recursively re-querying knowledge bases (Alangari and Team 2025; Alshaikh et al. 2025a). Beyond QA, agents are now used for simulation. Omarov et al. (2025) introduced an Agent-Based Modelling (ABM) framework in which LLM-driven agents simulate Zakat policy dynamics, modelling the behaviour of payers, beneficiaries, and regulators under Sharia constraints. In educational tutoring settings, agentic control can also be used to constrain generation: Hasan (2025) propose sparse-checklist prompting, where the model selects pedagogical hints from a fixed inventory instead of generating unconstrained explanations, improving accuracy while reducing tokens.

5.5 Alignment & Bias Mitigation Strategies

Beyond retrieval, specific strategies are employed to ensure theological fidelity and reduce bias as described below.

Prompt Engineering.

Prompt-level controls can reduce harmful stereotypes and shape style, but they are fragile against deep representational bias and adversarial prompting. Instructing models to reflect on bias serves as a first line of defense, reducing anti-Muslim sentiment and steering the "moral compass" (Asseri et al. 2025; Li et al. 2025). These techniques are validated against cultural benchmarks like PalmX and CamelEval (Alwajih et al. 2025a; Qian et al. 2024), though they are limited against deep representational biases (Seth et al. 2025; Magdy et al. 2025; Sadallah et al. 2025). Newer benchmarks like *IslamTrust* (Lahmar et al. 2025) utilize rigorous guideline-based prompting to expose deep misalignments in how models interpret Islamic ethics.

Domain-Specific Adaptation.

Deep alignment is achieved by training on religious corpora and fine-tuning on vetted QA pairs (Patel et al. 2023). Projects like NileChat and Swan demonstrate that dialect-aware specialization outperforms generalist models (Mekki et al. 2025; Bhatia et al. 2024), respecting intra-Muslim diversity (Al-Monef et al. 2025).

Interpretability & Governance.

Interpretability and governance approaches treat safety as an engineering problem. Model audits, mechanistic analyses, and safety indices translate "alignment" into measurable constraints that can be monitored over time. Mechanistic analyses (e.g., sparse autoencoders) can surface pathways associated with bias or harmful associations (Prakash and Roy 2024; Simbeck and Mahran 2025). Complementing this, safety governance relies on indices such as ASAS and AraTrust, now integrated into leaderboards like OALL 2.0 (Alghamdi et al. 2025; El Filali et al. 2025).

5.6 System Design Patterns

The reviewed literature suggests that high-performing Islamic-knowledge systems are rarely a single technique. They are typically *stacks* that combine orchestration, retrieval grounding (RAG), and attribution (Alghamdi et al. 2025). To translate the method landscape in earlier subsections into actionable guidance, we summarize recurring system design patterns observed across Quran/Hadith QA, fatwa support, and jurisprudential reasoning. Each pattern is described in terms of *(i)* the failure mode it targets, *(ii)* the architectural components it introduces, and *(iii)* evaluation signals that can provide evidence it is working. In addition, we illustrate them through case studies that method choices (Section 5) and evaluation objectives in Section 6.

5.6.1 Design Patterns for Reliable Islamic AI Systems

Across the reviewed systems, we repeatedly observe patterns aimed at deployment-critical failure modes, including fabrication, misattribution, disagreement collapse, and overconfident answering under ambiguity. In Table 3, we briefly summarize them.

Evidence-first RAG with a citation policy. Many systems treat retrieved Quran/Hadith/tafsīr evidence as the primary source of truth, rather than relying on parametric memory. This pattern enforces claim-level support from retrieved spans and requires citations in a fixed, machine-checkable format. Supporting evidence for effectiveness can include citation precision/recall and entailment/faithfulness checks against cited spans.

Authority-aware retrieval and filtering. Reported systems often degrade when retrieval draws from mixed-quality sources (e.g., web pages, forums, weak reports). A common response is to constrain retrieval via source whitelists, collection metadata, and Hadith authenticity/grade filters. One can then examine changes in false attributions alongside coverage–recall trade-offs.

Query decomposition and multi-hop retrieval for composite questions. For queries that are composed of multiple premises (e.g., ritual rules plus contextual constraints), many systems decompose the question before answering. They retrieve evidence per sub-claim and synthesize only after coverage is established. Evidence for reliability can be reported via per-claim faithfulness and end-to-end coherence.

Verification layer (entailment + citation verification). In high-stakes settings, retrieval alone often leaves residual hallucinations. Several deployments therefore add a post-generation verification layer that checks claim entailment against retrieved evidence and citation-to-span correctness. This pattern can be probed using span-level fabrication benchmarks and targeted adversarial prompts.

Calibrated abstention/deferral as a first-class capability. We observe frequent cases where queries are underspecified, evidence is missing, or disagreement is central. A robust response is to define explicit abstention triggers (e.g., low retrieval confidence, conflicting evidence, missing parameters, contested rulings) and treat refusal/deferral as a first-class outcome. One can report abstention quality and calibration, rather than penalizing refusal by default.

Tool-augmented deterministic reasoning for rule-based subproblems. Several systems route executable subproblems (e.g., inheritance shares, finance constraints, prayer-time calculations) to verified solvers, then generate explanations grounded in solver outputs. This reduces brittle free-form reasoning while preserving interpretability. Evidence for correctness can include execution accuracy and consistency between solver outputs and the generated narrative.

Design pattern	**Primary goal / failure mode**	**Evaluation**
P1. Evidence-first RAG + citation contract	Prevent fabricated or misattributed scripture/hadith; enforce provenance	**Retrieval & Grounding:** Citation Precision/Recall; Entailment Faithfulness; Span-level error / factuality checking. **Doctrinal & Cultural:** Scholar-in-the-loop adjudication for "fatwa-grade" outputs.
P2. Authority-aware retrieval & filtering	Reduce low-quality/weak sources and context collapse in retrieval	**Retrieval & Grounding:** Citation Precision/Recall; Entailment Faithfulness. **Doctrinal & Cultural:** Scholar rubric checks (source appropriateness, correct attribution).
P3. Query decomposition + multi-hop retrieval	Handle composite questions; reduce context drift across sub-claims	**Retrieval & Grounding:** Per-claim entailment/faithfulness; span-level factuality errors. **Capabilities:** Standardized testing (when framed as MCQ sub-questions).
P4. Verification layer (citation + entailment)	Detect/repair hallucinations and unsupported claims post-generation	**Retrieval & Grounding:** Span-level factuality; Citation verification; Entailment faithfulness.
P5. Calibrated abstention / deferral (first-class)	Avoid overconfident answers under missing evidence, ambiguity, or contested issues	**Doctrinal & Cultural:** Scholar-in-the-loop rubric includes appropriateness of uncertainty/deferral. **Reasoning & Calibration:** Penalize overconfidence; reward abstention when warranted (report calibration/abstention rates).
P6. Tool-augmented deterministic reasoning	Correctness on rule-based tasks (inheritance, calculations); avoid arithmetic/logic errors	**Capabilities:** Symbolic verification (execution accuracy against rule engines); Standardized testing (MCQ accuracy).
P7. Pluralism-aware answering (school-conditioned / multi-position)	Prevent disagreement collapse; preserve legitimate plurality	**Doctrinal & Cultural:** Scholar-in-the-loop adjudication with explicit rubric for pluralistic framing (qualify by school, show positions with evidence).
P8. Safety + etiquette guardrails (adab)	Prevent toxic/stigmatizing content; maintain culturally grounded safe behavior	**Safety & Governance:** Red-teaming attack success rate; Mechanistic analysis (latent bias/violence associations). **Doctrinal & Cultural:** Cultural probing alignment score; Dialectal robustness where relevant.

Table 3: Design patterns for Islamic-knowledge AI systems and their evaluation hooks, aligned to the metric families summarized in Table 4 (capabilities; retrieval/grounding; doctrinal/cultural alignment; safety/governance).

Pluralism-aware answer formatting (school-conditioned or multi-position outputs). In *fiqh*-oriented settings, legitimate disagreement is expected. Systems therefore present multiple positions with their evidences, or condition answers on a requested *madhhab*. Suitable evaluation approaches include multi-reference or rubric-based protocols that avoid rewarding "disagreement collapse."

Safety and style guardrails aligned with domain norms. Community-facing assistants frequently incorporate guardrails for hate/stigma content, respectful tone (*adab*), and risk-tiered responses. These guardrails shape both *what is said* and *how it is said* under uncertainty or risk. A practical validation approach combines targeted red-teaming with culturally grounded safety benchmarks.

These patterns provide a compact "reference approaches" for mapping methods to deployment requirements and for specifying what should be measured when a system claims to be "reliable" in Islamic-knowledge settings.

5.7 Current Architectures in Islamic-Knowledge Systems

Across Quran/Hadith QA, fatwa support, and jurisprudential assistants, the dominant engineering trend is toward *stacked* systems: retrieval grounding, provenance constraints, and post-hoc validation are composed under an orchestration layer rather than treated as a single prompting trick. In this section we summarize three architecture families that recur across recent Islamic QA datasets, shared tasks, and system papers, then instantiate them in two deployment-relevant case studies.

Curated retrieval-grounded QA with explicit attribution.

A common baseline is retrieval-augmented generation over curated corpora (e.g., Quran with tafsīr links, authenticated hadith collections, vetted fatwa repositories), coupled with an explicit requirement that answers cite the retrieved evidence in a consistent format. This style is reported in Quran/tafsīr QA proposals and empirical studies of RAG for Quranic QA, where grounding is used to reduce hallucinations and to make outputs auditable (Khalila et al. 2025b). In fatwa-oriented resources, similar pipelines rely on domain-specific corpora and explicit provenance discipline to maintain source appropriateness (Mohammed et al. 2025a; Aleid and Azmi 2025).

Multi-stage pipelines that verify and/or correct religious attributions.

A second family makes validation a first-class stage. Systems detect whether a claimed quotation or attribution is Quranic or hadith, verify it against authoritative references, and then either correct the text or refuse when verification fails. This architecture is strongly represented in the IslamicEval shared task, which operationalises hallucination detection and correction for Islamic content, and in participating system descriptions that implement explicit detection–verification–correction workflows (Mubarak et al. 2025). These pipelines shift evaluation away from end-answer plausibility toward measurable verification outcomes (e.g., quotation-match accuracy, correction success, and refusal appropriateness).

Agentic retrieval as iterative evidence seeking and answer revision.

More recent work treats retrieval as an iterative decision process. The model repeatedly plans, calls domain tools, inspects retrieved passages, and revises until evidence coverage is sufficient or deferral is warranted. Bhatia et al. (2026) frames this as a shift from conventional RAG to *agentic RAG* for faithful Islamic question answering, emphasizing tool-mediated inspection and iterative refinement rather than single-shot retrieval. While their focus is Quran-grounded QA, the architectural principle generalizes to broader Islamic-knowledge deployments that require multi-hop evidence gathering and explicit stopping criteria.

Tool-augmented deterministic reasoning for rule-executive subproblems. Certain Islamic tasks are primarily computational rather than textual. Islamic inheritance is a canonical example where correct outputs require deterministic rule execution over structured heir sets and exclusions. The QIAS shared task institutionalizes this as a benchmark target and shows how systems combine LLM understanding with constrained reasoning components (Bouchekif et al. 2025a). Empirical analyses further demonstrate that fluent LLM outputs can be systematically incorrect on inheritance even when responses appear persuasive (Bouchekif et al. 2025b). Competitive approaches therefore often combine retrieval support and targeted adaptation with structured execution (AL-Smadi 2025).

5.8 Case studies

5.8.1 Fatwa-Style QA Pipeline

Fatwa-like questions are a prototypical high-stakes Islamic-knowledge setting in which users may act on outputs, and correctness is evidence- and context-dependent. The literature therefore tends to position the LLM primarily as an *controller and explainer*, with reliability arising from curated retrieval, provenance discipline, verification, and calibrated deferral (Mohammed et al. 2025a; Aleid and Azmi 2025). Figure 8 visualizes these stages end-to-end, highlighting the single generative module and the surrounding safety and verification gates.

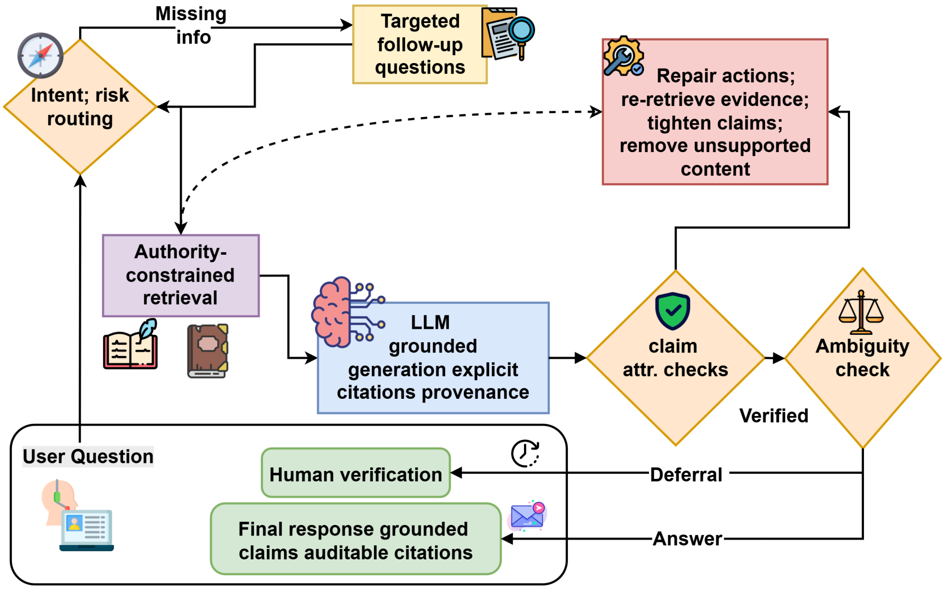

Missing info
Targeted follow-up questions
Intent; risk routing
Repair actions; re-retrieve evidence; tighten claims; remove unsupported content
Authority-constrained retrieval
LLM grounded generation explicit citations provenance
claim attr. checks
Ambiguity check
Verified
User Question
Human verification
Deferral
Final response grounded claims auditable citations
Answer

Fig. 8: Fatwa-style question answering pipeline with authority-constrained retrieval, grounded generation, verification/repair, and calibrated deferral. The generative component is explicitly isolated as a single module within the overall control flow.

Pipeline.

1. **Intent and risk routing; parameter elicitation.** The system first classifies the query type (informational vs. normative; low- vs. high-impact) and checks whether ruling-changing parameters are missing. If underspecified, it asks brief follow-ups rather than guessing.
2. **Authority-constrained retrieval over curated corpora.** Retrieval is restricted to trusted collections (Quran/tafsīr, authenticated hadith with grading metadata, vetted fatwa repositories). This reduces contamination from mixed-quality sources and supports auditable grounding (Mohammed et al. 2025a; Aleid and Azmi 2025).
3. **Evidence-grounded drafting with explicit citations.** The system generates a structured answer where non-trivial claims are explicitly supported by retrieved spans and accompanied by consistent citations, separating *(i)* directly evidenced statements and *(ii)* statements contingent on missing user context.
4. **Verification and repair.** A verifier checks that citations actually support the associated claims (e.g., quotation correctness, attribution validity, and claim-to-evidence consistency). If verification fails, the system revises by tightening claims, re-retrieving, or removing unsupported content. IslamicEval-style pipelines provide a concrete template for operationalizing quotation verification and correction (Mubarak et al. 2025).
5. **Deferral under ambiguity or contested evidence.** If retrieval is empty/low-confidence, evidence conflicts remain unresolved, or critical parameters are missing, the system defers (rather than producing a confident ruling) and recommends consultation with qualified scholarship for fatwa-grade decisions.

Evaluation signals.

Typical evaluations report retrieval relevance and citation correctness, claim-to-evidence consistency, rates of false attribution, and deferral calibration on underspecified or adversarial prompts. Shared-task protocols (e.g., IslamicEval's detection/correction setup) make these objectives measurable and comparable across systems (Mubarak et al. 2025).

5.8.2 Inheritance Reasoning Pipeline

Islamic inheritance illustrates a different reliability regime. the dominant failure mode is incorrect rule execution, not merely missing evidence. Benchmarks and shared tasks show that general-purpose LLMs can be fluent yet wrong, and often fail to request missing heir information (Bouchekif et al. 2025b,a). A robust system therefore separates language understanding and explanation from deterministic computation. Figure 9 depicts this separation explicitly, with the solver producing a trace that the generative module verbalizes and a checker enforcing exact agreement.

Pipeline.

1. **Structured intake and completeness checks.** The system extracts a structured heir schema (spouse(s), descendants, ascendants, siblings, exclusions) and

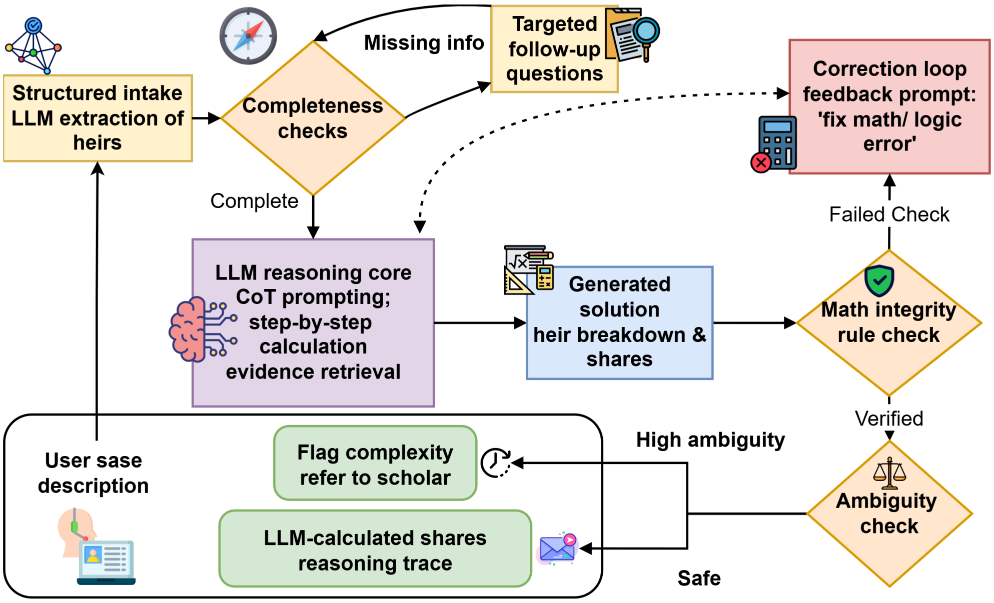


Fig. 9: Inheritance reasoning pipeline that separates structured intake and completeness checks from deterministic share computation, followed by trace-grounded explanation, consistency validation, and deferral for exceptional cases. The generative component is confined to explanation rather than arithmetic.

checks for missing discriminators (e.g., number of children and sons vs. daughters; surviving parents; debts and bequests). If incomplete, it asks targeted follow-ups.

2. **Deterministic execution via a verified solver.** The structured case is routed to a rule engine that computes shares, residuary allocations, and exclusion logic deterministically.
3. **Grounded explanation from solver trace.** The system generates an explanation by narrating the solver outputs rather than re-deriving arithmetic free-form. It may optionally retrieve authoritative references for transparency, but the numerical computation remains externalized.
4. **Consistency validation.** A checker verifies that the natural-language explanation matches solver outputs exactly (heir set, fractions, and totals). Any discrepancy triggers regeneration of the explanation, not re-computation.
5. **Deferral for ambiguous or out-of-scope cases.** If parsing confidence is low, inputs conflict, or exceptional cases are outside the encoded ruleset, the system defers rather than improvising.

Evaluation signals.

Evaluation emphasizes solver accuracy against gold computations, robustness to paraphrase/noisy descriptions, explanation-to-solver consistency, and deferral behavior under underspecification. QIAS provides standardized inheritance-focused evaluation contexts and system reports illustrating practical combinations of adaptation and retrieval support (Bouchekif et al. 2025a; AL-Smadi 2025).

5.9 Choosing a Method Stack

Overall, the literature suggests a risk-tiered design principle. For low-stakes educational or exploratory queries, lightweight grounding and cultural alignment may suffice. As stakes increase (fatwa-like guidance, financial rulings, or sensitive social questions), systems should progressively add *(i)* stronger retrieval with explicit provenance, *(ii)* authority-aware filtering, *(iii)* tool-based reasoning for deterministic subproblems, and *(iv)* verification and abstention policies. This layered perspective sets up our evaluation discussion. Each added component reduces a class of failures; however, it also introduces new evaluation requirements.

Evaluation Category	**Methodology**	**Metrics & Description**	**Key Benchmarks**
Linguistic & Reasoning Capabilities	*N-gram Matching*	**BLEU, ROUGE, METEOR**: Measures lexical overlap with reference text. Limited utility for theological nuance.	OALL (El Filali et al. 2025), AraGen (El Filali et al. 2024b)
	Symbolic Verification	**Execution Accuracy**: Validates mathematical derivations (e.g., inheritance shares) against rule-based engines.	QIAS (Bouchekif et al. 2025a), GATMath (AlBallaa et al. 2025)
	Standardized Testing	**Normalized Accuracy**: Performance on multiple-choice questions (MCQs) across diverse subjects (STEM, Humanities).	ArabicMMLU (Koto et al. 2024), QuranBench (Aljaji et al. 2025)
Retrieval & Grounding (RAG)	*Factuality Checking*	**Span-Level Error Rate**: Percentage of generated text spans containing fabricated content.	Halwasa (Mubarak et al. 2024), HalluVerse (Abdaljalil et al. 2025)
	Citation Verification	**Citation Precision/Recall**: Accuracy of retrieving specific Qur'anic Ayahs or Hadith to support a claim.	IslamicEval (Mubarak et al. 2025), Hajj-FQA (Aleid and Azmi 2025)
	Entailment	**Faithfulness Score**: Measures if the generated answer is logically entailed by the retrieved context.	FARSIQA (Asl and Bidgoli 2025), AraHalluEval (Alansari and Luqman 2025), IslamicFaithQA (Bhatia et al. 2026)
Doctrinal & Cultural Alignment	*Scholar-in-the-Loop*	**Adjudication Score**: Human expert evaluation of *Fatwa* correctness, *Adab* (etiquette), and *Hikmah* (wisdom).	FiqhQA (Atif et al. 2025), Iqra'Eval (El Kheir et al. 2025) , ADAB (Faruk et al. 2025)
	Cultural Probing	**Alignment Score**: Degree of conformity to Arab-Islamic norms vs. Western-centric values.	IslamTrust (Lahmar et al. 2025), PalmX (Alwajih et al. 2025b)
	Dialectal Robustness	**Dialectal Accuracy**: Performance consistency across Modern Standard Arabic (MSA) and regional dialects.	AraDiCE (Mousi et al. 2025), Absher (Al-Monef et al. 2025)
Safety & Governance	*Red Teaming*	**Attack Success Rate (ASR)**: Vulnerability to prompt injection or generating prohibited content (e.g., hate speech).	ASAS (aiastrolabe 2025), AraTrust (Alghamdi et al. 2024)
	Mechanistic Analysis	**Latent Activation**: Identification of internal neuronal pathways associated with bias or violence.	Simbeck and Mahran (2025)

Table 4: Evaluation methodologies and metrics for Islamic AI systems. The table categorizes evaluation strategies into four primary domains: general capabilities, retrieval & grounding, doctrinal faithfulness, and safety & ethics..

6 Evaluation

6.1 Evaluation as Assurance

Evaluation in the Islamic domain cannot be reduced to generic NLP benchmarks. The core question is not fluency but whether a system is *safe and well-grounded.* It should ground claims in authoritative evidence, preserve provenance, and abstain when evidence is missing or contested. We therefore frame evaluation as layered assurance that mirrors the system stack in Section 5, building on broader work in ethics, bias, and safety assessment for LLMs. Leaderboards such as OALL (El Filali et al. 2024a, 2025) and frameworks like AraGen (El Filali et al. 2024b) and BALSAM (Al-Matham et al. 2025) provide useful Arabic capability baselines, however, religious settings require targeted tests of faithfulness, doctrinal correctness, calibrated uncertainty, and culturally grounded safety.

6.2 Automatic Metrics

Standard n-gram metrics (e.g., BLEU, ROUGE) are increasingly viewed as insufficient for evaluating the semantic validity of religious text. Consequently, current state-of-the-art approaches rely on LLM-as-a-Judge paradigms, where strong models (e.g., GPT-4) score generated outputs against gold references (Dubois et al. 2023; Habib et al. 2023). However, because Islamic-domain errors have *asymmetric costs*, where low-stakes issues (e.g., grammar or phrasing) differ from high-stakes failures (e.g., misattributed sources or incorrect legal/financial calculations such as inheritance shares), automatic metrics should be interpreted as preliminary signals rather than final certification.

6.3 Human Evaluation & Scholar-in-the-loop

Despite advances in automated metrics, human evaluation remains the gold standard for high-stakes religious AI. Scholar-in-the-loop methodologies (e.g., involving experts in Fiqh) are essential for validating "Fatwa-grade" outputs, including assessing *Adab* (respect and politeness) and *Hikmah* (wisdom), dimensions that automated scorers often miss. Because *Adab* is central yet potentially subjective, studies should specify an explicit reviewer rubric (e.g., respectful tone, avoidance of harshness, appropriate uncertainty/deferral, and non-presumptive guidance) so that judgments are reproducible. Crucially, human evaluation must be structured such as defining clear rubrics, report inter-annotator agreement, and specify whether reviewers assess a single school of thought, a consensus position, or a pluralistic framing. This ensures that human judgments are comparable across studies rather than purely anecdotal. Frameworks such as EAGLE (Kaneko et al. 2024) and LocalValueBench (Meadows et al. 2024) help formalise this process by collecting human judgments on open-ended interactions to assess genuine alignment with user intent and values.

6.4 Faithfulness & Provenance

In religious contexts, hallucination is not a trivial issue Mubarak et al. (2025); Bhatia et al. (2026). Fabricating verses or misattributing hadith can directly mislead users.

Faithfulness evaluation therefore emphasizes detection and verification. The Halwasa benchmark (Mubarak et al. 2024) provides the first dedicated framework for quantifying hallucinations in Arabic LLMs, revealing that even high-performing models frequently generate non-factual content. This effort is complemented by AraHalluEval (Alansari and Luqman 2025) and HalluVerse25 (Abdaljalil et al. 2025), which employ fine-grained span-level detection to identify subtle corruptions in text. For RAG-based systems, provenance is evaluated by checking if the generated response correctly entails the retrieved evidence. Studies on Quranic RAG emphasize the need for strict entailment scoring to ensure that theological claims are directly supported by the retrieved *Ayat* (Khalila et al. 2025a; Asl and Bidgoli 2025).

6.5 Doctrinal Faithfulness

Beyond citation accuracy, systems must be evaluated for doctrinal and legal correctness. Responses should align with established theological norms and jurisprudential practice. Benchmarks such as IslamicEval (Mubarak et al. 2025) and PalmX (Alwajih et al. 2025b) probe cultural and religious knowledge through multiple-choice questions derived from Islamic curricula, and show that general-purpose multilingual models often miss key Fiqh nuances compared with domain-tuned variants. Related proxies include legal reasoning benchmarks such as ArabLegalEval (Hijazi et al. 2024) and MizanQA (Bahaj and Ghogho 2025), which test rule application to concrete cases and parallel aspects of fatwa-style reasoning over Sharīah evidence. Task-specific resources such as Hajj-FQA further target ritual QA, where accuracy is non-negotiable (Aleid and Azmi 2025). Finally, a recurring failure mode is the collapse of scholarly disagreement into a single oversimplified answer. We frame this as a *multi-perspective alignment* problem, where evaluation should reward systems that *(i)* recognize legitimate plurality, *(ii)* qualify claims by school or interpretive tradition (e.g., "according to school X"). Addressing these challenges, evaluation protocols should distinguish between *(i)* objective errors (fabrication, misattribution, invalid reasoning) and *(ii)* acceptable plurality, where the system should either qualify its answer or present multiple positions with their evidence.

6.6 Reasoning Quality & Calibration

Many high-impact Islamic tasks are reasoning-heavy such as inheritance, prayer time computations, or analogy-based rulings. For these settings, evaluation should measure not only final answers but also whether intermediate steps follow valid derivations, and whether the system appropriately delegates deterministic subproblems to tools. The Qiyas benchmark (Al-Khalifa and Al-Khalifa 2024) and GATMath (AlBallaa et al. 2025) evaluate the mathematical and logical reasoning capabilities of Arabic LLMs, while AraTable (Alshaikh et al. 2025b) assesses tabular reasoning often found in classical texts. Agentic pipelines dealing with inheritance calculations are evaluated not just on the final answer, but on the step-by-step derivation of shares, often necessitating the decomposition of complex queries into sub-problems (Phuc and Thin 2025). Equally important is calibration in which models should be penalized for overconfident answers in ambiguous cases and rewarded for abstention when warranted. This turns "I don't know" (abstention) into a measurable capability rather than a perceived weakness.

6.7 Bias/Fairness & Safety

Safety in the Islamic domain encompasses both the prevention of toxic content and alignment with cultural values (*Haya'* and *Adab*). CamelEval (Qian et al. 2024) and PalmX (Alwajih et al. 2025b) explicitly measure the cultural alignment of LLMs, penalizing "westernized" responses that conflict with Arab-Islamic norms. Trustworthiness is quantified using benchmarks like AraTrust (Alghamdi et al. 2024) and the AI Astrolabe Safety Index (ASAS) (aiastrolabe 2025). Specific datasets for LLM safeguards (Ashraf et al. 2025) allow researchers to "red team" models against prompt injection or the generation of prohibited content. To ensure inclusivity across the Arab world, benchmarks like Absher (Saudi dialects) (Al-Monef et al. 2025) and AL-QASIDA (Robinson et al. 2025) assess performance on non-standard Arabic, mitigating bias against dialectal speakers (Keleg et al. 2025).

7 Findings, Critical Gaps and Future Directions

7.1 Findings

Provenance failure patterns in in-scripture QA. Across Quran/Hadith question answering, a recurring deployment-critical failure mode is *provenance breakage*. Models produce fluent answers whose claimed scriptural support is missing, misattributed, or non-entailed by the cited evidence. Dedicated hallucination and fabrication benchmarks show that even strong Arabic-capable LLMs can generate non-factual religious content, including subtle corruptions and fabricated passages (Abdaljalil et al. 2025). This motivates evaluation that goes beyond surface citation presence to *verification*-checking that generated claims are strictly supported (entailed) by retrieved verses/narrations and that citations are precise at the span level (Asl and Bidgoli 2025; Khalila et al. 2025a). In short, in-scripture QA reliability depends less on generic factuality and more on enforcing and measuring end-to-end grounding constraints.
Failure conditions in fiqh tasks. Fiqh-oriented tasks expose failure modes that retrieval alone does not address, including multi-step legal/mathematical reasoning (e.g., inheritance), delegation of deterministic subproblems to tools/solvers, and ambiguity handling via calibrated abstention. Recent benchmarks show that general-purpose models often fail on structured jurisprudential reasoning and do not abstain when queries are underspecified (Bouchekif et al. 2025b). This motivates risk-tiered architectures in which the LLM handles language and retrieval, while rule execution is verified externally (Al-Khalifa and Al-Khalifa 2024). Accordingly, Fiqh evaluation should jointly score correctness, tool-verified derivations, and refusal/deferral behavior.
Pluralism & school-aware disagreement is under-evaluated. A common doctrinal failure mode is *flattening*, where legitimate juristic disagreement (*ikhtilāf*) is reduced to a single unqualified answer. This reflects an evaluation gap, since most protocols do not reward systems that recognize plurality, qualify responses by *madhhab*/tradition, or present multiple supported positions (Atif et al. 2025). Although school-labeled resources make pluralism and abstention measurable, pluralism-aware evaluation remains the exception. Evaluation should therefore distinguish objective

failures (fabrication, misattribution, invalid reasoning) from acceptable plurality, where the appropriate behavior is qualified or multi-perspective answering (Atif et al. 2025).

Multimodality is under-evaluated end-to-end. While multimodal work is growing—spanning recitation assessment and mispronunciation diagnosis, *tajwīd*-focused audio classification, Quranic OCR, and verse-level multimodal alignment—evaluation is still largely component-level rather than end-to-end (El Kheir et al. 2025; Mazid and Ahmad 2025; Salman et al. 2025). In practice, user-facing systems increasingly require full pipelines (ASR/OCR → retrieval → grounded response → verification), yet benchmarks rarely test the compounded error surface across these stages. Advancing multimodal Islamic assistants therefore requires integrated evaluations that measure not only recognition accuracy, but also downstream grounding fidelity, citation correctness, and abstention when multimodal inputs are noisy or ambiguous (Alansari and Luqman 2025).

7.2 Current Challenges

Aligning LLMs with Islamic ethics involves navigating a unique set of challenges, including data scarcity, bias, and high-stakes hallucination risks.

Data scarcity and quality. The training corpora for state-of-the-art models remain disproportionately sparse regarding Arabic, Islamic studies, and Muslim lived experiences (Naous and Xu 2025; Rhel and Roussinov 2025). Arabic content is often dominated by web-scraped noise rather than curated scholarship (Sibaee et al. 2025). This scarcity is compounded by standard tokenization methods that often disadvantage morphologically rich languages, inflating costs and degrading performance (Sun et al. 2025).

Representational bias. Biases can be embedded in model parameters. Mechanistic analyses suggest that latent features linking "Islam" with *security-related or conflict-focused* language can be more sharply separated than those for other religions (Simbeck and Mahran 2025; Chandna et al. 2025). Empirical evaluations further show that models may default to a homogenized, Western-centric worldview (Shankar et al. 2025; Seth et al. 2025), which can yield disproportionate refusal rates or stigmatizing framing when prompted about Islamic topics.

Hallucination and evaluation. The linguistic complexity of Classical Arabic exacerbates hallucination (Zhao et al. 2025), leading to the dangerous fabrication of Qur'anic verses or Hadith (Khalila et al. 2025a). On structured legal tasks like inheritance, general-purpose models often achieve less than 50% accuracy and fail to abstain from ambiguous queries (Atif et al. 2025). Furthermore, existing ethical benchmarks (Hendrycks et al. 2020) largely reflect Western values; a unified framework auditing Islamic theology (*Aqīdah*) and jurisprudence (*Fiqh*) remains fragmented despite emerging localized efforts.

7.3 Open Problems

The path toward trustworthy **Islam-centric LLMs** requires addressing three fundamental "open problems" that currently limit the field as discussed below.

Data authenticity crisis. Most current models are trained on indiscriminate web scrapes (e.g., Common Crawl), which mix authoritative scholarship with forum discussions, paraphrases, and sectarian noise. We distinguish "Scholarly" versus "Web" content primarily by *data provenance*. Scholarly data is traceable to curated editions, recognized publishers, archival collections, or verified institutional repositories (with stable metadata and attribution), whereas web data is often unverified, anonymously authored, or *decontextualized* (e.g., community forums and social media). Under this framing, the challenge reduces to a *source-reliability* problem that is familiar across data-intensive fields (comparable to preferring peer-reviewed archives over unmoderated discussion boards). The near-term roadmap therefore prioritizes "Golden Corpora" as verified digital editions of classical heritage (*Turath*) and contemporary fatwas to serve as ground truth for training and retrieval, shifting from "Web Islam" toward "Scholarly Islam".

Reasoning and abstention gap. Islamic jurisprudence often functions like a logical code (e.g., rule-based inheritance shares or prayer-time constraints), however, LLMs are probabilistic systems that struggle with strict logic. A major open problem is enabling models to reliably perform symbolic reasoning or, crucially, to abstain (*Tawaqquf*) when an answer is ambiguous. This goes beyond *calibrated uncertainty*: the system may be highly confident in the retrieved evidence yet still need to defer because the question is underspecified, operative facts are missing, legitimate juristic disagreement remains unresolved, or the derivation requires methodological steps the system cannot verify. Future work should explore neuro-symbolic architectures where the LLM handles language understanding and evidence retrieval, while legal logic is delegated to deterministic solvers or verified rule sets, alongside explicit *deferral policies* triggered by missing premises, unresolved *ikhtilaf*/tarjīḥ, or unverified inference steps.

Latent alignment and interpretability. Bias against Islamic concepts is often deeply embedded in the model's internal representations (vector space) due to low or toxic representation of digital Islamic content. Merely fine-tuning a model on surface-level data is not enough if the underlying associations remain biased. The roadmap forward involves "mechanistic interpretability": developing tools to visualize and surgically edit these internal associations to ensure the model's "subconscious" aligns with the ethical objectives (*Maqāṣid*) of dignity, fairness, and truthfulness.

7.4 Shared-task directions

To accelerate convergence, the community would benefit from shared tasks that bundle methods and evaluation into end-to-end requirements. Examples include: *(i)* citation-grounded QA where every claim must be supported by retrieved spans; *(ii)* disagreement-aware answering where systems must either qualify by school or present multiple positions with evidences; *(iii)* abstention-calibrated jurisprudential reasoning (e.g., inheritance) where both correctness and refusal behavior are scored; and *(iv)* adversarial safety evaluations targeted to religious contexts (prompt injection, hate framing, selective quoting). Shared tasks should publish not only leaderboards but also error taxonomies and reproducible pipelines for provenance checks.

7.5 Key Takeaways

Three takeaways follow from the open problem mentioned above.

- **Reliability over fluency:** In the Islamic domain, a model's ability to generate fluent content is less important than its safety. Current generic models frequently hallucinate verses or miscalculate inheritance shares, posing significant theological and ethical risks.
- **Citation is mandatory, not optional:** For low-stake tasks, "prompt engineering" is insufficient for religious alignment. However, for high-stake tasks, systems should ground every answer with verified scripture (Quran/Hadith), explicitly citing sources rather than relying on model memory.
- **A shift in evaluation standards:** We are moving past simple translation tasks. The new standard for success is "CoT" reasoning: can the model handle complex jurisprudence (*Fiqh*) and cultural nuance without falling into Western-centric biases?

8 Conclusion

This survey consolidates a rapidly growing body of research on AI systems for Islamic knowledge. Following a PRISMA-ScR guided process, we curated and screened the literature and ultimately reviewed 160 papers published over the past decade. We introduce layered taxonomies that distinguish epistemic disciplines, AI task families, and a domain-task mapping that connects them, providing a unified view of how computational methods have advanced across different Islamic knowledge areas. Across the surveyed work, three foundations recur as drivers of progress: *(i)* curated, high-quality datasets, *(ii)* models and system designs tailored to knowledge-intensive and normatively sensitive tasks, and *(iii)* evaluation methodologies that measure not only accuracy but also faithfulness, uncertainty handling, and pluralism. Building on this structure, we summarize the current landscape of datasets, model capabilities (including LLMs), system design patterns, and evaluation practices. We also provide case studies of representative end-to-end pipelines that can serve as reusable blueprints for future frontier systems.

Acknowledgements and Funding Declaration. 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.

```
SYSTEM_INSTRUCTION = """
You are an expert reviewer for a survey paper on Islamic Knowledge in AI.
Evaluate the provided paper metadata.
"""

CRITERIA_SELECTED = [
  "Focuses on Core Sources: Quran, Hadith/Sunnah, Tafsir.",
  "Focuses on Islamic Sciences: Fiqh, Fatwas, Aqidah, Tajweed, Qira'at.",
  "Focuses on Islamic Ethics/Alignment: bias regarding Muslims, Islamic
    values in LLMs.",
  "Focuses on Classical Heritage (Turath): explicitly dealing with classical
    religious archives."
]

CRITERIA_NOT_SELECTED = [
  "General Arabic NLP (MSA/Dialects) with no religious focus (e.g., news
    classification).",
  "Purely sociological studies without an AI/Computational component."
]

OUTPUT = "Return ONLY: SELECTED or NOT_SELECTED."
```

Fig. A1: Screening prompt used for PRISMA-ScR paper selection.

Appendix A Appendix

A.1 LLM Prompts for Systematic Review

To ensure the reproducibility of our systematic review pipeline and to foster transparency regarding the automated components of our methodology described in Section 2 and Section 3, we provide the prompts used during the filtering and categorization phases.

A.1.1 Relevance Screening Prompt

This prompt was utilized in the **Inclusion** phase to evaluate the full text or detailed metadata of papers. The goal was to rigorously distinguish between general Arabic NLP papers and those specifically addressing Islamic knowledge domains.

A.1.2 Taxonomy Categorization Prompt

Following the selection of the final corpus, we employed a hierarchical labeling prompt (as shown in Figure A2) to categorize each paper into a Domain (Level 1), Main Topic (Level 2), and Subtopics. This prompt enforced a strict JSON output format to facilitate the downstream quantitative analysis presented in Section 3.

```
SYSTEM_INSTRUCTION = """
You are assisting with a literature survey titled:
"Discovering and Understanding Islamic Knowledge with AI". The dataset consists
     of academic papers on AI, NLP, and speech technologies applied to Islamic
     sources and contexts.
For EACH paper, produce:
- One broad 'domain' label (level 1).
- One specific 'main_topic' label (level 2).
- Zero to five 'subtopic' labels (secondary aspects).
- A short rationale for EACH label.
"""
LABELING_PRINCIPLES = {
  "DOMAIN_LEVEL_1": {
    "description": "Broad area capturing the primary source/corpus/context.",
    "format": "UPPER_SNAKE_CASE, 1-3 words",
    "allowed": [
      "QURAN", "HADITH", "FIQH", "ISLAMIC_HERITAGE",
      "GENERAL_ISLAMIC_NLP", "ISLAMIC_EDTECH", "ETHICS_ALIGNMENT",
      "BENCHMARKS_EVALUATION", "MULTISOURCE_ISLAMIC_CONTENT"
    ],"constraint": "Reuse domains when possible; create a new one only if
     needed."
  },
  "MAIN_TOPIC_LEVEL_2": {
    "description": "Core AI/computational contribution.",
    "format": "UPPER_SNAKE_CASE, ˜2-6 words",
    "examples": [
      "QURAN_TEXT_LEXICAL_RESOURCES", "HADITH_CHAIN_ANALYSIS", "FIQH_QA_SYSTEMS"
    ]},
  "SUBTOPICS": {
    "description": "0-5 secondary labels (tasks, methods, resources).",
    "constraint": "Do not restate MAIN_TOPIC in different words."},
  "RATIONALES": {"description": "1-3 sentences per label, grounded in
     title/abstract/keywords."}
}
OUTPUT_FORMAT = """ Return ONLY a single valid JSON object with this exact
     structure:
{
  "paper_id": "<PAPER_ID>",
  "domain": "<DOMAIN_LABEL>",
  "domain_rationale": "<rationale text>",
  "main_topic": "<MAIN_TOPIC_LABEL>",
  "main_topic_rationale": "<rationale text>",
  "subtopics": [{"label": "<SUBTOPIC_LABEL>", "rationale": "<rationale text>"}]
}
If there are no meaningful subtopics, output: "subtopics": []. Do NOT include any
     text outside the JSON object.
"""
```

Fig. A2: Labeling prompt for hierarchical domain/topic coding in the survey.

A.2 Search Prompts

To maximize recall and reduce false negatives caused by inconsistent transliteration and punctuation, we expanded each seed query into a family of English orthographic variants and near-synonyms. In particular, we normalized punctuation (e.g., `Qur'an` vs. `Quran`), hyphenation (e.g., `Al-Qur'an` vs. `Al Quran`), and capitalization, and we included commonly used transliteration alternatives. The primary normalization and variant families used throughout the search include:

- **Qur'an variants:** Quran, Qur'an, Qur'an, Koran, Al-Quran, Al Quran, Al-Qur'an, Al Qur'an.
- **Hadith variants:** Hadith, Hadeeth, Hadis, Ahadith, Ahaadeeth, plus closely associated terms such as isnad, sanad, and matn.
- **Fiqh and law variants:** fiqh, fiqhi, Islamic jurisprudence, Sharia, Shari'a, Shariah, Shari'ah, and supporting methodology terms such as usul al-fiqh, ijtihad, and qiyas.

Within the *core Islamic NLP* pillar, we operationalized the domain using task-oriented keywords that appear in titles, abstracts, and keyword fields, while also accounting for community-preferred synonyms (e.g., "ruling" for "fatwa") and knowledge-representation terminology used in semantic resources. The following query families were used (with the above orthographic expansions applied where relevant):

- **Islamic QA and assistants:** Islamic question answering, Islamic QA, religious question answering, Islamic chatbot, conversational agent, dialogue system, assistant.
- **Fatwa and legal reasoning:** fatwa generation, fatwa retrieval, Islamic ruling, ifta, mufti, fiqh LLMs, Islamic legal reasoning, legal reasoning, jurisprudence.
- **Knowledge resources and semantics:** Islamic ontology, Quran ontology, Hadith ontology, fiqh ontology, knowledge graph, semantic web, RDF, OWL, linked data.

For the *Arabic language modeling*, *evaluation resources*, and *cultural alignment* pillars, we used model-centric and dataset-centric terms in combination with Arabic-specific preprocessing and evaluation vocabulary. We additionally included alignment and cultural-safety terminology to retrieve work that treats normative or faith-sensitive behavior as an evaluation dimension:

- **Arabic modeling and preprocessing:** Arabic large language model, Arabic LLM, Arabic NLP transformer, Arabic text generation, Classical Arabic NLP, Qur'anic Arabic, MSA, Modern Standard Arabic, plus components such as tokenization, morphology, lemmatization, and diacritization/diacritisation.
- **Benchmarks and datasets:** Arabic LLM benchmark, Arabic evaluation benchmark, Islamic NLP dataset, Quran corpus dataset, Hadith corpus, Hadith dataset, and related metadata terms including annotation, gold standard, leaderboard, and human evaluation.
- **Culture, ethics, and alignment:** cultural awareness LLM, religious knowledge language model, Arabic cultural AI, Islamic ethics AI,

faith-sensitive, religious sensitivity, ethical AI, value alignment, safety, harm prevention.

References

Abdaljalil S, Kurban H, Serpedin E (2025) Halluverse25: Fine-grained multilingual benchmark dataset for llm hallucinations. URL https://arxiv.org/abs/2503.07833

Abouzied A, Alam F, Ali R, et al (2025) Combating misinformation in the arab world: Challenges and opportunities. Communications of the ACM 68(10):48–53

Abuz Y (2025) Islamic data: A comprehensive resource for islamic knowledge. https://www.kaggle.com/datasets/yousefabuz17/islamic-data

Ahmad KA, Shohibuddin WAJ, Eldeib AAM (2025) Artificial intelligence (ai) in quranic education: Systematic review. Quranica 14(1):58–69. URL https://ejournal.um.edu.my/index.php/quranica/article/view/64955

aiastrolabe (2025) Redteaming frontier llms with ai astrolabe arabic safety index (asas). Accessed: Oct. 3, 2025

Akra D, Hammouda T, Jarrar M (2025) QuranMorph: Morphologically Annotated Quranic Corpus. Tech. rep., Birzeit University, URL https://arxiv.org/pdf/2506.18148

Al Adel A, Bakr Soliman A, Sakher Sawan M, et al (2025) BurhanAI at IslamicEval 2025 shared task: Combating hallucinations in LLMs for islamic content; evaluation, correction, and retrieval-based solution. In: Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks. Association for Computational Linguistics, Suzhou, China, pp 503–508, https://doi.org/10.18653/v1/2025.arabicnlp-sharedtasks.69

Al-Khalifa S, Al-Khalifa H (2024) The qiyas benchmark: Measuring chatgpt mathematical and language understanding in arabic. arXiv preprint arXiv:240700146

Al-Khalifa S, Durrani N, Al-Khalifa H, et al (2025) The landscape of arabic large language models. Communications of the ACM

Al-Matham R, Darwish K, Al-Rasheed R, et al (2025) Balsam: A platform for benchmarking arabic large language models. URL https://arxiv.org/abs/2507.22603

Al-Monef R, Alhuzali H, Alturayeif N, et al (2025) Absher: A benchmark for evaluating large language models' understanding of saudi dialects. URL https://arxiv.org/abs/2507.10216

AL-Smadi M (2025) In: Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks. Association for Computational Linguistics, Suzhou, China, https://doi.org/10.18653/v1/2025.arabicnlp-sharedtasks.123, URL https://aclanthology.org/2025.arabicnlp-sharedtasks.123/

Alangari N, Team N (2025) N&N at QIAS 2025: Chain-of-thought ensembles with retrieval-augmented framework for classical arabic islamic MCQs. In: Proceedings of the ArabicNLP 2025 Shared Tasks. Association for Computational Linguistics

Alansari A, Luqman H (2025) AraHalluEval: A fine-grained hallucination evaluation framework for Arabic LLMs. In: Darwish K, Ali A, Abu Farha I, et al (eds) Proceedings of The Third Arabic Natural Language Processing Conference. Association for Computational Linguistics, Suzhou, China, pp 148–161, https://doi.org/10.18653/v1/2025.arabicnlp-main.12, URL https://aclanthology.org/2025.arabicnlp-main.12/

Alatas SF (2013) 3 ibn khaldun on education and knowledge. In: Ibn Khaldun. Oxford University Press, https://doi.org/10.1093/acprof:oso/9780198090458.003.0003, URL https://doi.org/10.1093/acprof:oso/9780198090458.003.0003

AlBallaa S, AlTwairesh N, AlSalman A, et al (2025) Gatmath and gatlc: Comprehensive benchmarks for evaluating arabic large language models. PLoS One 20(9):e0329129

Aleid HA, Azmi AM (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

Alghamdi EA, Masoud RI, Alnuhait D, et al (2024) Aratrust: An evaluation of trustworthiness for llms in arabic. URL https://arxiv.org/abs/2403.09017

Alghamdi EA, Masoud RI, Alnuhait D, et al (2025) AraTrust: An evaluation of trustworthiness for LLMs in Arabic. In: Rambow O, Wanner L, Apidianaki M, et al (eds) Proceedings of the 31st International Conference on Computational Linguistics. Association for Computational Linguistics, Abu Dhabi, UAE, pp 8664–8679, URL https://aclanthology.org/2025.coling-main.579/

Alhammad N, Awae F, Yussuf A (2025) Integrating artificial intelligence in islamic education: A review on pedagogical approaches and learning outcomes. International Journal of Academic Research in Business and Social Sciences 15(7):563–579. https://doi.org/10.6007/IJARBSS/v15-i7/25947

Aljaji H, Mohamed R, Ibrahim R, et al (2025) Benchmarking generative ai on quranic knowledge. In: Proceedings of the 5th Muslims in ML Workshop at NeurIPS 2025, URL https://www.musiml.org/events/2025-NeurIPS/accepted_papers.html

Alnefaie S, Atwell E, Alsalka MA (2023) Islamic question answering systems survey and evaluation criteria. International Journal on Islamic Applications in Computer Science And Technology 11(1):9–18

Alrabiah M, Al-Salman A, Atwell E (2014) Ksucca: A key to exploring arabic historical linguistics. International Journal of Computational Linguistics (IJCL) 5(2):27–36

Alshaikh R, Alghanmi I, Jeawak S (2025a) Aratable: Benchmarking llms' reasoning and understanding of arabic tabular data. URL https://arxiv.org/abs/2507.18442

Alshaikh R, Alghanmi I, Jeawak S (2025b) AraTable: Benchmarking LLMs' reasoning and understanding of arabic tabular data. arXiv preprint arXiv:250718442

Altammami S (2023) Quran_hadith_datasets. GitHub repository, URL https://github.com/ShathaTm/Quran_Hadith_Datasets

Altammami S, Atwell E, Alsalka A (2020) The arabic–english parallel corpus of authentic hadith. In: International Conference on Islamic Applications in Computer Science and Technologies (IMAN 2019), pp 1–10, URL https://eprints.whiterose.ac.uk/id/eprint/160497/, published online: 2020-06-30 (IJASAT proceedings record via White Rose Research Online)

Alwajih F, El Mekki A, Magdy SM, et al (2025a) Palm: A culturally inclusive and linguistically diverse dataset for Arabic LLMs. In: Che W, Nabende J, Shutova E, et al (eds) Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Vienna, Austria, pp 32871–32894, https://doi.org/10.18653/v1/2025.acl-long.1579, URL https://aclanthology.org/2025.acl-long.1579/

Alwajih F, El Mekki A, Mubarak H, et al (2025b) PalmX 2025: The first shared task on benchmarking LLMs on Arabic and islamic culture. In: Darwish K, Ali A, Abu Farha I, et al (eds) Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks. Association for Computational Linguistics, Suzhou, China, pp 774–789, https://doi.org/10.18653/v1/2025.arabicnlp-sharedtasks.107, URL https://aclanthology.org/2025.arabicnlp-sharedtasks.107/

Alwajih F, Mekki AE, Magdy SM, et al (2025c) Palm: A Culturally Inclusive and Linguistically Diverse Dataset for Arabic LLM. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Vienna, Austria, URL https://arxiv.org/pdf/2503.00151

Alzubaidi A, Alsuwaidi S, Boussaha BEA, et al (2025) Evaluating arabic large language models: A survey of benchmarks, methods, and gaps. arXiv preprint arXiv:251013430

Antoun W, Baly F, Hajj H (2020) Arabert: Transformer-based model for arabic language understanding. In: Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection. European Language Resources Association (ELRA), pp 9–15

Plaza-del Arco FM, Curry AC, Paoli S, et al (2024) Divine LLaMAs: Bias, stereotypes, stigmatization, and emotion representation of religion in large language models. In: Al-Onaizan Y, Bansal M, Chen YN (eds) Findings of the Association for Computational Linguistics: EMNLP 2024. Association for Computational Linguistics, Miami, Florida, USA, pp 4346–4366, https://doi.org/10.18653/v1/2024.findings-emnlp.251, URL https://aclanthology.org/2024.findings-emnlp.251/

Ashour AF, Rashdan W (2025) Heritage-aware generative AI workflow for islamic geometry in interiors. Heritage 8(11):486. https://doi.org/10.3390/heritage8110486

Ashraf Y, Wang Y, Gu B, et al (2025) Arabic dataset for LLM safeguard evaluation. In: Chiruzzo L, Ritter A, Wang L (eds) Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). Association for Computational Linguistics, Albuquerque, New Mexico, pp 5529–5546, https://doi.org/10.18653/v1/2025.naacl-long.285, URL https://aclanthology.org/2025.naacl-long.285/

Asl MA, Bidgoli BM (2025) Farsiqa: Faithful and advanced rag system for islamic question answering. 251025621v1 URL http://arxiv.org/abs/2510.25621v1 [cs.CL]

Asma Abdul-Qader Abdullah Al-Ani IAKSAK (2024) The holy quran. Mesopotamian journal of Quran studies URL https://www.semanticscholar.org/paper/d15deb526ca59e3d8891e32bf18b6bad9f994e14

Asseri B, Abdelaziz E, Al-Wabil A (2025) Prompt engineering techniques for mitigating cultural bias against arabs and muslims in large language models: A systematic review. 250618199v2 URL http://arxiv.org/abs/2506.18199v2 [cs.CL]

Atif F, Askarbekuly N, Darwish K, et al (2025) Sacred or synthetic? evaluating llm reliability and abstention for religious questions. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, pp 217–226

Azmi AM, Al-Qabbany AO, Hussain A (2019) Computational and natural language processing based studies of hadith literature: a survey. Artificial Intelligence Review 52(2):1369–1414. https://doi.org/10.1007/s10462-019-09692-w, URL https://doi.org/10.1007/s10462-019-09692-w

Badry M, Hassan H, Bayomi H, et al (2018) Qtid: Quran text image dataset. International Journal of Advanced Computer Science and Applications 9(3):385–391. https://doi.org/10.14569/IJACSA.2018.090351, URL https://thesai.org/Publications/ViewPaper?Code=IJACSA&Issue=3&SerialNo=51&Volume=9

Bahaj A, Ghogho M (2025) Mizanqa: Benchmarking large language models on moroccan legal question answering. arXiv preprint arXiv:250816357

Bari MS, Alnumay Y, Alzahrani NA, et al (2025) ALLam: Large language models for arabic and english. In: The Thirteenth International Conference on Learning

Representations, URL https://openreview.net/forum?id=MscdsFVZrN

Bashir MH, Azmi AM, Nawaz H, et al (2021) Arabic natural language processing for qur'anic research: a systematic review. Artificial Intelligence Review 56(Suppl 1):13951–13993

Bashir MH, Azmi AM, Nawaz H, et al (2023) Arabic natural language processing for Qur'anic research: a systematic review. Artificial Intelligence Review 56(7):6801–6854. https://doi.org/10.1007/s10462-022-10313-2, URL https://doi.org/10.1007/s10462-022-10313-2

Bellino F (2014) The classification of sciences in an ottoman arabic encyclopaedia: Taşköprüzāde's Miftāḥ al-Sa'āda. Quaderni di Studi Arabi pp 161–180

Ben Ayed H, Najari M, Makni W, et al (2025) A comparative study of arabic embedding models in rag-based fatwa retrieval. In: ResearchGate, conference Paper

Bhatia G, Nagoudi EMB, Mekki AE, et al (2024) Swan and arabicmteb: Dialect-aware, arabic-centric, cross-lingual, and cross-cultural embedding models and benchmarks. 241101192v2 URL http://arxiv.org/abs/2411.01192v2 [cs.CL]

Bhatia G, Nagoudi EMB, El Mekki A, et al (2025) Swan and ArabicMTEB: Dialect-aware, Arabic-centric, cross-lingual, and cross-cultural embedding models and benchmarks. In: Chiruzzo L, Ritter A, Wang L (eds) Findings of the Association for Computational Linguistics: NAACL 2025. Association for Computational Linguistics, Albuquerque, New Mexico, pp 4654–4670, https://doi.org/10.18653/v1/2025.findings-naacl.263, URL https://aclanthology.org/2025.findings-naacl.263/

Bhatia G, Mubarak H, Jarrar M, et al (2026) From rag to agentic rag for faithful islamic question answering. arXiv preprint arXiv:260107528 https://doi.org/10.48550/arXiv.2601.07528

Bouchekif A, Rashwani S, Mohamed ESA, et al (2025a) QIAS 2025: Overview of the shared task on islamic inheritance reasoning and knowledge assessment. In: Darwish K, Ali A, Abu Farha I, et al (eds) Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks. Association for Computational Linguistics, Suzhou, China, pp 851–860, https://doi.org/10.18653/v1/2025.arabicnlp-sharedtasks.117, URL https://aclanthology.org/2025.arabicnlp-sharedtasks.117/

Bouchekif A, Rashwani S, Sbahi H, et al (2025b) Assessing large language models on islamic legal reasoning: Evidence from inheritance law evaluation. arxiv URL http://arxiv.org/abs/2509.01081v2

Brown J (2008) How we know early hadīth critics did matn criticism and why it's so hard to find. Islamic Law and Society 15(2):143 – 184. https://doi.org/10.1163/156851908X290574, URL https://brill.com/view/journals/ils/15/2/article-p143_1.xml

Chandna B, Bashir Z, Sen P (2025) Dissecting bias in llms: A mechanistic interpretability perspective. 250605166v2 URL http://arxiv.org/abs/2506.05166v2 [cs.CL]

Comanici G, Bieber E, Schaekermann M, et al (2025) Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. URL https://arxiv.org/abs/2507.06261, arXiv:2507.06261

Dubois Y, Li X, Taori R, et al (2023) Alpacafarm: A simulation framework for methods that learn from human feedback

Dukes K, Habash N (2010) Morphological annotation of quranic arabic. Language Resources and Evaluation 44:453–482

El Filali A, Alobeidli H, Fourrier C, et al (2024a) Open arabic llm leaderboard. https://huggingface.co/spaces/OALL/Open-Arabic-LLM-Leaderboard

El Filali A, Sengupta N, Abouelseoud, et al (2024b) Rethinking llm evaluation with 3c3h: Aragen benchmark and leaderboard. https://https://huggingface.co/blog/leaderboard-3c3h-aragen

El Filali A, ALOUI M, Husaain T, et al (2025) The open arabic llm leaderboard 2. https://huggingface.co/spaces/OALL/Open-Arabic-LLM-Leaderboard

El Kheir Y, Meghanani A, Toyin HO, et al (2025) Iqra'eval: A shared task on qur'anic pronunciation assessment. In: Darwish K, Ali A, Abu Farha I, et al (eds) Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks. Association for Computational Linguistics, Suzhou, China, pp 443–452, https://doi.org/10.18653/v1/2025.arabicnlp-sharedtasks.61, URL https://aclanthology.org/2025.arabicnlp-sharedtasks.61/

Elsharif WAO, Alzubaidi M, She J, et al (2025) Ara-pic: A framework for enhancing arabic cultural representation in AI-generated images. In: 2025 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), https://doi.org/10.1109/ICMEW68306.2025.11152135

Elston M (2022) Becoming turāth: the islamic tradition in the modern period. Die Welt des Islams 63(4):441 – 473. https://doi.org/10.1163/15700607-20220026, URL https://brill.com/view/journals/wdi/63/4/article-p441_003.xml

Farhana Khandaker Mursheda MKA (2025) Artificial intelligence (ai) opens a new horizon in the study of the holy quran. Theoretical and applied technological science review URL https://www.semanticscholar.org/paper/2eda0e8ada48489bb86ba7ac1c212315e9830df7

Faruk KMTM, Talha MR, Ahamad HMK, et al (2025) Adab: A culturally-aligned automated response generation framework for islamic app reviews by integrating absa and hybrid rag. In: Proceedings of the 5th Muslims in ML Workshop at NeurIPS 2025, URL https://openreview.net/pdf?id=PnWmDdwTXE

Fawzi M, Ross B, Magdy W (2024) ”the prophet said so!”: On exploring hadith presence on arabic social media. URL https://arxiv.org/abs/2412.20581, arXiv:2412.20581

Fawzi M, Ross B, Magdy W (2025) Fabricating holiness: Characterizing religious misinformation circulators on arabic social media. URL https://arxiv.org/abs/2508.07845, arXiv:2508.07845

Gema AP, Hägele A, Chen R, et al (2025) Inverse scaling in test-time compute. 250714417v1 URL http://arxiv.org/abs/2507.14417v1 [cs.AI]

Grattafiori A, Dubey A, Jauhri A, et al (2024) The llama 3 herd of models. URL https://arxiv.org/abs/2407.21783

Guo G, Naous T, Wakaki H, et al (2025) CARE: Multilingual human preference learning for cultural awareness. In: Christodoulopoulos C, Chakraborty T, Rose C, et al (eds) Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Suzhou, China, pp 32854–32883, https://doi.org/10.18653/v1/2025.emnlp-main.1669, URL https://aclanthology.org/2025.emnlp-main.1669/

Gürer DZ, Atlamaz U, Özateş ŞB (2025) Text extraction and script completion in images of arabic script-based calligraphy: A thesis proposal. In: Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop. Association for Computational Linguistics

Habib N, Fourrier C, Kydlíček H, et al (2023) Lighteval: A lightweight framework for llm evaluation. URL https://github.com/huggingface/lighteval

Hakim A, Anggraini P (2023) Artificial intelligence in teaching islamic studies: Challenges and opportunities. Molang: Journal Islamic Education 1(2):19–30. https://doi.org/10.32806/jm.v1i2.619, URL https://jurnalalkhairat.org/ojs/index.php/molang/article/view/619

Hamad ZT, Laouar MR, Bendib I, et al (2022) Arabic quran verses authentication using deep learning and word embeddings. The International Arab Journal of Information Technology URL https://www.semanticscholar.org/paper/533c68890751a45303b6d1a71724d02f9ea12347

Harrag F, Al-Nasser A, Al-Musnad A, et al (2020) Quran intelligent ontology construction approach using association rules mining URL http://arxiv.org/abs/2008.03232v2

Harvey S (2020) Factsheet: Islam. URL https://religionmediacentre.org.uk/factsheets/islam/

Hasan Z (2025) Sparse-checklist prompting for arabic grammar tutoring: Fast, token-efficient feedback. In: Proceedings of the 5th Muslims in ML Workshop at NeurIPS 2025, URL https://www.musiml.org/events/2025-NeurIPS/accepted_papers.html

Haval H. Ameen AK (2024) Ai model for parsing the text of holy quran sentences. arxiv URL https://www.semanticscholar.org/paper/b2bad1b1231ed4d1a2f07a08c21c91fc6c194e74

Hendrycks D, Burns C, Basart S, et al (2020) Aligning ai with shared human values. 200802275v6 URL http://arxiv.org/abs/2008.02275v6 [cs.CY]

Hijazi F, AlHarbi S, AlHussein A, et al (2024) Arablegaleval: A multitask benchmark for assessing Arabic legal knowledge in large language models. arXiv preprint arXiv:240807983

Hosseini M, Hosseini K, Bali S, et al (2025) Perhallueval: Persian hallucination evaluation benchmark for large language models. 250921104v1 URL http://arxiv.org/abs/2509.21104v1 [cs.CL]

Huang H, Yu F, Zhu J, et al (2024) Acegpt, localizing large language models in arabic. URL https://arxiv.org/abs/2309.12053

Huda A, Fauziah, Ratnawulan E, et al (2021) Arabic part of speech (pos) tagging analysis using bee colony optimization (bco) algorithm on quran corpus. arxiv URL https://www.semanticscholar.org/paper/63aa61a5147cc4b0ed89b47b8cdada0b942d3272

Hui Z, Dong YR, Shareghi E, et al (2025) TRIDENT: Benchmarking llm safety in finance, medicine, and law. arXiv preprint arXiv:250721134 URL https://arxiv.org/abs/2507.21134

Jarrar M (2021) The Arabic Ontology - An Arabic Wordnet with Ontologically Clean Content. Applied Ontology Journal 16(1):1–26. https://doi.org/10.3233/AO-200241

Jarrar M, Hammouda TH (2024) Qabas: An Open-Source Arabic Lexicographic Database. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). ELRA and ICCL, Torino, Italy, pp 13363–13370, URL https://aclanthology.org/2024.lrec-main.1170.pdf

Kanaan G, Al-Shalabi R, Gharaibeh S (2013) A comparison of various neural network approaches for arabic text classification. International Journal of Advanced Research in Artificial Intelligence 2(2)

Kaneko M, Bollegala D, Baldwin T (2024) Eagle: Ethical dataset given from real interactions. 240214258v1 URL http://arxiv.org/abs/2402.14258v1 [cs.CL]

Keleg A (2025) Llm alignment for the arabs: A homogenous culture or diverse ones? 250315003v1 URL http://arxiv.org/abs/2503.15003v1 [cs.CL]

Keleg A, Goldwater S, Magdy W (2025) Revisiting common assumptions about arabic dialects in nlp. arXiv preprint arXiv:250521816

Khalila Z, Nasution AH, Monika W, et al (2025a) Investigating retrieval-augmented generation in quranic studies: A study of 13 open-source large language models. 250316581v1 https://doi.org/10.14569/IJACSA.2025.01602134, URL http://arxiv.org/abs/2503.16581v1, international Journal of Advanced Computer Science and Applications(IJACSA), 16(2), 2025 [cs.CL]

Khalila Z, Nasution AH, Monika W, et al (2025b) Investigating retrieval-augmented generation in quranic studies: A study of 13 open-source large language models. (IJACSA) International Journal of Advanced Computer Science and Applications 16(2). URL https://thesai.org/Downloads/Volume16No2/Paper_134-Investigating_Retrieval_Augmented_Generation_in_Quranic_Studies.pdf

Khan RS, Rahman A (2025) Computationally distinguishing quran and pre-islamic arabic poetry. 2025 Eighth International Women in Data Science Conference at Prince Sultan University (WiDS PSU) URL https://www.semanticscholar.org/paper/5238b3456f49e1186f87c6302c2aea2fcf5cf994

Koto F, Li H, Shatnawi S, et al (2024) ArabicMMLU: Assessing massive multitask language understanding in Arabic. In: Ku LW, Martins A, Srikumar V (eds) Findings of the Association for Computational Linguistics: ACL 2024. Association for Computational Linguistics, Bangkok, Thailand, pp 5622–5640, https://doi.org/10.18653/v1/2024.findings-acl.334, URL https://aclanthology.org/2024.findings-acl.334/

Koubaa A, Ammar A, Ghouti L, et al (2024) Arabiangpt: Native arabic gpt-based large language model. URL https://arxiv.org/abs/2402.15313

Lahmar A, Arafat ME, Farou Z, et al (2025) Islamtrust: A benchmark for llms alignment with islamic values. In: Proceedings of the 5th Muslims in ML Workshop at NeurIPS 2025, URL https://openreview.net/forum?id=PBcv90iKFB

Li J, Li Y, Wan X (2025) Analyzing cognitive differences among large language models through the lens of social worldview. 250501967v1 URL http://arxiv.org/abs/2505.01967v1 [cs.CL]

M. Othman YMEHM. A. Al-Hagery (2020) Arabic text processing model: Verbs roots and conjugation automation. arxiv URL https://www.semanticscholar.org/paper/ea17a3edbf5b13d26d517f50d6e8f4447ab901a2

Magdy SM, Kwon SY, Alwajih F, et al (2025) JAWAHER: A multidialectal dataset of Arabic proverbs for LLM benchmarking. In: Chiruzzo L, Ritter A, Wang L (eds) Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the

Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). Association for Computational Linguistics, Albuquerque, New Mexico, pp 12320–12341, https://doi.org/10.18653/v1/2025.naacl-long.613, URL https://aclanthology.org/2025.naacl-long.613/

Mahdi MS (2026) Islam. https://www.britannica.com/topic/Islam, encyclopaedia Britannica. Last updated 2026-01-05. Accessed 2026-01-10

Malhas R, Mansour W, Elsayed T (2022) Qur'an qa 2022: Overview of the first shared task on question answering over the holy qur'an. In: Proceedings of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection, URL https://aclanthology.org/2022.osact-1.9/

Malhas R, Mansour W, Elsayed T (2023) Qur'an qa 2023 shared task: Overview of passage retrieval and reading comprehension tasks over the holy qur'an. In: Proceedings of ArabicNLP 2023, URL https://aclanthology.org/2023.arabicnlp-1.76/

Martínez AG (2025) A computational system to handle the orthographic layer of tajwid in contemporary quranic orthography URL http://arxiv.org/abs/2505.11379v1

Mashaabi M, Al-Khalifa S, Al-Khalifa H (2024) A survey of large language models for arabic language and its dialects. arXiv preprint arXiv:241020238

Mazid N, Ahmad M (2025) Tajweedai: A hybrid asr-classifier for real-time qalqalah detection in quranic recitation. In: Proceedings of the 5th Muslims in ML Workshop at NeurIPS 2025, URL https://openreview.net/forum?id=AauWmDPOIf

Meadows GI, Lau NWL, Susanto EA, et al (2024) Localvaluebench: A collaboratively built and extensible benchmark for evaluating localized value alignment and ethical safety in large language models. 240801460v1 URL http://arxiv.org/abs/2408.01460v1 [cs.CY]

Mekki AE, Atou H, Nacar O, et al (2025) Nilechat: Towards linguistically diverse and culturally aware llms for local communities. 250518383v3 URL http://arxiv.org/abs/2505.18383v3 [cs.CL]

Mghari M, Bouras O, El Hibaoui A (2022) Sanadset 650k: Data on hadith narrators. Data in Brief 44:108540. https://doi.org/10.1016/j.dib.2022.108540

Mhnaa D, Dayoub Y, Salman J (2025) Development of an intelligent system for recognizing islamic religious visual signs in the arabic language. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, pp 4874–4882

Modarres Kamaly A (2025) Towards inclusive nlp: Evaluating llms on low-resource indo-iranian languages. In: Proceedings of the 5th Muslims in ML Workshop at NeurIPS

2025, URL https://www.musiml.org/events/2025-NeurIPS/accepted_papers.html

Mohammed MY, Ali SA, Ali SK, et al (2025a) Aftina: enhancing stability and preventing hallucination in ai-based islamic fatwa generation using llms and RAG. Neural Computing and Applications 37(25):20957–20982. https://doi.org/10.1007/S00521-025-11229-Y, URL https://doi.org/10.1007/s00521-025-11229-y

Mohammed MY, Ali SA, Ali SK, et al (2025b) Aftina: enhancing stability and preventing hallucination in ai-based islamic fatwa generation using llms and rag. Neural Computing and Applications 37(25):20957–20982. https://doi.org/10.1007/s00521-025-11229-y, URL http://dx.doi.org/10.1007/s00521-025-11229-y

Mousi B, Durrani N, Ahmad F, et al (2025) AraDiCE: Benchmarks for dialectal and cultural capabilities in LLMs. In: Rambow O, Wanner L, Apidianaki M, et al (eds) Proceedings of the 31st International Conference on Computational Linguistics. Association for Computational Linguistics, Abu Dhabi, UAE, pp 4186–4218, URL https://aclanthology.org/2025.coling-main.283/

Mubarak H, Al-Khalifa H, Alkhalefah KS (2024) Halwasa: Quantify and analyze hallucinations in large language models: Arabic as a case study. In: Calzolari N, Kan MY, Hoste V, et al (eds) Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). ELRA and ICCL, Torino, Italia, pp 8008–8015, URL https://aclanthology.org/2024.lrec-main.705/

Mubarak H, Malhas R, Mansour W, et al (2025) IslamicEval 2025: The first shared task of capturing LLMs hallucination in islamic content. In: Darwish K, Ali A, Abu Farha I, et al (eds) Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks. Association for Computational Linguistics, Suzhou, China, pp 480–493, https://doi.org/10.18653/v1/2025.arabicnlp-sharedtasks.67, URL https://aclanthology.org/2025.arabicnlp-sharedtasks.67/

Mushtaq A, Naeem R, Elmahjub E, et al (2025) Can llms write faithfully? an agent-based evaluation of llm-generated islamic content. 251024438v1 URL http://arxiv.org/abs/2510.24438v1 [cs.CL]

Namoun A, Humayun MA, Nawaz W (2024) A multimodal data scraping tool for collecting authentic islamic text datasets. International Journal of Advanced Computer Science and Applications 15(12). URL https://thesai.org/Downloads/Volume15No12/Paper_24-A_Multimodal_Data_Scraping_Tool.pdf

Naous T, Xu W (2025) On the origin of cultural biases in language models: From pre-training data to linguistic phenomena. 250104662v1 URL http://arxiv.org/abs/2501.04662v1 [cs.CL]

Naous T, Ryan MJ, Ritter A, et al (2023) Having beer after prayer? measuring cultural bias in large language models. 230514456v4 URL http://arxiv.org/abs/2305.14456v4

[cs.CL]

NoorSoft (2024) Noor digital library site (noorlib) with more than 115,000 book volumes. https://www.noorsoft.org/en/News/View/111588/Noor-digital-library-site-%28Noorlib%29-with-more-than-115%2C000-book-volumes, accessed 2026-01-25

Omarov Z, Sultimov R, Volkov A, et al (2025) Llm agent-based modeling for zakat policy simulation in islamic finance. In: Proceedings of the 5th Muslims in ML Workshop at NeurIPS 2025, URL https://www.musiml.org/events/2025-NeurIPS/accepted_papers.html

Omayrah A, Alkhereyf S, Abdelali A, et al (2025) HUMAIN at IslamicEval 2025 shared task 1: A three-stage LLM-based pipeline for detecting and correcting hallucinations in Quran and Hadith. In: Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks. Association for Computational Linguistics, Suzhou, China, pp 509–514, https://doi.org/10.18653/v1/2025.arabicnlp-sharedtasks.70, URL https://aclanthology.org/2025.arabicnlp-sharedtasks.70/

OpenAI (2023) Chatgpt. https://chat.openai.com/chat, accessed: 2 October 2025

OpenAI, Achiam J, Adler S, et al (2024) Gpt-4 technical report. URL https://arxiv.org/abs/2303.08774

Oshallah I, Mohamed Basem AH, Mohammed A (2025) Cross-language approach for quranic qa. Proceedings of Tenth International Congress on Information and Communication Technology (ICICT 2025) URL http://arxiv.org/abs/2501.17449v1

Page MJ, McKenzie JE, Bossuyt PM, et al (2021) The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 372

Patel S, Kane H, Patel R (2023) Building domain-specific llms faithful to the islamic worldview: Mirage or technical possibility? 231206652v1 URL http://arxiv.org/abs/2312.06652v1 [cs.AI]

Pavlova V (2025) Multi-stage training of bilingual islamic llm for neural passage retrieval. 250110175v1 URL http://arxiv.org/abs/2501.10175v1 [cs.CL]

Peuriekeu YM, Noyum VD, Feudjio C, et al (2021) A text mining discovery of similarities and dissimilarities among sacred scriptures URL https://www.semanticscholar.org/paper/f8f1a49a709fac1d82af680fe970d10a71cd2ecf

Phuc NX, Thin DV (2025) PuxAI at QIAS 2025: Multi-agent retrieval-augmented generation for islamic inheritance and knowledge reasoning. In: Proceedings of the ArabicNLP 2025 Shared Tasks. Association for Computational Linguistics

Prakash N, Roy LKW (2024) Interpreting bias in large language models: A feature-based approach. 240612347v1 URL http://arxiv.org/abs/2406.12347v1 [cs.CL]

Premasiri D, Ranasinghe T, Zaghouani W, et al (2022) Dtw at qur'an qa 2022: Utilising transfer learning with transformers for question answering in a low-resource domain URL https://www.semanticscholar.org/paper/a56c8e6b2db32abe2c38bbc4a78a4a895137d15d

Qian Z, Altam F, Alqurishi M, et al (2024) Cameleval: Advancing culturally aligned arabic language models and benchmarks. URL https://arxiv.org/abs/2409.12623

Raghad Salameh MAM (2024) Quranic audio dataset: Crowdsourced and labeled recitation from non-arabic speakers. arxiv URL http://arxiv.org/abs/2405.02675v1

Rhel H, Roussinov D (2025) Large language models and Arabic content: A review. In: International Conference on AI: Current Research, Industry Trends, and Innovations, Springer, pp 402–419

Ridoy SZ, Wasi AT, Tonmoy KA (2025) Bengalimoralbench: A benchmark for auditing moral reasoning in large language models within bengali language and culture. 251103180v1 URL http://arxiv.org/abs/2511.03180v1 [cs.CL]

Rippin A (2022) The Qur'an and its interpretative tradition. Routledge

Robinson NR, Abdelmoneim S, Marchisio K, et al (2025) Al-qasida: Analyzing llm quality and accuracy systematically in dialectal arabic. In: Findings of the Association for Computational Linguistics: ACL 2025, URL https://aclanthology.org/2025.findings-acl.1137.pdf

Rushdi AA (2025) Technical vs cultural: Evaluating llms in arabic. In: Proceedings of the 5th Muslims in ML Workshop at NeurIPS 2025, URL https://www.musiml.org/events/2025-NeurIPS/accepted_papers.html

Sadallah A, Tonga JC, Almubarak K, et al (2025) Commonsense reasoning in Arab culture. In: Che W, Nabende J, Shutova E, et al (eds) Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Vienna, Austria, pp 7695–7710, https://doi.org/10.18653/v1/2025.acl-long.380, URL https://aclanthology.org/2025.acl-long.380/

Sahebi A, Hemmatyar M, Asgari E (2025) Context-aware extraction of quranic references: A hybrid language model- and rule-based approach. In: Proceedings of the 5th Muslims in ML Workshop at NeurIPS 2025, URL https://www.musiml.org/events/2025-NeurIPS/accepted_papers.html

Salameh R, Al Mdfaa M, Askarbekuly N, et al (2024) Quranic audio dataset: Crowdsourced and labeled recitation from non-arabic speakers. Procedia Computer Science 246:2684–2693. https://doi.org/10.1016/j.procs.2024.09.404

Saleh A, Al-Khalifa H (2020) Shamela: A large-scale historical arabic corpus. In: Proceedings of the LREC 2020 Workshop on Language Resources and Evaluation. ELRA

Salman MU, Qazi MA, Alam MT (2025) Quran-md: A fine-grained multilingual multimodal dataset of the quran. In: Proceedings of the 5th Muslims in ML Workshop at NeurIPS 2025, URL https://openreview.net/forum?id=NQ6er5I4PK

Schmidtke S (2016) Introduction. In: The Oxford Handbook of Islamic Theology. Oxford University Press, https://doi.org/10.1093/oxfordhb/9780199696703.013.48, URL https://doi.org/10.1093/oxfordhb/9780199696703.013.48

Sengupta N, Sahu SK, Jia B, et al (2023) Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. URL https://arxiv.org/abs/2308.16149

Seth A, Choudhary M, Sitaram S, et al (2025) How deep is representational bias in llms? the cases of caste and religion. 250803712v1 URL http://arxiv.org/abs/2508.03712v1 [cs.CL]

Shankar H, P VS, Cavale T, et al (2025) Sometimes the model doth preach: Quantifying religious bias in open llms through demographic analysis in asian nations. 250307510v1 URL http://arxiv.org/abs/2503.07510v1 [cs.CY]

Sibaee S, Nacar O, Ammar A, et al (2025) From guidelines to practice: A new paradigm for arabic language model evaluation. arxiv URL https://www.semanticscholar.org/paper/e5bdc7c06581bbc5871432f95a9036850c541dc1

Simbeck K, Mahran M (2025) Mechanistic interpretability with saes: Probing religion, violence, and geography in large language models. 250917665v1 URL http://arxiv.org/abs/2509.17665v1 [cs.LG]

Soliman AB, Eissa K, El-Beltagy SR (2017) Aravec: A set of arabic word embedding models for use in arabic nlp. In: Procedia Computer Science, vol 117. Elsevier, pp 256–265

Sumayli A, Alkaoud M (2025) Handwritten arabic calligraphy generation: A systematic literature review. International Journal of Advanced Computer Science and Applications 16(3). https://doi.org/10.14569/IJACSA.2025.0160381

Sun M, Yin Y, Xu Z, et al (2025) Idiosyncrasies in large language models. 250212150v2 URL http://arxiv.org/abs/2502.12150v2 [cs.CL]

Sunnah.com (2024) Sunnah.com: The hadith of the prophet muhammad (pbuh). https://sunnah.com

Team F, Abbas U, Ahmad MS, et al (2025a) Fanar: An arabic-centric multimodal generative ai platform. 250113944 URL https://arxiv.org/abs/2501.13944 [cs.CL]

Team G, Kamath A, Ferret J, et al (2025b) Gemma 3 technical report. URL https://arxiv.org/abs/2503.19786

Team S (2024) Silma. URL https://www.silma.ai

Umme Hani AR (2024) Predicting revelation periods of verses of the quran via deep learning URL https://www.semanticscholar.org/paper/4ce8831ba0b6517cc0e1f2842a925e737d81f6c9

Versteegh K (2014) The Arabic Linguistic Tradition, Edinburgh University Press, p 107–125

Wen B, Yao J, Feng S, et al (2025) Know your limits: A survey of abstention in large language models. Transactions of the Association for Computational Linguistics 13:529–556. https://doi.org/10.1162/tacl_a_00754, URL https://aclanthology.org/2025.tacl-1.26/

Yang A, Li A, Yang B, et al (2025) Qwen3 technical report. URL https://arxiv.org/abs/2505.09388

Yasser Shohoud SAMaged Shoman (2023) Quranic conversations: Developing a semantic search tool for the quran using arabic nlp techniques. arxiv URL http://arxiv.org/abs/2311.05120v1

Youssef O (2025) islamic-qa-egyptian-arabic. Hugging Face dataset, URL https://huggingface.co/datasets/Omar-youssef/islamic-qa-egyptian-arabic, accessed: 2025-12-30. License: Apache-2.0

Yu H, Jeong S, Pawar S, et al (2025) Entangled in representations: Mechanistic investigation of cultural biases in large language models. 250808879v1 URL http://arxiv.org/abs/2508.08879v1 [cs.CL]

Yudiono IRI, Permadi DP (2025) Artificial intelligence and spirituality: Can ai understand the divine in sufism? SUHU: Journal of Sufism and Humanities 1(1):19–30. URL https://suhu.lakaspia.org/index.php/suhu/article/view/28

Zerrouki T, Balla A (2017) Tashkeela: Novel corpus of arabic vocalized texts, data for auto-diacritization systems. Mendeley Data, V2, https://doi.org/10.17632/45bi5514-2

Zhao JX, Hooi B, Ng SK (2025) Test-time scaling in reasoning models is not effective for knowledge-intensive tasks yet. 250906861v1 URL http://arxiv.org/abs/2509.06861v1 [cs.AI]

Zhong T, Yang Z, Liu Z, et al (2024) Opportunities and challenges of large language models for low-resource languages in humanities research. 241204497v3 URL http://arxiv.org/abs/2412.04497v3 [cs.CL]