Foundations of Domain-specific Large Language Models for Islamic Studies: A Comprehensive Review

Mohamed Yassine El Amrani*, Arshad Vakayil, Feroz Mohammed & Faisal Al Amri

Department of Computer and Information Technology, Jubail Industrial College, Royal Commission of Jubail and Yanbu - Jubail, Al Huwailat Road, Jubail Industrial City, Saudi Arabia, 35718.
*E-mail: amranim@rcjy.edu.sa

Abstract. Large language models (LLMs) have undergone rapid evolution and are highly effective in tasks such as text generation, question answering, and context-driven analysis. However, the unique requirements of Islamic studies, where textual authenticity, diverse jurisprudential interpretations, and deep semantic nuances are critical, present challenges for general LLMs. This article reviews the evolution of neural language models by comparing the historical progression of general LLMs with emerging Islamic-specific LLMs. We discuss the technical foundations of modern Transformer architectures and examine how recent advancements, such as GPT-4, DeepSeek, and Mistral, have expanded LLM capabilities. The paper also highlights the limitations of standard evaluation metrics like perplexity and BLEU in capturing doctrinal, ethical, and interpretative accuracy. To address these gaps, we propose specialized evaluation metrics to assess doctrinal correctness, internal consistency, and overall reliability. Finally, we outline a research roadmap aimed at developing robust, ethically aligned, and jurisprudentially precise Islamic LLMs.

Keywords: *bias mitigation; ethical AI; fiqh; Islamic studies; large language models; natural language processing; transformer architecture.*

1 Introduction

General-purpose large language models (LLMs) such as GPT, BERT, T5, and DeepSeek have advanced natural language processing (NLP) by enabling breakthroughs in machine translation, information retrieval, and text summarization [1-3]. However, when applied to religious or culturally sensitive domains, these models sometimes fall short [4][5]. Islamic studies, especially in *tafsir* (Quranic exegesis), Hadith authentication, and *fiqh* (Islamic jurisprudence), require a depth of nuance and contextual understanding that generic LLMs may lack.

This paper explains why domain-specific LLMs for Islamic studies are both necessary and timely. Models trained on general web data may misinterpret legal

Received April 25th, 2025, 1st Revision September 22nd, 2025, 2nd Revision October 20th, 2025, Accepted for publication October 22nd, 2025.

texts or misrepresent key religious sources. Islamic jurisprudence demands a precise reading of canonical texts, respect for diverse schools of thought, and an ethical commitment to avoid misinformation. For example, when asked a jurisprudential question such as "Is gold liable for *zakat*, and what are the applicable *nisab* and rate?", a general-purpose LLM can produce imprecise or internally inconsistent responses: for instance, implying that zakat on gold is optional or omitting the established nisab and the standard zakat rate (commonly understood as 2.5% per lunar year). Such omissions or inaccuracies may mislead non-expert users seeking legal guidance. In contrast, a domain-adapted model trained on curated jurisprudential corpora and verified sources is better positioned to reproduce doctrinally accurate answers and supply primary citations, thereby reducing the risk of doctrinal error and improving trustworthiness [23].

Errors or biases in an Islamic LLM can directly affect religious practice and understanding. Ethical frameworks in AI, including transparency, fairness, privacy, and accountability are therefore critical [6-9]. Context-specific ethical guidelines, curated data practices, and ongoing scholarly oversight are also essential [10][11].

2 Evolution of Large Language Models

2.1 Evolution of General LLMs

The field of language modeling has progressed through several distinct generations (Figure 1). Early approaches relied on statistical methods and n-gram frequency counts to predict text, grounded in information theory. These were gradually succeeded by neural network models; simple recurrent neural networks (RNNs) and long short-term memory (LSTM) networks allowed sequential processing with some memory of context. However, RNN-based models struggled with long-range dependencies due to vanishing gradients. A breakthrough came with the introduction of the Transformer architecture by Vaswani *et al.* in 2017 [12].

Transformers use self-attention mechanisms to process input in parallel and capture long-distance relationships more effectively. On this foundation, a new wave of large language models emerged. For example, BERT employed a bidirectional Transformer encoder for natural language understanding [1], while GPT used an autoregressive Transformer decoder for text generation [2]. Unified frameworks like T5 recast all NLP tasks into a text-to-text format, further demonstrating the versatility of Transformers [3].

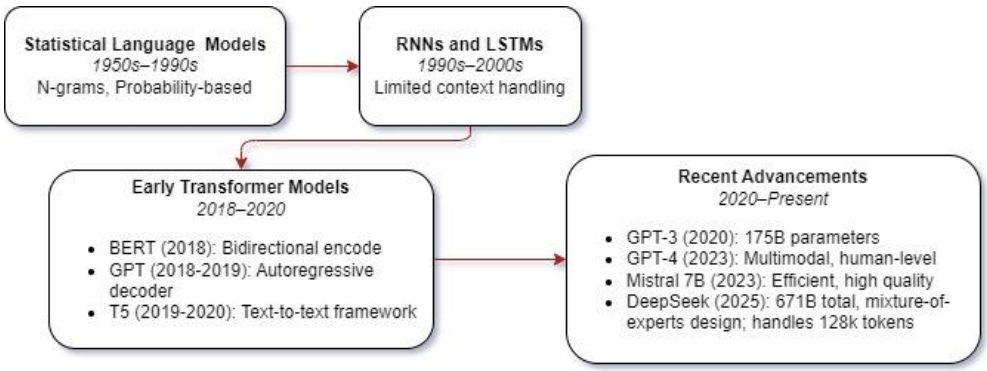

Figure 1 Timeline of the evolution of general LLMs.

As these architectures matured, model scale became a key factor. The number of parameters and training data size grew exponentially in the late 2010s and early 2020s, leading to models like GPT-3 with 175 billion parameters that exhibited emergent abilities in zero-shot learning. This scaling trend culminated in even larger and more capable models such as GPT-4, introduced in 2023, which is a multimodal model capable of processing both text and images and achieving human-level performance on many benchmarks [13].

Alongside proprietary models, open-source efforts have produced efficient LLMs. Mistral 7B, released in 2023, demonstrated that a 7-billion-parameter model can match or exceed the performance of larger models by training on high-quality data [14]. At the extreme end, mixture-of-experts architectures such as DeepSeek, announced in 2025, leverage a massive parameter count (671 billion total) while activating only a subset of 37 billion parameters per query, resulting in state-of-the-art performance and context handling of up to 128k tokens [15]. These latest advancements illustrate the diverse paths in LLM evolution, from efficient small models to large-scale, innovative architectures.

Recent surveys have further highlighted the diverse applications of LLMs in domains such as healthcare, legal reasoning, and education, demonstrating the broad potential of these models when tailored for specific contexts [43][44]. Similar approaches have been adopted in other domains to improve factuality and task performance. For instance, Med-PaLM and BioGPT specialize in medical question answering [45][46], while Legal-BERT has been fine-tuned for legal text classification and contract analysis [47]. These efforts show that domain specialization significantly improves model performance on high-stakes tasks, motivating the development of equally specialized models for Islamic studies.

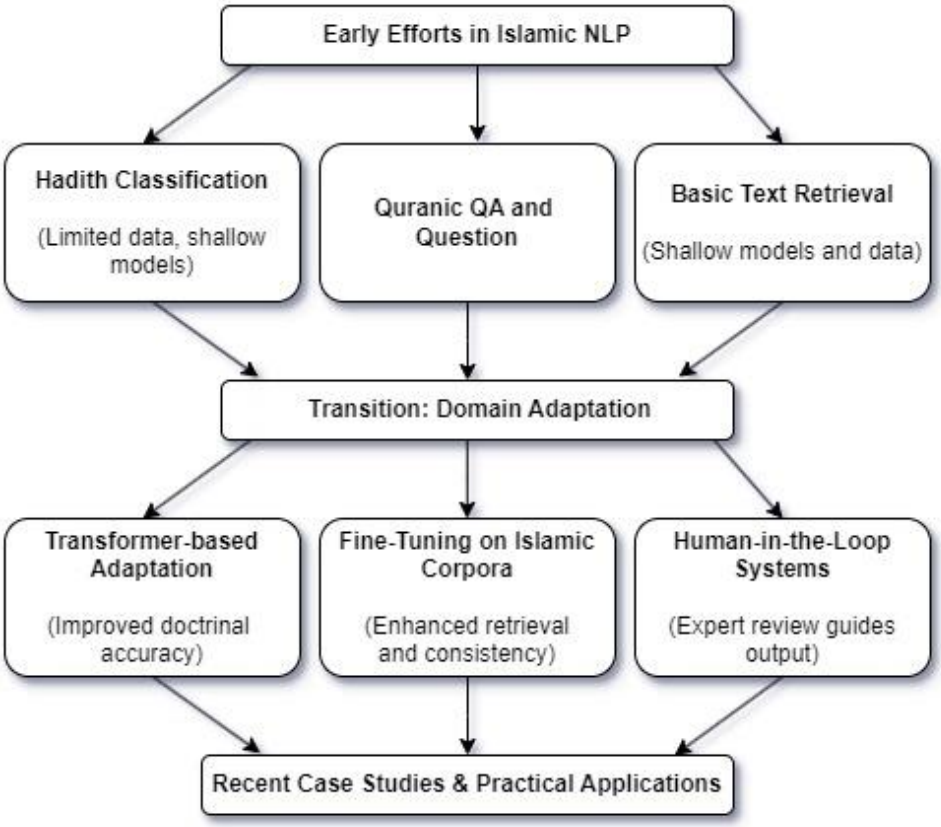

Figure 2 Evolution and practical applications of Islamic LLMs.

2.2 Evolution of Islamic LLMs

The development of LLMs tailored to Islamic studies is promising and growing rapidly. Early computational efforts applied NLP techniques to specific tasks rather than building general-purpose Islamic LLMs.

Researchers initially focused on automated Hadith classification and verification systems, as well as Quranic question-answering systems that leveraged machine learning on limited datasets [4][5]. A survey by Alnefaie *et al.* in [19] cataloged domain-specific QA systems for Arabic Islamic questions, highlighting the feasibility of applying NLP in this domain [16]. Early systems often used shallow models due to data limitations and focused on retrieving and organizing knowledge from canonical texts (see Figure 2).

More recent efforts have adapted modern Transformer architectures to Islamic content. Researchers have fine-tuned models like BERT and GPT on Islamic textual data to generate interpretations or answer jurisprudence questions while attempting to maintain doctrinal accuracy [17]. For example, AlZahrani and Al-Yahia [27] concluded that pretrained transformer models fine-tuned using the Islamic legal dataset, showed significant results in applying authorship attribution to Islamic legal texts. Ibrahim *et al.* [26] introduced computational methods for Hadith authentication by analyzing chains of narrators, while [28] explored AI-based fatwa systems that generate legal opinions under strict oversight [19][20]. In addition, community-driven efforts to digitize and annotate classical Islamic texts have begun to yield large-scale, high-quality datasets that can be used for domain-specific model training [21].

A notable recent case study is the development of a bilingual Islamic LLM for neural search, which used a multi-stage training process starting from a multilingual Transformer such as XLM-R and further pre-trained it on Islamic texts, resulting in a model that outperformed larger generic models on in-domain retrieval tasks [22]. Although these Islamic LLMs are still in early phases of development, they show promising potential to address both linguistic and doctrinal challenges, particularly when integrated with expert feedback and specialized evaluation metrics as described later. While research on Islamic LLMs has grown rapidly, public accessibility remains limited.

Most existing systems, including bilingual Islamic LLMs for neural search [22], Hadith authentication models [26], and AI-based fatwa generation systems [28], are either research prototypes or restricted to academic collaborations. None currently match the level of public availability of large general-purpose models such as ChatGPT [2] or Gemini. Nevertheless, community initiatives such as the Quranic Arabic Corpus and other open digitization efforts [29][31] are creating foundational datasets that may enable future publicly accessible, open-source Islamic LLMs. Until such models are released, access to domain-specific LLMs for Islamic studies is primarily limited to researchers and developers involved in ongoing projects. The following Table 1 provides a summary of key, prominent initiatives based on available information:

Table 1 Key Characteristics of Prominent Islamic LLM Projects and Prototypes.

Project / Model Name	Architecture	Primary Task	Description	Status & Accessibility
Quranic Semantic Embedding Search	Transformer-based embeddings	Semantic search	Focused on meaning-based retrieval of Quranic verses using contextual embeddings instead of keyword matching.	Research prototype; not a conversational LLM [4]
Bilingual Islamic LLM for Neural Search	Multilingual Transformer	Neural search & retrieval	Multi-stage training on Islamic corpora. Outperformed larger generic LLMs in retrieval tasks.	Academic case study; limited accessibility [22]
Hadith Authentication Systems	Transformer models (BERT-based classifiers)	Chain-of-narration analysis & classification	Analyzes isnad (chains of narrators) to verify authenticity, grouping narrations by reliability (sahih, hasan, daif).	Research systems; limited datasets and academic access [26]
Islamic Knowledge Classification System	Transformer-based classifier (AraBERT / BERT)	Text classification & organization	Categorizes large corpora of Islamic texts into thematic categories, supporting better retrieval and annotation.	Published academic system; not publicly deployed [27]
AI-Based Fatwa Generation System	Fine-tuned Transformer + human-in-the-loop	Fatwa draft generation & legal opinion assistance	Generates draft fatwas with citations, reviewed by qualified scholars to ensure doctrinal accuracy.	Prototype under research; expert-supervised use only [28]

Comparative evaluations reported in [22] show that a bilingual Islamic LLM significantly outperformed larger general-purpose models (e.g., mBERT, XLM-R) on in-domain retrieval tasks, achieving higher MRR and MAP scores. Similarly, [27] demonstrated significant improved classification accuracy when AraBERT is fine-tuned on Islamic knowledge categories compared to generic pretrained models. The systems described above represent existing implementations or prototypes. The following sections transition from reporting current efforts to outlining conceptual directions and future design considerations for Islamic LLMs.

3 Islamic Studies and NLP

Islamic studies are founded on a rich textual tradition comprising the Quran, the Hadith (prophetic traditions), and centuries of scholarly commentary. Fiqh

(Islamic jurisprudence) derives legal rulings from the Quran and the Hadith using principles such as *qiyas* (analogy) and *ijmaʿ* (consensus), with various schools of law (Hanafi, Maliki, Shafiʿi, Hanbali, etc.) employing distinct interpretative methods [23][24]. Hadith science involves the meticulous verification of narrations through an analysis of the chain of transmitters and the content, categorizing narrations as *sahih* (authentic), *hasan* (good), *daʿif* (weak), or *mawduʿ* (fabricated) [25]. These processes demand not only a deep understanding of classical Arabic but also contextual and theological expertise.

NLP applications in Islamic studies have been increasingly explored in recent years. One major area is Hadith analysis. Researchers have applied machine learning techniques for automated Hadith classification, grouping narrations by topic or authenticity. Methods include both content-based approaches (processing the text itself) and chain-based approaches (analyzing transmission metadata). Some systems have built narration graphs to track relationships among narrators, thereby assisting scholars in verifying authenticity [26]. For the Quran, NLP techniques have been applied to develop specialized morphological analyzers, part-of-speech taggers, and syntactic parsers tailored to Quranic Arabic, which significantly differs from modern standard Arabic. Additionally, several studies have developed systems for automatic tafsir assistance, retrieving classical commentaries relevant to specific verses.

Advances in Arabic NLP have underpinned many of these efforts. The development of Arabic-specific models such as AraBERT has resulted in improved performance on tasks like question answering, named entity recognition, and sentiment analysis compared to earlier multilingual models [28]. The availability of annotated resources such as the Quranic Arabic Corpus has further supported these applications [29]. However, while general NLP systems typically deal with open-domain language, Islamic NLP must address additional challenges. Data availability is limited due to the relatively small size of annotated Islamic texts compared to vast general-domain datasets. Moreover, many Islamic texts exhibit complex linguistic features, including honorifics and archaic expressions, which are not present in modern texts. Thus, while NLP in general can rely on large-scale, well-annotated corpora, Islamic NLP must often work with highly specialized and limited datasets. This necessitates the development of domain-specific models and benchmarks that ensure both linguistic fluency and doctrinal accuracy. In practice, most Islamic LLMs adopt encoder-decoder or decoder-only Transformer architectures similar to BERT or GPT variants, with domain adaptation performed through continued pretraining on Islamic corpora followed by supervised fine-tuning on QA, classification, and retrieval tasks [17][22][26]. Parameter-efficient techniques such as LoRA and adapters are often employed to reduce compute cost while preserving doctrinal fidelity [40].

4 Challenges and Ethical Considerations in Islamic LLMs

Building a domain-specific LLM for Islamic studies presents a unique set of technical and ethical challenges (Figure 3). A primary challenge is data scarcity and quality. Although digitization projects have made many canonical texts available, the volume of high-quality, annotated Islamic texts is modest compared to general web corpora. Many historical manuscripts remain undigitized or inconsistently annotated, particularly in the case of hadith chains where details about transmitters are crucial [30]. Moreover, while the Arabic Quran and primary hadith collections are well-represented, resources in other languages are limited, complicating efforts to create multilingual models [31]. Another significant challenge is bias and representation. General LLMs trained on internet data can absorb societal and cultural biases. In the context of Islamic texts, a model might inadvertently learn orientalist misrepresentations, favor one jurisprudential perspective over another, or reflect other demographic biases. Since Islamic scholarship is inherently pluralistic, it is ethically imperative to design models that do not privilege one interpretation at the expense of others [32]. Mitigating bias requires curating a balanced dataset representing various schools of thought and employing techniques such as adversarial training and bias regularization. It is also essential to integrate expert feedback loops so that scholars can review and correct model outputs. Doctrinal accuracy and ethical use are principal concerns. In Islamic applications, hallucinations or fabrications, common issues in general LLMs, can have severe consequences. A model that produces incorrect hadiths or misinterprets a Quranic verse could mislead users in matters of faith. To prevent this, models must be constrained to rely on verified sources and their outputs should include citations from authentic texts. Furthermore, Islamic questions often have multiple valid interpretations; a one-size-fits-all answer is unacceptable. The system should ideally acknowledge diverse scholarly opinions rather than presenting a single, definitive ruling [33].

Addressing these challenges requires a combination of technical strategies and governance frameworks. On the technical side, data scarcity can be alleviated by targeted data augmentation and domain-specific corpus expansion. For instance, synthetic data generation and expert-verified translations can enrich existing datasets. Techniques such as transfer learning and continued pre-training on domain-specific texts (e.g., using models like XLM-R further trained on Islamic corpora) have shown promise [34]. To mitigate bias, curated data and adversarial training are essential, along with periodic audits by subject-matter experts. Human-in-the-loop systems, as proposed in [30], can help refine outputs in real time [35]. On the governance side, an ethical framework tailored to Islamic AI is critical. This framework should incorporate general AI ethics principles: beneficence, non-maleficence, autonomy, justice, and explicability, along with Islamic ethical values such as *maslaha* (welfare) and *amanah* (trust). An

oversight board composed of Islamic scholars and AI experts can periodically audit the system and provide guidelines for acceptable output. Transparency is also essential. End-users should be clearly informed that the model is an AI tool, with limitations and potential biases, and that final religious decisions should be verified by qualified scholars.

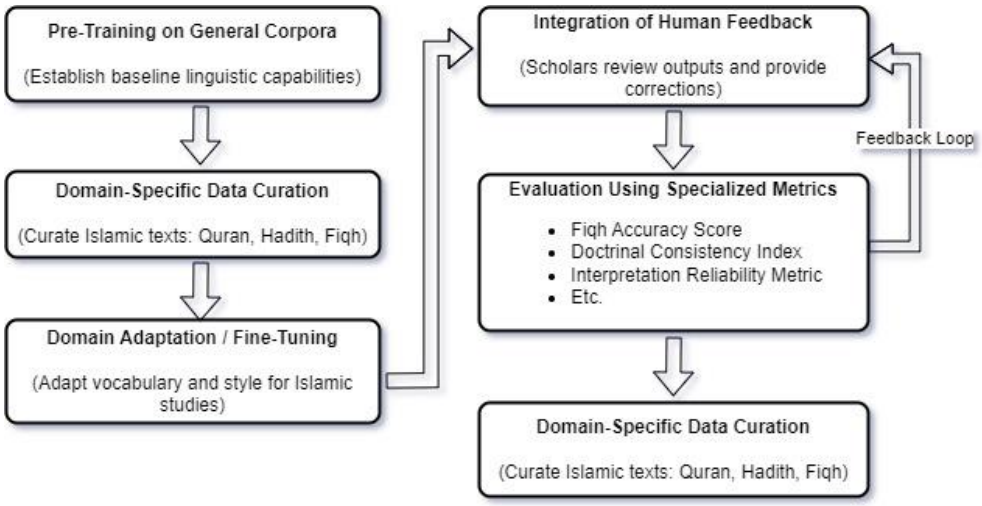

Figure 3 Workflow diagram for developing an Islamic LLM.

Addressing multilingualism is crucial, as many Muslim-majority regions rely on non-Arabic languages such as Urdu, Bahasa Indonesia, and Hausa. Current efforts include multilingual continued pretraining of XLM-R models and crowdsourced translation projects to expand parallel Islamic corpora [22][31]. Future research should prioritize building benchmark datasets across these languages to ensure inclusivity. Finally, specialized evaluation metrics, discussed in Section 5, are proposed as potential solutions to measure doctrinal fidelity and consistency. These metrics, namely Fiqh Accuracy Score, Doctrinal Consistency Index, and Interpretation Reliability Metric, can provide quantitative feedback during development and refinement. They are intended to guide future research and further testing in collaboration with Islamic scholars to ensure that the LLM's outputs meet the high standards required in religious contexts.

5 Potential Impact and Evaluation Metrics for an Islamic LLM

This section examines two interconnected aspects. First, we detail three specialized evaluation metrics designed to capture the unique requirements of Islamic textual interpretation. Second, we discuss the potential impact of an Islamic LLM on education, legal consultation, and academic research. We also

outline a validation strategy to empirically test these metrics and ensure their robustness. Finally, we also present a validation framework to ensure these metrics are empirically reliable. This section introduces a forward-looking framework. The architecture and governance mechanisms discussed here are proposed constructs intended to guide future development rather than descriptions of deployed systems.

5.1 Specialized Evaluation Metrics

Standard NLP metrics such as perplexity and BLEU [21] are widely used to assess language models. Perplexity measures how well a model predicts word sequences, while BLEU evaluates the overlap of n-grams between generated text and reference texts. Both metrics provide useful insight into linguistic fluency and surface-level similarity. However, they focus mainly on syntactic and lexical accuracy and do not capture deeper semantic meanings, doctrinal nuances, and ethical responsibilities.

In Islamic studies, texts carry layers of meaning that require strict adherence to established interpretations. An LLM in this domain must align with the doctrinal standards of various fiqh schools and respect the ethical guidelines endorsed by recognized scholars. Relying solely on perplexity and BLEU risks overlooking errors in theological accuracy or subtle interpretative shifts. This gap has led to the development of specialized metrics such as the Fiqh Accuracy Score, Doctrinal Consistency Index, and Interpretation Reliability Metric to ensure that models are evaluated on both language quality and doctrinal precision. The proposed framework is designed for interactive dialogue, enabling users to explore multiple jurisprudential opinions when ambiguity exists, rather than presenting a single static answer.

5.1.1 Fiqh Accuracy Score

The Fiqh Accuracy Score quantifies how closely a model's output aligns with established Islamic jurisprudence. It measures doctrinal alignment by comparing generated interpretations against a curated gold-standard dataset of scholarly opinions. It also evaluates contextual fidelity across various fiqh schools (e.g., Hanafi, Maliki, Shafiʿi, Hanbali) through expert ratings and automated semantic similarity measures. This score is crucial during both model training and quality assurance to ensure that outputs adhere to recognized religious interpretations [25][29].

5.1.2 Doctrinal Consistency Index

The Doctrinal Consistency Index assesses the internal coherence of an LLM's outputs across related queries. It focuses on ensuring that the model consistently

applies doctrinal principles when responding to similar or follow-up questions. The metric evaluates cross-context consistency using statistical divergence measures and expert review to verify uniformity in interpretations. Maintaining such consistency prevents contradictory or mixed doctrinal outputs and builds user trust [28][30].

5.1.3 Interpretation Reliability Metric

The Interpretation Reliability Metric evaluates the trustworthiness of the model's outputs relative to recognized scholarly opinions. It is based on expert agreement scores, where domain experts rate the generated interpretations, and on confidence scoring mechanisms that the model may internally assign. The metric also monitors longitudinal reliability to ensure that model performance remains stable over time and after updates. This dual approach provides a robust quantitative measure and guides iterative improvements [10][11][30].

5.1.4 Empirical Validation of Specialized Metrics

To ensure that the Fiqh Accuracy Score, Doctrinal Consistency Index, and Interpretation Reliability Metric are applicable and reliable, we propose a multi-phase empirical validation strategy:

1. Gold-Standard Benchmark Creation: A balanced, multi-*madhhab* benchmark dataset will be curated, containing questions and answers drawn from the Quran, canonical hadith collections, and fiqh references [23][25][29]. Each item will include contextual metadata (time, place, custom) to capture jurisprudential nuance, following best practices for domain-specific benchmark construction [40].
2. Expert Annotation and Reliability Analysis: Multiple qualified scholars per madhhab will independently rate doctrinal alignment and contextual fidelity. Inter-rater reliability will be computed using Krippendorff's α and Cohen's κ [41], ensuring statistically robust agreement before using these annotations as ground truth.
3. Convergent Validity Testing: Metric scores will be compared with expert judgments and task-level performance indicators such as retrieval MRR/MAP and QA F1 [22]. Strong and statistically significant correlations will provide evidence that the metrics reflect expert judgments of doctrinal and contextual correctness.
4. Discriminant Validity via Ablations: Controlled ablation studies (e.g., disabling source citation, removing madhhab conditioning) will verify that the metrics are sensitive to doctrinal errors, showing measurable degradation when key fidelity-preserving features are removed [18].
5. Cross-Domain Generalizability: The metrics will be applied across multiple task types (tafsir retrieval, hadith classification, fiqh QA) [19][26], and across

Arabic and English subsets to confirm that they generalize beyond a single dataset or language setting.
6. Temporal Stability and Drift Monitoring: Successive model versions will be re-evaluated using bootstrap confidence intervals [42] to monitor longitudinal consistency and detect regressions over time.
7. Iterative Human-in-the-Loop Refinement: Divergences between metric outputs and expert ratings will be systematically reviewed. Using an adaptive governance approach [30], metric definitions and weightings will be iteratively refined until alignment with scholarly judgment is maximized.

This validation approach transforms the proposed metrics into empirically grounded, reproducible measures, providing a robust foundation for continuous model development and scholarly oversight.

5.2 Impact on Education, Legal Consultation, and Research

A well-designed Islamic LLM incorporating these specialized metrics can benefit multiple areas. In Islamic education, students often seek instant clarification on Quranic verses or Hadith references. A model trained on curated texts can provide accurate context and citations, thereby supplementing traditional learning methods [35][36]. In legal or fiqh consultation, semi-automated systems could help muftis and legal experts rapidly compile relevant Quranic verses, Hadith, and scholarly opinions. To enhance scalability, expert-in-the-loop pipelines can incorporate active learning, prioritizing only the most uncertain cases for human review, thereby reducing annotation load while maintaining doctrinal reliability [30].

The Doctrinal Consistency Index ensures that legal reasoning remains uniform across interpretations, while the Interpretation Reliability Metric strengthens confidence in the outputs. These systems can also highlight minority or less-cited perspectives, fostering balanced legal guidance [37][38]. Additionally, academic researchers can use such systems for cross-school comparative studies and thematic analysis, thus bridging the gap between classical texts and modern scholarship [39]. Clear disclaimers must accompany these outputs to stress that final religious judgments remain to be validated by knowledgeable scholars [10], [11].

In addition, a representative Islamic AI oversight board is envisaged to supervise both dataset curation and model evaluation. Such a board could include scholars from each of the four major Sunni madhahib, AI ethicists, and technical experts. Decisions would aim for consensus, and where divergence persists, majority and minority positions could be documented, allowing model outputs to present multiple valid interpretations transparently. This approach would operationalize doctrinal plurality while maintaining accountability in model governance.

6 Conclusion

This paper has shown that domain-specific LLMs are necessary to address the complexity, sensitivity, and diversity inherent in Islamic studies. General-purpose models, while fluent, may miss crucial jurisprudential details and ethical dimensions, which can lead to misinformation. By reviewing the evolution of LLMs, detailing the intricacies of Islamic jurisprudence, and outlining the challenges in data curation and bias mitigation, we make a strong case for specialized Islamic LLMs.

Recent advances, as seen in models such as GPT-4, combine context sensitivity with improved doctrinal precision. However, challenges like data scarcity, bias, and ethical responsibility remain. Continued collaboration between AI practitioners and Islamic scholars is essential. A robust Islamic LLM with rigorous scholarly oversight could transform education, legal research, and scholarly collaboration. Future research should focus on refining methodological frameworks, developing tailored evaluation metrics (including the Fiqh Accuracy Score, Doctrinal Consistency Index, and Interpretation Reliability Metric), and validating real-world use cases through comprehensive user feedback.

References

[1] Devlin, J., Chang, M., Lee, K. & K. Toutanova, *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,* in Proc. NAACL-HLT, pp. 4171-4186, 2019. DOI: 10.18653/v1/N19-1423.

[2] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. & Sutskever, I., *Language Models are Unsupervised Multitask Learners,* OpenAI technical report, 2019. (PDF: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf).

[3] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. & Liu, P.J., *Exploring the limits of transfer learning with a unified text-to-text transformer*. J. Mach. Learn. Res., **21**(1), pp. 5485-5551, 2020. DOI: 10.48550/arXiv.1910.10683.

[4] Tariq, M., Awan, M.A. & Khaleeq, D., *Developing a Quranic QA System: Bridging Linguistic Gaps in Urdu Translation using NLP and Transformer Model.* IJIST, **7**(1), pp. 493-505. Mar, 2025.

[5] Kamali, M.H., *Principles of Islamic Jurisprudence*, 3rd ed. Cambridge, U.K.: Islamic Texts Society, 2003. ISBN: 9780946621811. ark:/13960/t8pd3br6c

[6] Hagendorff, T., *The Ethics of AI Ethics: An Evaluation of Guidelines,* Minds Mach., **30**(1), pp. 99-120, 2020. DOI: 10.1007/s11023-020-09517-8.

[7] Falletti, E., *Algorithmic Discrimination and Privacy Protection.* JDTL, **1**(2), pp. 387-420, 2023. DOI: 10.21202/jdtl.2023.16.

[8] Mittelstadt, B., *Principles Alone Cannot Guarantee Ethical AI,* Nat. Mach. Intel., **1**, pp. 501-507, 2019. DOI: 10.1038/s42256-019-0114-4.

[9] Brundage, M., Avin, Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., Khlaaf, H., Yang, J., Toner, H., Fong, R., Maharaj, T., Koh, P.W., Hooker, S., Leung, J., Trask, A., Bluemke, E., Lebensold, J., O'Keefe, C., Koren, M., Ryffel, T., Rubinovitz, J.B., Besiroglu, T., Carugati, F., Clark, J., Eckersley P., de Haas, S., Johnson, M., Laurie, B., Ingerman, A., Krawczuk, I., Askell, A., Cammarota, R., Lohn, A., Krueger, D., Stix, C., Henderson, P., Graham, L., Prunkl, C., Martin, B., Seger, E., Zilberman, N., Ó hÉigeartaigh, S., Kroeger, F., Sastry, G., Kagan, R., Weller, A., Tse, B., Barnes, E., Dafoe, A., Scharre, P., Herbert-Voss, A., Rasser, M., Sodhani, S., Flynn, C., Gilbert, T.K., Dyerm L., Khan, S., Bengio, Y. & Anderljung, M., *Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims*, arXiv e-prints, Art. no. arXiv:2004.07213, 2020. DOI: 10.48550/arXiv.2004.07213.

[10] Bryson, J., *Patiency is Not a Virtue: The Design of Intelligent Systems and Systems Of Ethics,* Ethics Inf. Technol., **20**, pp. 15-26, 2018. DOI: 10.1007/s10676-018-9448-6.

[11] Berendt, B., *AI for the Common Good?! Pitfalls, Challenges, and Ethics Pen-Testing,* Paladyn, J. Behav. Robot., **10**(1), pp. 44-65, 2019. DOI: 10.1515/pjbr-2019-0004.

[12] Mikolov, T., Chen, K., Corrado, G. & J. Dean, *Efficient Estimation of Word Representations in Vector Space,* in Proc. ICLR, 2013. DOI: 10.48550/arXiv.1301.3781.

[13] Pennington, J., Socher, R. & Manning, C. *GloVe: Global Vectors for Word Representation,* in Proc. Empir. Methods Nat. Lang. Process., (EMNLP), pp. 1532-1543, Doha, Qatar, 2014. DOI: 10.3115/v1/D14-1162.

[14] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. & Polosukhin, I., *Attention Is All You Need*, in Proc. 31st Int. Conf. Neural Inf. Process. Syst. (NIPS), pp. 5998-6008, Long Beach, CA, USA, 2017. DOI: 10.48550/arXiv.1706.03762.

[15] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S.,. So, C.H & Kang, J., *BioBERT: A Pre-Trained Biomedical Language Representation Model for Biomedical Text Mining,* Bioinformatics, **36**(4), pp. 1234-1240, 2020. DOI: 10.1093/bioinformatics/btz682.

[16] Alsentzer, E., Murphy, J., Boag, W., Wei, W.-H., Jin, D., Naumann, T. & McDermott, M., *Publicly Available Clinical BERT Embeddings,* ClinicalNLP (NAACL), 2019. DOI: 10.48550/arXiv.1904.03323.

[17] Huang, A.-H., Wang, H. & Yang, Y., *FinBERT - A Large Language Model for Extracting Information from Financial Text,* Cont. Acc. Res., 2020. DOI: 10.2139/ssrn.3910214.

[18] Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H. & Neubig, G., *Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing,* ACM Comput. Surv., **55**(9), pp. 1-35, 2023. DOI: 10.1145/3560815.

[19] Alnefaie, S., Atwell, E. & Alsalka, M.A., *Islamic Question Answering Systems Survey and Evaluation Criteria,* Int. J. Islam. App. CS & Tech. (IJIACST), **11**(1), pp. 9-18, 2023. oai:eprints.whiterose.ac.uk:206757.

[20] Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I., *Improving Language Understanding by Generative Pre-training,* OpenAI, 2018.

[21] Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J., *BLEU: A Method for Automatic Evaluation of Machine Translation,* in Proc. 40th Annu. Meet. Assoc. Comput. Linguist. (ACL), pp. 311-318, 2002. DOI: 10.3115/1073083.1073135.

[22] Zhou, Y. & Srikumar, V., *A Closer Look at How Fine-tuning Changes BERT,* in Proc. 60th Ann. Meet. ACL Conf., pp. 1046-1061, Dublin, Ireland, 2022. DOI: 10.18653/v1/2022.acl-long.75.

[23] Hallaq, W.B., *An Introduction to Islamic Law*. New York: Cambridge University Press, 2009. DOI: 10.1017/CBO9780511801044.

[24] Kamali, M.H., *Shariah Law: An Introduction,* Oneworld Publications, 2008. ISBN 978-1-85168-565-3.

[25] Binbeshr, F., Kamsin, A. & Mohammed, M., *A Systematic Review on Hadith Authentication and Classification Methods,* ACM Trans. Asian Low-Resour. Lang. Inf. Process., **20**(2), pp 1-17, 2021. DOI: 10.1145/3434236.

[26] Ibrahim, N.K., Noordin, M.F., Samsuri, S., Abu Seman, M.S. & Ali, A.E.B., *Isnad Al-Hadith Computational Authentication: An Analysis Hierarchically,* Int. Conf. Inf. & Comm. Tech. Muslim World (ICT4M), Jakarta, Indonesia, pp. 344-348, 2016. DOI: 10.1109/ICT4M.2016.075.

[27] AlZahrani F.M. & Al-Yahya, M., *A Transformer-based Approach to Authorship Attribution in Classical Arabic Texts."* Appl. Sci., **13**(12), 2023. DOI: 10.3390/app13127255.

[28] Usmonov, M., *From Human Scholars to AI Fatwas: Media, Ethics, and the Limits of AI in Islamic Religious Communication.* J. Cntm. Isl. Comm., **5**(1), 2025. DOI: 10.33102/jcicom.vol5no1.125.

[29] Kathir, I., *Tafsir Ibn Kathir* (abridged), Translated by Shaykh Ṣafī al-Raḥmān al-Mubārakpūrī. 10 vols. Riyadh: Darussalam, 2000.

[30] Chen, X., Wang, X. & Qu, Y., *Constructing Ethical AI based on the "Human-in-the-Loop" System,* Systems, **11**(11), 2023. DOI: 10.3390/systems11110548.

[31] Abdoh, E., *Utilizing Modern Technology for the Preservation of Ancient Manuscripts and Rare Books: The Digitization Project at King Abdulaziz Complex for Endowment Libraries as a Model.* Restaurator. Int. J. Pres. Lib. Arch. Mat., **46**(1), pp. 35-58, 2025. DOI: 10.1515/res-2024-0016.

[32] Ghozali, N.I.M., Mansor, N.S., Awang, H. & Yusof, S.M., *Automated Translation Tools for Mualaf: Artificial Intelligence Solutions for Accessing Islamic Texts and Resources Across Languages.* Int. J. Islamic Theo. & Civ., **2**(3), pp. 60-64, 2024, DOI: 10.5281/zenodo.13943175.

[33] Garg, N., Schiebinger, L., Jurafsky, D. & Zou, J., *Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes,* Proc. Natl. Acad. Sci. U.S.A., **115**(16), pp. E3635-E3644, 2018. DOI: 10.1073/pnas.1720347115.

[34] Floridi, L. & Cowls, J., *A Unified Framework of Five Principles for AI in Society,* Harvard Data Sc. Rev., **1**(1), 2019. DOI: 10.1162/99608f92.8cd550d1.

[35] Alhammad, N., Awae, F. Yussuf, A., Al-Awami, A.Y.M., Hayiwaesorhoh, M. & Chehama, A., *Using E-learning Platforms in Teaching Islamic Education,* J. Islam. Edu. Research, **11**(1), pp. 63-69, 2025. DOI: 10.22452/jier.vol11no1.6.

[36] Chukhanov, S. & Kairbekov, N., *The Importance of a Semantic Approach in Understanding the Texts of the Holy Quran and Sunnah,* Pharos J. Theo., **105**(3), pp. 1-11, 2024. DOI: 10.46222/pharosjot.105.36.

[37] Mustafa, M. & Agbaria, A.K., *Islamic Jurisprudence of Minorities (Fiqh al-Aqalliyyat): The Case of the Palestinian Muslim Minority in Israel*, J. Musl. Min. Aff., **36**(2), pp. 184-201, 2016. DOI: 10.1080/13602004.2016.1180889.

[38] Osman, A.M.S., *Adopting Comparative Fiqh Methodology in Islamic Jurisprudence: Facing Contemporary Challenges with Ethical Considerations.* Al-Mazaahib, **11**(2), pp. 115-138. 2023. DOI: 10.14421/al-mazaahib.v11i2.3203.

[39] AlSajri, A., *Challenges in Translating Arabic Literary Texts using Artificial Intelligence Techniques,* EDRAAK, **2023**, pp. 5-10, 2023. DOI: 10.70470/EDRAAK/2023/002.

[40] Ethayarajh, K., & Jurafsky, D., *Utility is in the Eye of the User: A Critique of NLP Leaderboards,* In Proc. Emp. Meth. NLP (EMNLP), pp. 4846-4853, 2020. DOI: 10.18653/v1/2020.emnlp-main.393.

[41] Cohen, J., *A Coefficient of Agreement for Nominal Scales,* Educ. Psychol. Meas., **20**(1), pp. 37-46, 1960. DOI: 10.1177/001316446002000104.

[42] Efron B. & Tibshirani, R., *Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy,* Statist. Sci., **1**(1), pp. 54-75, 1986. DOI: 10.1214/ss/1177013815.

[43] Hang, C.N., Yu, P.-D. & Tan, C.W., *TrumorGPT: Graph-based Retrieval-augmented Large Language Model for Fact-checking,* in IEEE Trans. on AI, pp. 1-15, 2025. DOI: 10.1109/TAI.2025.3567369.

[44] Zeng, J., Dai, Z., Liu, H., Varshney, S., Liu, Z., Luo, C., Li, Z., He, Q. & Tang, X., *Examples as the Prompt: A Scalable Approach for Efficient LLM*

Adaptation in E-Commerce, In Proc. Int. ACM SIGIR Conf. on R&D in Info. Ret. (SIGIR). New York, NY, USA, pp. 4244-4248, 2025. DOI: 10.1145/3726302.3731941.

[45] Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., y Arcas, B.A., Webster, D., Corrado, G.S., Matias, Y., Chou, K., Gottweis, J., Tomasev, N., Liu, Y., Rajkomar, A., Barral, J., Semturs, C., Karthikesalingam, A. & Natarajan, V., *Large Language Models Encode Clinical Knowledge*, Nature, **620**, pp. 172-180, 2023. DOI: 10.1038/s41586-023-06291-2.

[46] Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H. & Liu, T.-Y., *BioGPT: Generative Pre-trained Transformer for Biomedical Text Generation & Mining*, Briefings in Bioinformatics, **23**(6), 2022. DOI: 10.1093/bib/bbac409.

[47] Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N. & Androutsopoulos, I., *Legal-BERT: The Muppets Straight Out of Law School,* Proc. of Findings of ACL: EMNLP, pp. 2898-2904, 2020. DOI: 10.18653/v1/2020.findings-emnlp.261.