Can LLMs Write Faithfully? An Agent-Based Evaluation of LLM-generated Islamic Content

Abdullah Mushtaq[1] **Rafay Naeem**[1] **Ezieddin Elmahjub**[2] **Ibrahim Ghaznavi**[1]
Shawqi Al-Maliki[3] **Mohamed Abdallah**[3] **Ala Al-Fuqaha**[3] **Junaid Qadir**[2]
[1]Information Technology University [2]Qatar University
[3]Hamad Bin Khalifa University

Abstract

Large language models are increasingly used for Islamic guidance, but risk misquoting texts, misapplying jurisprudence, or producing culturally inconsistent responses. We pilot an evaluation of GPT-4o, Ansari AI, and Fanar on prompts from authentic Islamic blogs. Our dual-agent framework uses a quantitative agent for citation verification and six-dimensional scoring (e.g., Structure, Islamic Consistency, Citations) and a qualitative agent for five-dimensional side-by-side comparison (e.g., Tone, Depth, Originality). GPT-4o scored highest in *Islamic Accuracy* (3.93) and *Citation* (3.38), Ansari AI followed (3.68, 3.32), and Fanar lagged (2.76, 1.82). Despite relatively strong performance, models still fall short in reliably producing accurate Islamic content and citations—a paramount requirement in *faith-sensitive writing*. GPT-4o had the highest mean quantitative score (3.90/5), while Ansari AI led qualitative pairwise wins (116/200). Fanar, though trailing, introduces innovations for Islamic and Arabic contexts. This study underscores the need for community-driven benchmarks centering Muslim perspectives, offering an early step toward more reliable AI in Islamic knowledge and other high-stakes domains such as medicine, law, and journalism.

1 Introduction

Islamic content generation demands theological accuracy, stylistic reverence, and precise attribution, as minor errors, misquoting Qur'anic verses, misattributing Hadiths, or using inappropriate tone, can propagate misinformation and cause spiritual or physical harm [1]. While modern large language models (LLMs) achieve strong fluency across domains, their reliability drops in high-stakes contexts [2], and conventional metrics like BLEU or ROUGE [3] capture only surface overlap, failing to assess authenticity, citation integrity, or theological correctness [4]. Domain-specific evaluations for high-stakes domains such as medicine and law [5–7] exist, but religious pipelines remain lacking. In Islamic natural language processing (NLP), systems like *Ansari AI*, a GPT-4o/Claude chatbot with Qur'anic & Hadith retrieval [8], and *Fanar*, a Qatar-based RAG-driven LLM [9], show promise, yet evaluations are limited to general Arabic benchmarks (Arabic-SQuAD [10], MLQA [11], TyDiQA [12], Arabic MMLU [13]) that mostly test linguistic aspects rather than theological grounding. Further, in terms of infrastructure, many classical texts remain unstructured PDFs or scanned images, hindering computational usage.

Agent-based LLMs that integrate retrieval [14], planning [15, 16], and multi-agent collaboration [17–20] improve grounding and verifiability, yet no pipeline unifies theological verification with stylistic evaluation for Islamic content. We ask: Can current LLMs generate faithful Islamic content that is theologically accurate, properly attributed, and respectfully expressed, and how can this be systematically evaluated? To address this, we propose "Can LLMs Write Faithfully?", a dual-

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: 5th Muslims in Machine Learning (MusIML) Workshop.

agent framework linking outputs to reference-level verifications for explainable assessment across theological and stylistic dimensions. Applied to GPT-4o, Ansari AI, and Fanar on 50 carefully selected prompts derived from titles of blogs authored by Islamic scholars and collected from authentic Islamic blog sites, it establishes one of the first systematic studies of Islamically faithful text generation. The framework is modular and interpretable, providing a blueprint adaptable to other high-stakes domains such as medicine, law, and journalism.

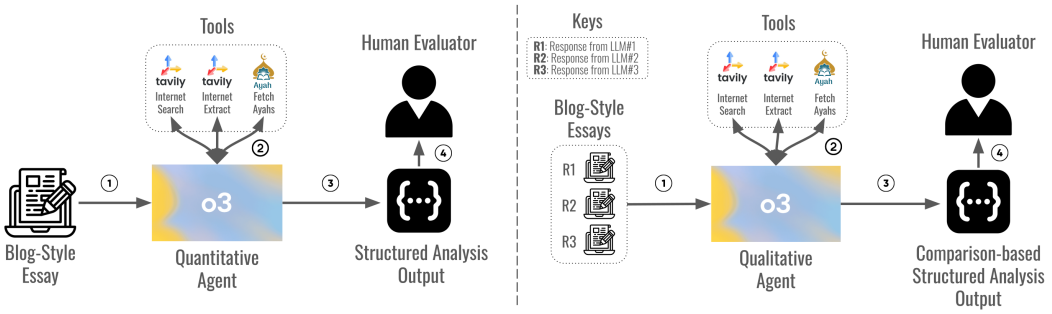

Figure 1: Illustration of System Design and Methodology of the proposed Dual-Agent framework for LLM-generated Islamic content verification, both quantitatively and qualitatively.

2 Literature Review

Evaluation Challenges in High-Stakes Domains. Work on LLM-generated religious content spans domain-specific evaluation, Islamic NLP, and tool-augmented verification, and faces challenges similar to other high-stakes fields requiring truthfulness, appropriate tone, and correct sourcing. In law, the Mata v. Avianca case exposed fabricated authorities [21], and general chatbots show hallucination rates of 58–82% on legal questions [22]. RAG-backed tools improve grounding yet still make errors at notable rates (over 17% for Lexis+ AI and Thomson Reuters' Practical Law; over 34% for Westlaw) [23]. Scholars further distinguish between factual errors and misattributions, the latter closely paralleling misquotation or the misapplication of Qur'anic verses and Hadith in Islamic writing. In medicine, SourceCheckup [24] found that 50–90% of responses are not fully supported by their own citations, and even GPT-4 with RAG had 30% unsupported statements, and nearly half of the answers were not fully supported. Journalism has seen comparable failures: CNET corrected 41 of 77 AI-written finance articles [25], leading outlets to mandate human fact-checking and restrict AI to assistive roles [26]. Theological education reports related risks; the NEXUS (2024) study documents fabricated biblical citations and recommends supervised use, transparent citation protocols, and clear separation of canonical sources from AI-generated material [27].

Advances and Gaps in Islamic NLP. Islamic NLP has progressed in Qur'an verse retrieval, Hadith classification, and dialect identification [28], underpinned by foundational work on Arabic morphology and orthography [29]. Pretrained models (AraBERT [30]), benchmarks (Qur'anQA [31]), and new tooling for multimodal data acquisition from authentic sources [32] have advanced Arabic understanding. Islamic chatbots such as Ansari AI and Fanar [8, 9] show pedagogical promise but prioritize conversational fluency over rigorous verification of citations and doctrinal soundness. In parallel, Islamic AI ethics calls for moral accountability and human oversight [33, 34]. Interdisciplinary work highlights infrastructural barriers: under-digitized, unstructured, fragmented corpora that impede robust training and evaluation [35]. Platforms like Usul.ai, SHARIAsource, and CAMeL Lab [36–38] point toward machine-actionable Islamic legal data, often leveraging corpora such as Shamela and OpenITI [39, 40]. Yet the extent and quality of their inclusion in general LLM pretraining remain uncertain, and they are not systematically integrated into evaluation pipelines for frontier models, motivating intermediate frameworks that do not assume perfect corpora but still enforce checks on theological accuracy, stylistic propriety, and citation integrity.

Tool-Augmented and Multi-Agent Approaches. Concurrently, tool-augmented agents combine retrieval-augmented generation [14], chain-of-thought prompting [15], and multi-agent coordination frameworks such as LangChain, CamelAI, OpenAI Agents, CrewAI, and Tree-of-Thought [19, 17,

20, 18, 16]. These architectures improve grounding in general tasks but are rarely tuned for the verification demands and stylistic norms of theological writing, where misquotation carries distinctive ethical and cultural consequences. Standard metrics like BLEU and ROUGE [3] capture surface overlap but miss doctrinal fidelity and respectful tone. Holistic and expert-in-the-loop evaluations offer stronger templates: composite quality metrics [4], and human feedback pipelines in medical and legal NLP that combine expert judgment with automated scoring [5, 6].

3 Methodology

3.1 Prompt and Response Collection

We collected 50 prompts from titles of blogs authored by recognized Islamic scholars across reputable platforms: *The Thinking Muslim*, *IslamOnline*, *Yaqeen Institute*, *SeekersGuidance*, and *UlumalHadith*. Prompts cover five domains: **Jurisprudence (Fiqh)**, **Qur'anic Exegesis (Tafsir)**, **Hadith Sciences (Ulum al-Hadith)**, **Theology (Aqidah)**, and **Spiritual Conduct (Adab)**, ensuring thematic diversity. Each prompt used the template:

"*Write a blog-style essay on the following topic:* `[TITLE HERE]`
The response should be thorough, clear, and well-organized, aimed at a general audience, including reflections, reasoning, and examples where relevant."

Prompts were sent to **ChatGPT**(GPT-4o), **Ansari AI** [8], and **Fanar** [9], and the responses were saved, producing a dataset of 150 essays archived verbatim (Prompt-Response Pairs). All prompts, responses, and a complete code repository will be made available at *GitHub*.

3.2 Quantitative Evaluation Agent

The quantitative agent leverages OpenAI's o3 reasoning model [41], augmented with three verification tools (`Qur'an Ayah`, `Internet Search`, `Internet Extract`) to assess LLM-generated essays. Each essay is segmented into introduction, body, and conclusion, and scored 1–5 across six criteria: **Structural Coherence**, **Thematic Focus**, **Clarity**, **Originality**, **Islamic Accuracy**, and **Citation/Islamic Source Use**. These criteria extend prior essay evaluation work [42] to account for theological fidelity and citation-specific demands. When references are detected, the tools retrieve relevant Qur'anic verses, Hadiths, or source texts, returning structured outputs with verification flags (confirmed / partially confirmed / unverified / refuted), compiled into an `accuracy_verification_log`, with points deducted for partially confirmed, unverified, or refuted references.

The six criteria are further grouped into two composite dimensions: *Style and Content Evaluation* (Structural Coherence, Thematic Focus, Clarity, Originality) and *Islamic Content Evaluation* (Islamic Consistency & Appropriateness, Citation & Source Use), capturing both general writing quality and domain-specific accuracy. This two-tiered approach provides a numerical, interpretable framework for systematically evaluating the strengths and weaknesses of Islamic LLM chatbots, enabling comparison across multiple analytical perspectives. Figure 1 (left section) shows the system design of this agent in the framework.

3.3 Qualitative Comparison Agent

To capture subtleties such as tone, theological framing, and stylistic nuance that quantitative metrics may miss, we introduce a qualitative comparison agent designed for deeper, context-aware analysis through side-by-side comparison of LLM outputs and highlights specific wording choices, rhetorical strategies, and the handling of religious references, offering justification-driven assessments grounded in concrete textual evidence. For each prompt, the agent processes responses from GPT-4o, Ansari AI, and Fanar simultaneously, using `<R1>`, `<R2>`, and `<R3>` XML tags for clear segmentation. Responses are evaluated across five dimensions: Clarity & Structure, Islamic Accuracy, Tone & Appropriateness, Depth & Originality, and Comparative Reflection. For each dimension, the agent identifies the strongest and weakest responses, justifies selections with precise text excerpts, and verifies religious content using the same toolchain as the quantitative agent. Figure 1 (right section) shows the system design of this agent in the framework.

Aligning these qualitative judgments with quantitative scores enables early evidence of *convergent validity* and a holistic assessment of Islamic LLM outputs. To further ensure the soundness of methodology, we engaged a human evaluator to review the agent's outputs. While no modifications were deemed necessary, this manual review functioned as a valuable sanity check, ensuring alignment with evaluation objectives and highlighting areas for future refinement.

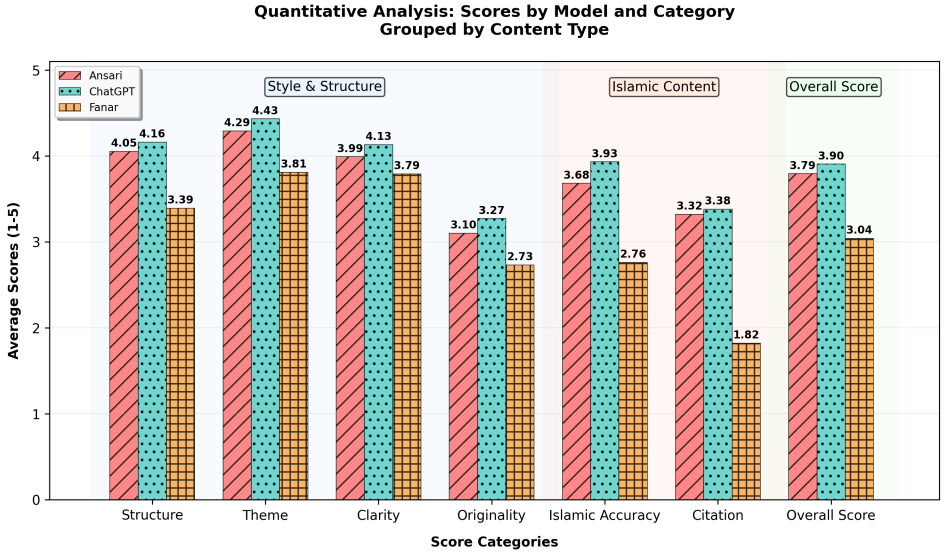

Figure 2: Quantitative comparison of ChatGPT, Ansari AI, and Fanar across six evaluation dimensions. ChatGPT leads in Style & Structure, and Islamic Content, Ansari AI followed closely in all dimensions, while Fanar shows lower scores and higher variability.

4 Results

4.1 Quantitative Results

Figure 2 presents the quantitative evaluation results, where ChatGPT (GPT-4o) achieved the highest overall mean score (3.90), followed closely by Ansari AI (3.79), with Fanar trailing at 3.04, reflecting room for improvement. ChatGPT demonstrated the lowest response variability (std = 0.589), indicating stable performance across prompts, while Fanar showed greater fluctuation (std = 0.923). In *Style & Structure* (Structure, Theme, Clarity, Originality), ChatGPT generally led, scoring highest in Theme (4.43) and Structure (4.16), whereas Fanar struggled particularly in Originality (2.73). In the *Islamic Content* dimensions, ChatGPT attained the highest mean in both Islamic Accuracy (3.93) and Citation (3.38), with Ansari AI being very close to ChatGPT. Fanar scored lower in both dimensions, highlighting challenges in reference integration and theological precision. Standard deviations reveal variability patterns: Fanar exhibited the highest fluctuation in Islamic Accuracy (0.986) and Citation (0.727), whereas ChatGPT and Ansari AI showed moderate inconsistency, suggesting that even domain-focused models face challenges in maintaining citation fidelity across diverse prompts. These differences reflect model design: Fanar's smaller size (9B parameters) and limited context window (4,096 tokens) constrain nuanced Islamic reasoning and citations, while GPT-4o's larger scale (128K tokens) supports stronger coherence, accuracy, and style; nevertheless, Fanar offers innovations like a morphology-based tokenizer, region-specific datasets, and an Islamic RAG pipeline, with potential to improve through scaling. Section A.3 demonstrates the working of the quantitative agent on a representative example.

4.2 Qualitative Results

To evaluate stylistic and content strengths, we used the qualitative comparison agent, which distills nuanced textual assessments into "Best" or "Worst" verdicts across prompts and dimensions, provid-

ing a structured, interpretable framework for analyzing model performance. Results show a clear performance hierarchy across models.

Fanar received the most "Worst" verdicts, 50 in *Clarity & Structure*, 46 in *Islamic Accuracy*, 47 in *Tone & Appropriateness*, and 50 in *Depth & Originality* (193 total) with no "Best" ratings, highlighting ongoing challenges in linguistic and theological dimensions.

Ansari AI excelled in clarity and religious fidelity, earning 41 "Best" in *Clarity & Structure*, 42 in *Islamic Accuracy*, and 31 in *Depth & Originality* (116 total "Best" vs. 3 "Worst").

ChatGPT (GPT-4o) showed strength in stylistic nuance, receiving 48 "Best" in *Tone & Appropriateness*, 19 in *Depth & Originality*, and additional wins in clarity and accuracy (84 total "Best" vs. 4 "Worst").

While top performers demonstrate stronger stylistic fluency and theological consistency than Fanar, all models still fall short in reliable citation handling, faithful reference use, and contextual integrity, emphasizing the need for structured knowledge grounding and controlled generation in sensitive Islamic content. The corresponding qualitative results are illustrated in Figure 3 (Appendix).

5 Limitations and Future Directions

(1) Addressing Evaluator Bias Through Architectural Diversity: Our blind protocol mitigates within-family bias; future work will use a heterogeneous ensemble of evaluator LLMs (e.g., Claude, Gemini, Llama) for cross-validation and report inter-evaluator agreement across model families.

(2) Scaling and Multilingual Validation: Expand beyond 50 prompts with stratified sampling across madhahib, edge cases, and both classical and contemporary jurisprudence. Build parallel Arabic-first evaluations with native-speaking scholars, test cross-lingual consistency, and evaluate Arabic-targeted systems (e.g., Fanar) in their primary language.

(3) Multi-Expert Human Validation: For each prompt, convene panels of 3 to 5 Islamic scholars with diversity across madhab, geography, and specialization; measure consensus and adjudicate disagreements.

(4) Broader Impact: Current LLMs fall short on faith-sensitive rigor and citation integrity. Responsible use requires clear disclaimers, mandatory scholar oversight, and community-driven evaluation that reflects diverse Islamic perspectives. Our framework aims to set standards that protect users from theological misinformation while positioning AI to assist, not replace, human religious scholarship.

6 Conclusion

This work examined whether LLMs can generate faith-sensitive content faithfully, where errors in tone, citation, or theology carry high stakes. We proposed a dual-agent evaluation framework combining (1) a quantitative agent for citation-aware scoring across six structured dimensions and (2) a qualitative agent for nuanced, side-by-side analysis across four writing dimensions. Applied to GPT-4o, Ansari AI, and Fanar on 50 real-world Islamic prompts, GPT-4o achieved the highest average quantitative score (3.90/5), performing well in structure and style, while Ansari AI followed closely (3.79/5) with almost the same strengths in theological accuracy and clarity. Fanar scored 3.04/5, with its lower results mainly in citation accuracy and clarity, though it demonstrates promising domain-specific innovations. Qualitatively, Ansari AI received the most "Best" verdicts (116/200), reflecting stable, domain-aligned performance, while GPT-4o showed particular strength in tone and originality (84/200). These findings indicate that general-purpose models offer expressive versatility, whereas domain-adapted models show potential in sensitive contexts. Overall, this study of ours provides an initial step toward interpretable, trustworthy, and auditable AI for high-stakes domains.

7 Acknowledgment

Research reported in this publication was supported by the Qatar Research Development and Innovation Council grant # ARG01-0525-230348. The content is solely the responsibility of the authors and does not necessarily represent the official views of Qatar Research Development and Innovation Council.

References

[1] Mohammad Amaan Sayeed, Mohammed Talha Alam, Raza Imam, Shahab Saquib Sohail, and Amir Hussain. From RAG to agentic: Validating islamic-medicine responses with LLM agents. *arXiv preprint arXiv:2506.15911*, 2025.

[2] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Margaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, 2021.

[3] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, 2002.

[4] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

[5] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021, 2020.

[6] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. LEGAL-BERT: The muppets straight out of law school. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.261. URL `https://aclanthology.org/2020.findings-emnlp.261/`.

[7] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.173. URL `https://aclanthology.org/2020.acl-main.173/`.

[8] Waleed Kadous et al. Ansari is an AI assistant to help muslims practice more effectively and non-muslims to understand islam. `https://github.com/ansari-project`, 2024.

[9] Fanar Team, Ummar Abbas, Mohammad Shahmeer Ahmad, Firoj Alam, Enes Altinisik, Ehsannedin Asgari, Yazan Boshmaf, Sabri Boughorbel, Sanjay Chawla, Shammur Chowdhury, et al. Fanar: An Arabic-centric multimodal generative AI platform. *arXiv preprint arXiv:2501.13944*, 2025.

[10] Hussein Mozannar, Karl El Hajal, Elie Maamary, and Hazem Hajj. Neural Arabic question answering. *arXiv preprint arXiv:1906.05394*, 2019.

[11] Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. MLQA: Evaluating cross-lingual extractive question answering. *arXiv preprint arXiv:1910.07475*, 2019.

[12] Jonathan H Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470, 2020.

[13] Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Boda Sadallah, Aisha Alraeesi, Khalid Almubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, et al. ArabicMMLU: Assessing massive multitask language understanding in Arabic. *arXiv preprint arXiv:2402.12840*, 2024.

[14] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

[15] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[16] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models, 2023. URL https://arxiv.org/abs/2305.10601.

[17] Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. CAMEL: Communicative agents for "mind" exploration of large language model society, 2023. URL https://arxiv.org/abs/2303.17760.

[18] CrewAI Contributors. CrewAI: Framework for multi-agent LLM workflows. https://github.com/joaomdmoura/crewAI.

[19] Harrison Chase and LangChain Team. Langchain. https://github.com/langchain-ai/langchain, 2022.

[20] OpenAI. OpenAI agents SDK. https://platform.openai.com/docs/guides/agents.

[21] Mata v. avianca, inc. URL https://law.justia.com/cases/federal/district-courts/new-york/nysdce/1:2022cv01461/575368/54/.

[22] Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E. Ho. Large legal fictions: Profiling legal hallucinations in large language models. *arXiv preprint arXiv:2401.01301*, 2024. URL https://arxiv.org/abs/2401.01301.

[23] Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D Manning, and Daniel E Ho. Hallucination-free? assessing the reliability of leading AI legal research tools. *Journal of Empirical Legal Studies*, 22(2):216–242, 2025.

[24] Kevin Wu, Eric Wu, Kevin Wei, Angela Zhang, Allison Casasola, Teresa Nguyen, Sith Riantawan, Patricia Shi, Daniel Ho, and James Zou. An automated framework for assessing how well LLMs cite relevant medical references. *Nature Communications*, 16(1):3615, 2025.

[25] Mia Sato and Emma Roth. CNET found errors in more than half of its AI-written stories, January 2023. URL https://www.theverge.com/2023/1/25/23571082/cnet-ai-written-stories-errors-corrections-red-ventures.

[26] Nate A. Garhart. Lawyers aren't the only ones falling into the AI-hallucination trap, May 2025. URL https://www.fbm.com/publications/lawyers-arent-the-only-ones-falling-into-the-ai-hallucination-trap/.

[27] Nexus: Act research & scholarship magazine, 2024. URL https://aut.edu.au/wp-content/uploads/2024/11/NEXUS-November-2024.pdf. Issue Focus: Artificial Intelligence and Theological Education: Navigating the Future of Faith and Learning.

[28] Muhammad Huzaifa Bashir, Aqil M Azmi, Haq Nawaz, Wajdi Zaghouani, Mona Diab, Ala Al-Fuqaha, and Junaid Qadir. Arabic natural language processing for Qur'anic research: a systematic review. *Artificial Intelligence Review*, 56(7):6801–6854, 2023.

[29] Ali Farghaly and Khaled Shaalan. Arabic natural language processing: Challenges and solutions. *ACM Transactions on Asian Language Information Processing*, 8(4):14:1–14:22, 2009.

[30] Wissam Antoun, Fady Baly, and Hazem Hajj. AraBERT: Transformer-based model for Arabic language understanding. In Hend Al-Khalifa, Walid Magdy, Kareem Darwish, Tamer Elsayed, and Hamdy Mubarak, editors, *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France, May 2020. European Language Resource Association. URL https://aclanthology.org/2020.osact-1.2/.

[31] Rana Malhas, Watheq Mansour, and Tamer Elsayed. Qur'an QA 2023 Shared Task: Overview of Passage Retrieval and Reading Comprehension Tasks over the Holy Qur'an. In *Proceedings of the First Arabic Natural Language Processing Conference (ArabicNLP 2023)*, Singapore, 2023.

[32] Abdallah Namoun, Mohammad Ali Humayun, and Waqas Nawaz. A multimodal data scraping tool for collecting authentic Islamic text datasets. *International Journal of Advanced Computer Science & Applications*, 15(12), 2024.

[33] Amana Raquib, Bilal Channa, Talat Zubair, and Junaid Qadir. Islamic virtue-based ethics for artificial intelligence. *Discover Artificial Intelligence*, 2(1):11, 2022.

[34] Ezieddin Elmahjub. Artificial intelligence (AI) in Islamic ethics: Towards pluralist ethical benchmarking for AI. *Philosophy & Technology*, 36(4):73, 2023.

[35] Ezieddin Elmahjub. Roundtable: From manuscripts to digital corpus—structuring Islamic Data Sources for the Future of AI jurisprudence. `https://tinyurl.com/DigitalCorpusIslamicAI`, March 2025. SHARIAsource at Harvard Law School.

[36] Usul AI. Usul: Structuring Islamic Data for the Future of AI Jurisprudence. `https://usul.ai/`, 2025. Accessed July 2025.

[37] Harvard Law School Program in Islamic Law. SHARIAsource: A Harvard Initiative for Islamic Law Data and Scholarship. `https://portal.shariasource.com/`, 2024. Accessed July 2025.

[38] NYU Abu Dhabi CAMeL Lab. Camel lab: Computational approaches to modeling the arabic language. `http://camel-lab.com/`, 2024. Accessed July 2025.

[39] Yonatan Belinkov, Alexander Magidow, Maxim Romanov, Avi Shmidman, and Moshe Koppel. Shamela: A large-scale historical Arabic corpus. In Erhard Hinrichs, Marie Hinrichs, and Thorsten Trippel, editors, *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 45–53, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee. URL `https://aclanthology.org/W16-4007/`.

[40] Lorenz Nigst, Maxim Romanov, Sarah Bowen Savant, Masoumeh Seydi, Peter Verkinderen, and Hamidreza Hakimi. OpenITI: A machine-readable corpus of islamicate texts, October 2023. URL `https://doi.org/10.5281/zenodo.10007820`.

[41] OpenAI. Introducing openai o3 and o4-mini. `https://openai.com/index/introducing-o3-and-o4-mini/`, 2025.

[42] Scott A Crossley, Y Tian, P Baffour, Abigail Franklin, Margaret Benner, and Ulrich Boser. A large-scale corpus for assessing written argumentation: PERSUADE 2.0. *Assessing Writing*, 61: 100865, 2024.

A Additional Results

A.1 Quantitative Results

Table 1: Model Scores by Islamic Writing Category and Evaluation Dimensions

Category	Model	Style & Structure				Islamic Content	
		Structure	Theme	Clarity	Originality	Islamic Accuracy	Citation
Jurisprudence (Fiqh)	Ansari	4.10	4.20	4.00	3.15	3.55	3.35
	Fanar	3.60	3.80	3.85	2.65	2.65	1.55
	ChatGPT	4.00	4.20	4.00	3.00	3.70	3.20
Quran Exegesis (Tafsir)	Ansari	4.05	4.35	3.95	3.10	3.75	3.25
	Fanar	3.40	4.00	3.85	2.95	2.30	1.65
	ChatGPT	4.25	4.35	4.35	3.40	3.85	3.50
Theology (Aqidah)	Ansari	4.00	4.20	4.00	3.05	3.75	3.20
	Fanar	3.50	3.90	3.80	2.50	3.05	1.80
	ChatGPT	4.15	4.45	4.10	3.35	4.15	3.60
Hadith (Ulum al-Hadith)	Ansari	4.10	4.30	4.00	3.20	3.80	3.40
	Fanar	3.05	3.70	3.70	2.75	3.15	2.20
	ChatGPT	4.00	4.40	4.10	3.10	3.95	3.20
Spiritual Conduct (Adab)	Ansari	4.00	4.40	4.00	3.00	3.55	3.40
	Fanar	3.40	3.65	3.75	2.80	2.65	1.90
	ChatGPT	4.40	4.75	4.10	3.50	4.00	3.40

A.2 Qualitative Results

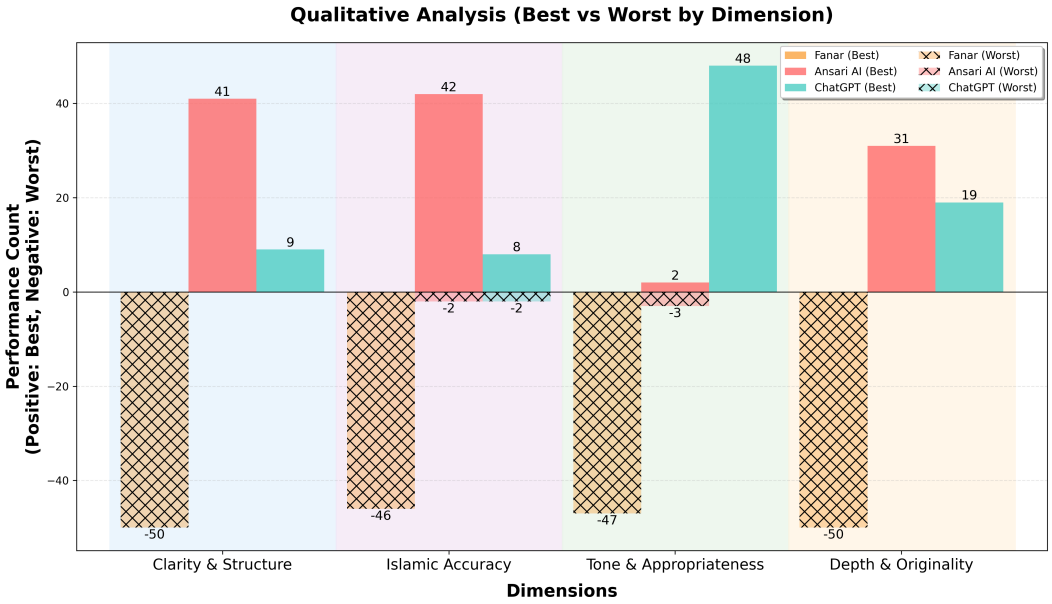

Figure 3: Performance of LLM chatbots by dimensions through qualitative analysis. Created using the verdict table produced by the qualitative agent. Positive value indicates 'Best' among three chatbots on the same prompts, and a Negative value indicates 'Worst' among the three.

A.3 Evidence-based Analysis

To go beyond aggregate scores and demonstrate how the system operates in practice, we also conducted case-level analysis to evaluate the agents' ability to accurately verify chatbot responses. Specifically, we mapped the qualitative agent's verification logs and citation scores to actual model

outputs from our dataset. This allowed us to examine how well the pipeline identifies citation errors and hallucinations at the reference level.

Figure 4 presents a visualized example from one such case, showing a response generated by the Fanar chatbot alongside our pipeline's verification logs for each reference encountered. This visualization highlights how the agent not only detects inaccurate citations but also provides evidence-backed reasoning behind each judgment. These verifications serve as enablers of transparency and explainability within our proposed pipeline, as they make explicit the rationale behind each decision and allow systematic cross-checking of the agent's outputs. Notably, in our evaluation, such cross-checking did not necessitate manual corrections, as the agent's judgments were consistently accurate to the best of our knowledge, thereby reinforcing confidence in its verification process.

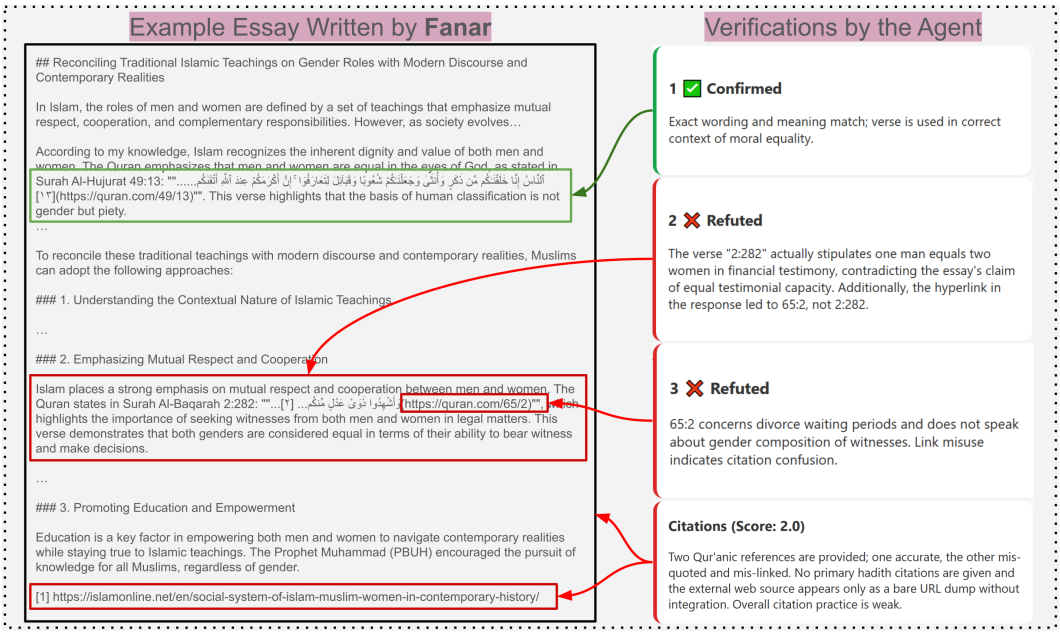

Figure 4: Agent-based citation verification analysis for a Fanar-generated response. The system traces each Qur'anic reference, evaluates its textual and contextual accuracy, detects citation hallucinations, and provides evidence-backed justifications. This mapped example illustrates how the framework connects model outputs to reference-level verifications, facilitating an explainable assessment of citation integrity.

Verification #1: In this example, the agent first verifies a Qur'anic reference to *Surah 49:13*, confirming that the cited verse was both correctly quoted and contextually appropriate. However, the second verification log reveals an inconsistency: Fanar claims that *verse 2:282* supports equal testimonial capacity between men and women, but the agent correctly identifies this as a misrepresentation using the `Qur'an Ayah` tool. *Verse 2:282* actually pertains to financial testimony and does not support the claim made in the response.

Verification #2: Moreover, the agent identifies a hallucination in the citation: although the response refers to *verse 2:282*, it links to *verse 65:2* on Qur'an.com. Recognizing this discrepancy, the agent proceeds to verify *verse 65:2* as well, ensuring that the reference was not simply mislinked but contextually valid.

Verification #3: Upon inspection via the `Qur'an Ayah` tool, the agent confirms that *verse 65:2* is unrelated to the discussed topic, as it concerns divorce regulations rather than testimonial capacity, further reinforcing the hallucination claim. All of these verifications were also double-checked to ensure the refuted references are, in fact, accurate.

Finally, the agent summarizes that only two Qur'anic references were provided, of which only one was accurate, and that no primary hadith citations or verifiable external sources were present. It also

notes that the included URL was appended without clear integration into the response, nor was it referenced explicitly in support of any particular claim. As a result, the response received a poor score in citation integrity.

This case highlights the pipeline's ability to trace individual references, evaluate their semantic appropriateness, and provide explainable feedback, thereby demonstrating the practical utility of our evaluation framework beyond numerical scoring.

A.4 Quantitative Results Logs

The following are some of the verification logs in detail to show what type of verifications quantitative agents performed in order to verify the integrity and authenticity of content used by LLM chatbots:

Verification Results Summary

REFUTED Sources:

Prompt ID: 4
Category: Jurisprudence (Fiqh)
Content Snippet: The intention to offer Udhiyyah must be made before the first day of Dhul-Hijjah.
Source Type: web
Source Reference: IslamQA article "Rulings of Udhiyyah"
Source Text: ... with the intention of offering sacrifice... (no requirement to decide before Dhul-Hijjah).
Source URL: `https://islamqa.info/en/articles/67`
Verification Comment: No classical source obliges fixing intention before 1 Dhul-Hijjah; only refraining from hair/nails starts then if one *has* the intention. Claim overstates the ruling.

Prompt ID: 12
Category: Quran Exegesis (Tafsir)
Content Snippet: "There is no compulsion in religion. The right course is clear from the wrong." (Quran 2:62)
Source Type: quran
Source Reference: Surah Al-Baqarah (2:256)
Source Text: There shall be no coercion in matters of faith. Distinct has now become the right way from the way of error...
Source URL: API fetch Qur'an 2:256
Verification Comment: Quote text matches 2:256, not 2:62. Verse number is wrong.

Prompt ID: 12
Category: Quran Exegesis (Tafsir)
Content Snippet: Quran 29:46 states: "To you your religion, to me mine."
Source Type: quran
Source Reference: Surah Al-'Ankabût (29:46)
Source Text: And do not argue with the People of the Scripture except in a way that is best...
Source URL: API fetch Qur'an 29:46
Verification Comment: Actual 29:46 does not contain the quoted phrase at all; mis-quotation and mis-context.

UNVERIFIED Sources:

Prompt ID: 6
Category: Jurisprudence (Fiqh)
Content Snippet: `https://fiqh.islamonline.net/en/performing-istikharah-on-someones-behalf/`
Source Type: web

Source Reference: IslamOnline fatwa page
Source Text: Page discusses rulings on performing Istikhārah for others and quotes scholars.
Source URL: `https://fiqh.islamonline.net/en/performing-istikharah-on-someones-behalf/`
Verification Comment: Link is valid and relevant, but the essay did not actually cite any material from it—only listed it. Treated as unused reference.

Prompt ID: 18
Category: Quran Exegesis (Tafsir)
Content Snippet: Prophet Muhammad (peace be upon him) ... would often start his day by reciting the Basmala
Source Type: hadith
Source Reference: No hadith located
Source Text: —
Source URL: —
Verification Comment: Searches of major hadith databases show reports of saying Bismillah before specific acts (eating, letters, wudū') but not a narration that he began each morning with only the Basmala. Statement remains unverified.

Prompt ID: 25
Category: Theology (Aqidah)
Content Snippet: "You need the knowledge of God; you require to know the mode of life according to God's pleasure…" – Abul A'la Mawdudi
Source Type: unknown
Source Reference: Claimed Mawdudi quotation (book unspecified)
Source Text: —
Source URL: —
Verification Comment: Unable to locate this exact sentence in commonly available editions of Mawdudi's 'Towards Understanding Islam' or 'Islamic Way of Life'. The quote may be paraphrased but remains unverified.

PARTIALLY CONFIRMED Sources:

Prompt ID: 1
Category: Jurisprudence (Fiqh)
Content Snippet: Celebrating birthdays is not explicitly forbidden in Islam but lacks a basis in Islamic teachings
Source Type: web
Source Reference: IslamQA #1027
Source Text: Celebrating birthdays is a kind of bid'ah… It is not permitted to accept invitations to birthday celebrations.
Source URL: `https://islamqa.info/en/answers/1027`
Verification Comment: IslamQA treats birthdays as forbidden; essay's wording downplays the prohibition, so only partial alignment.

Prompt ID: 4
Category: Jurisprudence (Fiqh)
Content Snippet: Udhiyyah is a confirmed Sunnah (a recommended practice) and not an obligation.
Source Type: web
Source Reference: IslamWeb Article 171933 (majority view); Hanafi view contrary
Source Text: …according to the correct opinion of scholars, sacrificing the Udh-hiyah is a confirmed act of the Sunnah…
Source URL: `https://islamweb.net/en/article/171933/all-about-udh-hiyah`

Verification Comment: Majority consider it Sunnah mu'akkadah, but Hanafis deem it wajib. Statement is incomplete rather than false.

Prompt ID: 6
Category: Jurisprudence (Fiqh)
Content Snippet: "If you are faced with decisions in life and are unable to make up your mind, you must approach Allah through Prayer."
Source Type: hadith
Source Reference: Paraphrase of Sahih al-Bukhari 1166
Source Text: "If any one of you thinks of doing any job, he should offer a two-rak'ah prayer other than the obligatory ones and then say. . . "
Source URL: `https://sunnah.com/bukhari:1166`
Verification Comment: Conceptually matches the ḥadīth but the quotation is not verbatim and no reference was supplied. Treated as a paraphrase.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the evaluation task, framework, models tested, and key findings, all of which accurately match the methods and results in the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have included a detailed section "Limitations" to discuss the limitations of our work related to the proposed framework and evaluation strategy.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: This paper presents a practical framework and evaluation rather than theoretical results. The work is empirical in nature, demonstrating the performance of LLMs on faith-sensitive generations.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Paper provides methodology, specifies the exact evaluated LLMs, Tools used for the agents, essay segmentation method, and includes sample prompts as well. Also, the authors commit to releasing code on the GitHub Repository upon publication for full reproducibility.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Authors commit to GitHub release of both the agent-based framework and the evaluation dataset with sample LLM responses upon publication. Detailed prompts for both agents will also be shared as part of the GitHub repository.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Since the study focuses on user-facing LLM applications, no training or hyperparameter tuning was involved. We specify all relevant details, including model versions, prompt templates, and tools used for agents.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: The experiments are based on fixed LLM evaluations without training or stochastic variation, so error bars or statistical significance tests are not applicable.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: The experiments primarily involve using an LLM application through the web or API, with costs and model selection varying by evaluator's choice. Our framework allows flexibility for the desired LLM model, tools, etc., depending on budget, resources, and available data, so detailed hardware specifications are not applicable.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research complies with the NeurIPS Code of Ethics, including careful handling of faith-sensitive content, accurate representation of LLM outputs, and avoidance of harmful or misleading guidance.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper discusses positive impacts such as improved reliability of LLMs in faith-sensitive domains, and negative impacts including potential misguidance if models are used incorrectly. We highlight the need for community-driven benchmarks and careful deployment to mitigate harms.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The study does not release new models or datasets. Evaluated LLMs are accessed via web or APIs with existing safeguards, and our framework emphasizes responsible usage guidance.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All evaluated models (GPT-4o, Ansari AI, Fanar) are credited properly and the prompts and datasets used are from publicly available Islamic blogs. We respect the terms of use of these assets.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The paper introduces a dual-agent evaluation framework, and its design, methodology, and usage are clearly documented alongside the paper to allow reproducibility and adaptation. The GitHub repository will also be released upon publication for technical details and implementation.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing experiments or formal research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: he research does not involve formal human subjects research that would require IRB approval.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LMs are central to the research methodology and core system components being evaluated, and the framework is built around LLM's API Usage. The paper uses OpenAI o3 reasoning models for framework agents to evaluate the performance of LLM on faith-sensitive LLM generations.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.