





# Generative AI for Islamic Texts: The EMAN Framework for Mitigating GPT Hallucinations

Amina El Ganadi<sup>1,3</sup><sup>a</sup>, Sania Aftar<sup>2</sup><sup>b</sup>, Luca Gagliardelli<sup>2</sup><sup>c</sup> and Federico Ruozzi<sup>1</sup><sup>d</sup>

<sup>1</sup>Department of Education and Humanities, University of Modena and Reggio Emilia, Reggio Emilia, Italy

<sup>2</sup>Department of Engineering, University of Modena and Reggio Emilia, Modena, Italy

<sup>3</sup>Department of Culture and Society, University of Palermo, Palermo, Italy  
{amina.elganadi, sania.aftar, luca.gagliardelli, federico.ruozzi}@unimore.it

**Keywords:** Generative AI Applications, Digital Humanities, Hallucinations, Religious Text Analysis, Bias Mitigation, Context-Aware Constraints, Prompt Engineering, Large Language Models (LLMs), GPT Builder, AI in Islamic Studies, Hadith Studies, Sahih Al-Bukhari.

**Abstract:** Recent advancements in large language models (LLMs) have facilitated specialized applications in fields such as religious studies. Customized AI models, developed using tools like GPT Builder to source information from authoritative collections such as *Sahih al-Bukhari* or the Qur'an, were explored as potential solutions to address inquiries related to Islamic teachings. However, initial evaluations highlighted significant limitations, including hallucinations and reference inaccuracies, which undermined their reliability for handling sensitive religious content. To address these limitations, this study proposes EMAN (Embedding Methodology for Authentic Narrations), a novel framework designed to enhance adherence to *Sahih al-Bukhari* through API-based integration. Three methodologies are examined within this framework: Zero-Shot Instructions, which guide the model without prior examples; Few-Shot Learning, which fine-tunes the model using a limited set of examples; and Embedding-Based Integration, which grounds the model directly in a verified Ahadith database. Results demonstrate that Embedding-Based Integration significantly improves performance by anchoring outputs in a structured knowledge base, reducing hallucination rates, and increasing accuracy. The success of this approach underscores its potential for enhancing LLM performance in precision-critical domains. This research provides a foundation for the ethical and accurate deployment of AI in religious studies, emphasizing accountability and fidelity to source material.

## 1 INTRODUCTION


Hadith literature, which includes the recorded sayings, actions, and approvals of the Prophet Muhammad, forms a foundational pillar of Islamic theology, law, and ethics. It is considered the second most important source of Islamic legislation after the Qur'an, providing essential context and elaboration for Quranic teachings.


The Ahadith (plural of Hadith) offer practical guidance on every aspect of life, from personal conduct to societal governance, making them indispensable for a comprehensive understanding and practice of Islam. The classification of Ahadith into


various categories of authenticity, *Sahih* (authentic), *Da'if* (weak), *Hasan* (good), and *Mawdu'* (fabricated) (Ilyas, 2018; Mghari et al., 2022) reflects the scholarly rigor and historical accuracy invested in the study of these texts. This system is crucial for ensuring the integrity of Hadith-based jurisprudence.


Among Hadith collections, *Al-Kutub al-Sittah* (the Six Canonical Books) are the foremost in Sunni Islam (Kamali, 2014), with *Sahih al-Bukhari* recognized as the most authentic and often regarded as the most reliable book after the Qur'an.

Despite the profound importance of Hadith literature, scholars and researchers face significant challenges due to its vast scope and lack of structured organization. The extensive collections contain thousands of Ahadith, spread across multiple volumes and often not systematically categorized. This complexity makes navigating, analyzing, and interpreting the texts a demanding task. The intricate networks of the

<sup>a</sup> <https://orcid.org/0000-0002-8196-2628>

<sup>b</sup> <https://orcid.org/0000-0001-8151-8941>

<sup>c</sup> <https://orcid.org/0000-0001-5977-1078>

<sup>d</sup> <https://orcid.org/0000-0003-2729-5016>

*Sanad* (chain of narrators who transmitted the Hadith) require meticulous examination to assess the reliability of each Hadith, while the *Matn* (the core text or the actual content of the Hadith) demands deep linguistic and contextual analysis to fully understand its implications (Aftar et al., 2024a,b).

The rise of Generative Artificial Intelligence (AI), particularly Large Language Models (LLMs) such as ChatGPT, (Brown et al., 2020), Falcon LLM (Almazrouei et al., 2023), Fanar<sup>1</sup>, and Claude (Templeton et al., 2024), presents transformative opportunities for Islamic studies. These models have demonstrated remarkable capabilities in processing and generating human-like text, offering significant potential in organizing, analyzing and improving access to Hadith literature.

## 2 PROBLEM STATEMENT

Despite potential benefits, the integration of Generative AI into interpreting Hadith literature faces critical challenges that must be addressed. The complexity of the texts and the risk of hallucinations require robust methodologies to ensure accuracy. Moreover, the quality and reliability of the data used to train AI models are concerning, as the internet is replete with incorrect Islamic legal rulings (*fatwas*) and interpretations. Such discrepancies can lead to serious misrepresentations, bias, and judicial inconsistencies in applying *Shari'a* (Islamic Law), particularly given the diverse interpretations across different Islamic schools of thought. Precise interpretation is crucial, as these decisions carry significant social, ethical, and legal implications within the Muslim community. Additionally, the inherent opacity of large language models complicates error correction, fostering skepticism among scholars and the wider Muslim community. Thus, maintaining a high level of precision and respecting diverse theological perspectives are essential to ensure that AI-supported applications do not inadvertently propagate errors or biases, potentially resulting in judicial inconsistencies.

## 3 RELATED WORK

Recent studies have investigated GPT models' applications and limits, highlighting their adaptability and challenges like hallucinations, while proposing strategies to improve their reliability in sensitive contexts.

<sup>1</sup><https://fanar.qa/en>

### 3.1 GPT in Practice

Advancements in large language models like GPT-3.5 and GPT-4 have been notable through few-shot learning, where these models occasionally surpass fine-tuned counterparts in specialized tasks (Brown et al., 2020). Few-shot learning enables GPT-4 to generalize effectively with minimal contextual input, providing a viable alternative to domain-specific models. Nevertheless, challenges persist in highly specialized domains, particularly in religious studies (Rizqullah et al., 2023).

Such domain-specific challenges surfaced again when evaluating the potential of ChatGPT for organizing an Islamic digital library. It demonstrated the ability to hierarchically categorize topics, while also revealing issues like interpretability, generalization, and hallucination (El Ganadi et al., 2023). These issues reflect broader difficulties in ensuring AI reliability in culturally and semantically nuanced domains (Rizqullah et al., 2023; Alnefaie et al., 2023), such as occasional miscategorization by ChatGPT, highlighting struggles in understanding nuanced relationships within religious texts.

The introduction of the QASiNa dataset for *Sirah Nabawiyah* in Indonesian provided new insights into language model performance (Rizqullah et al., 2023). Tests on mBERT and IndoBERT showed that XLM-R achieved the highest accuracy, while ChatGPT-3.5 and GPT-4, although less accurate, scored higher on Substring Match. This suggests that while ChatGPT excels in phrase matching, it struggles with the deeper semantic understanding required for religious question-answering, echoing previous findings in religious and linguistic studies (Alnefaie et al., 2023).

### 3.2 Mitigating Hallucinations in AI Models

Recent research has explored hallucinations in Large Language Models to understand their origins, implications, and mitigation strategies, emphasizing the importance of accuracy in sensitive domains (Reddy et al., 2024). LLMs often generate convincing but incorrect responses (Ji et al., 2023a), which can mislead users due to their natural language fluency.

Hallucinations have been defined as outputs that deviate from the original prompt, resulting in irrelevant or erroneous content (Gallifant et al., 2024). To address this issue, the Mistral 7B model was applied, leveraging few-shot learning and prompt engineering to improve the detection of hallucinations. This approach outperformed baseline models in both English and Swedish, demonstrating the effectiveness of tai-

lored techniques for enhancing the accuracy and reliability of large language models (Siino and Tinnirello, 2024).

An evaluation of ChatGPT and Gemini (Bard) for generating references in systematic reviews on shoulder rotator cuff pathology revealed frequent inaccuracies in citations (Chelli et al., 2024), raising concerns about their reliability in scientific applications where accuracy is essential.

To mitigate hallucinations, various strategies have been proposed, including cross-language summary accuracy (Qiu et al., 2023), contextual detection frameworks (Varshney et al., 2023), and analyzing causes such as self-contradiction (Mündler et al., 2023). Retrieval-Augmented Generation (Lewis et al., 2020) and frameworks like EVER (Wang et al., 2023) integrate external knowledge to enhance factuality, while techniques like self-reflection (Madaan et al., 2023) and RARR (Gao et al., 2022) refine outputs through iterative feedback or post-generation alignment. Knowledge graph-based methods (Ji et al., 2023b) further address hallucinations by grounding outputs in structured data. These approaches collectively contribute to advancing reliable solutions for reducing inaccuracies in LLM-generated information, particularly in high-stakes domains like medicine, legal research, and scientific citation.

## 4 BACKGROUND AND INITIAL MODEL APPROACH

The development of specialized GPT models<sup>2</sup>, including *Bukhari GPT*<sup>3</sup>, was inspired by the potential of AI to advance religious studies, particularly in addressing inquiries related to Ahadith and the Qur'an<sup>4</sup>. Our initial approach involved developing a customized model, *Bukhari GPT*, utilizing the GPT builder platform. This model utilized the *Sahih Al-Bukhari* collection, a revered corpus within Sunni Islam, as its authoritative knowledge base. Given the sensitive nature of religious content, especially Ahadith which are crucial for Islamic jurisprudence, we prioritized accuracy and strict adherence to the source texts to prevent misinterpretation.

*Bukhari GPT* was developed with stringent constraints: it was configured to generate responses solely from the *Sahih Al-Bukhari* collection, prohibited the use of external references, and employed a default non-availability response for queries lacking di-

rect textual support. Despite these rigorous measures, real-world testing revealed significant challenges, including hallucinated Hadith references, misattributed Hadith numbers, and difficulties in producing coherent responses to complex queries.

Parallel testing involved *Bukhari GPT* and general-purpose models such as ChatGPT-4, Claude, and Gemini in a controlled environment. All models were granted access to the *Sahih Al-Bukhari* texts: *Bukhari GPT* through CSV files and the general purpose models via online resources<sup>5</sup>. All models were bound by identical constraints, use only the designated dataset, provide no speculative content, and respond with a default non-availability message when applicable. Additionally, the general-purpose models utilized their internal base knowledge to supplement their responses, providing a broader context when addressing queries. This reliance on internal knowledge bases was inevitable since, unlike *Bukhari GPT*, these models are retrained continuously and are not strictly confined to a specific dataset. As such, we could not fully control their use of broader information beyond the *Sahih Al-Bukhari* texts. The evaluation of each model focused on their adherence to the constraints, their accuracy in referencing, and their capability to handle ambiguities without fabricating responses.

Although *Bukhari GPT* demonstrated better performance compared to other models, it still exhibited suboptimal precision (0.42) and recall (0.52). These metrics underscore the need for more reliable methods to accurately identify authentic Hadith while minimizing false positives. Concurrently, general-purpose models such as ChatGPT-4, Claude, and Gemini encountered parallel challenges, including misquotations, inaccuracies, and hallucinations. These issues highlight the difficulty of adapting general AI systems to the nuanced and high-stakes contexts of religious studies. Table 2 demonstrates the discrepancies in precision, recall, and F1 scores among these models, underscoring the need for improved technical infrastructure and better data handling when working with specialized content like religious texts.

Additionally, technical constraints in the GPT Builder architecture sometimes hindered *Bukhari GPT* from fully accessing the CSV-based Hadith data, causing it to revert to its broader pre-trained knowledge base. This fallback mechanism led to unintended violations of the “no external sources” rule and weakened the model’s ability to cite Hadith correctly (see Fig.1).

Inconsistencies were also observed in the use of default responses, with the model occasionally at-

<sup>2</sup>GPT sahih-al-bukhari

<sup>3</sup>Bukhari-gpt

<sup>4</sup>GPT Quran

<sup>5</sup>Sunnah.com/Sahih-al-Bukhari

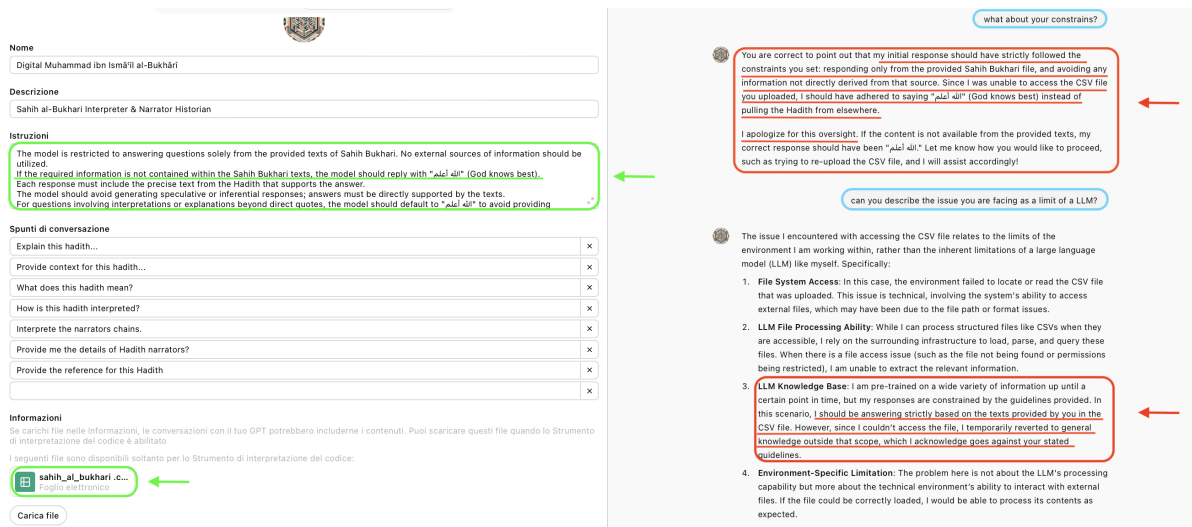


Figure 1: GPT Builder Limitations Experienced in *Bukhari GPT* Configuration.

tempting to answer questions despite lacking the necessary information, thereby undermining the reliability of the response protocol.

These challenges exposed the limitations of the experimental approach, highlighting the complex requirements of handling religious texts with generative AI models. The issues of hallucinations, misreferencing, and inconsistent fallback responses underscored the need for more structured methodologies. Notably, they illustrated the limitations of relying solely on prompt engineering and source restriction to manage sensitive content effectively. This evaluation informed the current study, which explores structured methods, such as embedding-based integration, to mitigate these issues and improve the model’s capacity to handle high-stakes religious content reliably.

The subsequent sections detail these advanced approaches, focusing on strategies to mitigate hallucinations, improve citation accuracy, and strengthen the overall integrity of AI models in religious studies applications.

## 5 METHODOLOGY

To address the identified limitations, this study introduces and evaluates the EMAN (Embedding Methodology for Authentic Narrations) framework. This novel approach is designed to enhance adherence to *Sahih al-Bukhari* as the authoritative source by minimizing hallucinations and improving response accuracy. The acronym EMAN (also transliterated as *Iman*), meaning “faith” or “belief” in Arabic, reflects the framework’s commitment to authentic narrations, particularly those from *Sahih al-Bukhari*.

Building on prior experiments conducted with the GPT Builder architecture and addressing its limitations, EMAN employs a systematic approach to evaluate and refine the model’s ability to generate accurate, contextually relevant, and constrained outputs. To achieve these objectives, the framework incorporates three structured strategies: **Zero-Shot Learning**, **Few-Shot Learning**, and **Embedding-Based Learning**. These strategies are designed to work in a complementary manner. Zero-Shot Learning leverages pre-trained knowledge to establish a baseline performance without requiring additional labeled data. Few-Shot Learning builds on this by introducing curated examples to refine contextual understanding and guide the model toward domain-specific outputs. Embedding-Based Learning further enhances the framework by anchoring responses in a verified corpus.

### 5.1 Dataset Preparation

This study relies on a dataset extracted from the *Sahih-al-Bukhari* collection<sup>6</sup>, comprising the original Arabic texts and their English translations. To ensure compatibility with approaches like zero-shot, few-shot, and embedding-based API integration, the data was structured in CSV files, enabling seamless integration and embedding creation. The dataset was curated and cleaned to ensure consistency and accuracy for use in all subsequent experiments.

<sup>6</sup>Sunnah.com/Sahih-al-Bukhari

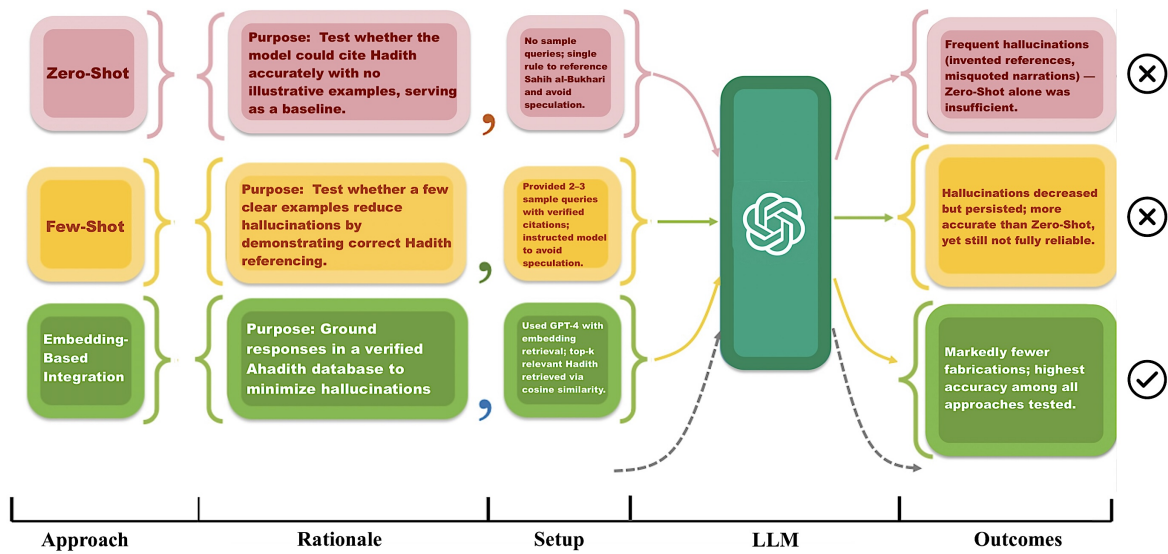


Figure 2: Overview of three tested approaches (Zero-Shot, Few-Shot, and Embedding-Based Integration).

## 5.2 Zero-Shot Learning

The zero-shot learning approach was employed to evaluate the ability of a pre-trained LLM to generate accurate and contextually relevant responses based solely on carefully designed instructions, without requiring additional labeled examples or fine-tuning (Liu et al., 2023). This approach was chosen for its scalability and flexibility in addressing diverse queries while ensuring strict adherence to *Sahih al-Bukhari* as the authoritative source. Zero-shot learning leverages the model’s pre-trained knowledge, enabling rapid deployment in contexts where extensive labeled datasets or domain-specific fine-tuning are unavailable, making it an appropriate choice for the initial stages of this study. The implementation utilized a structured prompt with explicit constraints to maintain fidelity to the source material. The model was required to directly quote from *Sahih al-Bukhari*, accurately cite Ahadith, and issue a default response indicating uncertainty for unsupported or out-of-scope queries. To mitigate hallucinations and speculative content, the prompt explicitly prohibited inferential reasoning, modern analogies, or commentary not explicitly grounded in the source. All responses included precise citations, specifying both book and Hadith numbers, with references verified using the “sunnah.com” platform. The OpenAI GPT-4 API was configured with a low-temperature setting of 0.2 to prioritize factual accuracy and deterministic outputs. While this framework demonstrated the model’s ability to generalize effectively within defined constraints, limitations such as occasional misquotations and inconsistent fallback mechanisms

highlighted the need for further refinement in prompt engineering and stricter enforcement of constraints to improve reliability.

## 5.3 Few-Shot Learning

This methodology was employed to address the limitations of the zero-shot approach by incorporating structured examples that guided the model’s responses while maintaining adherence to *Sahih al-Bukhari*. This method aimed to enhance contextual accuracy and response consistency in the analysis of Hadith. A carefully structured prompt template established the foundational guidelines for this approach. The model was required to directly quote from *Sahih al-Bukhari*, provide accurate Hadith citations, and issue a default response indicating uncertainty whenever the necessary information was unavailable, thereby ensuring adherence to constraints and preventing unsupported outputs. For queries involving multiple applicable Ahadith, the response format ensured that each reference was listed separately to preserve clarity and traceability. Examples were carefully curated to represent a diverse range of query types, including inquiries about charity, patience, and kindness to parents. These examples demonstrated the ideal response format, emphasizing concise, source-based answers with accurate citations. The implementation leveraged OpenAI’s GPT-4 API, configured with a low-temperature setting (0.2) to ensure deterministic outputs while minimizing response variability. By introducing task-specific examples, the few-shot learning approach significantly improved the model’s alignment with task requirements com-

pared to the zero-shot method. However, challenges such as occasional hallucinations, speculative content, and misquotations remained, indicating areas for further refinement in prompt engineering and example selection. Despite these limitations, the methodology effectively enhanced the model's ability to generate accurate and contextually appropriate responses.

#### 5.4 The EMAN Framework

The embedding-based approach was introduced to overcome the limitations of zero-shot and few-shot methodologies by anchoring responses in a semantically structured knowledge base derived from *Sahih al-Bukhari*. High-dimensional embeddings for each Hadith were generated using OpenAI's *text-embedding-ada-002 model*, which captured the semantic meaning of the texts while preserving critical metadata such as chain references for contextual accuracy. This embedding-based framework enabled precise retrieval of relevant Ahadith, minimizing the reliance on the model's internal generalized knowledge. A cosine similarity mechanism was employed to measure the relevance between user queries and precomputed Hadith embeddings. To ensure relevance, a similarity threshold filtered out irrelevant matches, and the top-k most relevant Ahadith were selected for inclusion in the response. If no Hadith meets the similarity threshold, the system provides a default response indicating uncertainty, which adheres to constraints and prevents unsupported outputs. Responses were structured to include direct quotes, Hadith numbers, and chain references, maintaining clarity and traceability. This approach demonstrated significant improvements in accuracy and contextual alignment compared to previous methods, effectively reducing hallucinations and misquotations by grounding responses in verified embeddings. However, occasional false negatives were observed due to the strict similarity threshold, which excluded some semantically relevant Ahadith with minor variations. Future refinements will focus on optimizing similarity thresholds, enhancing the embedding process, and incorporating advanced linguistic features to further improve retrieval accuracy ensuring precision and significantly reducing hallucinations. Together, these strategies create a cohesive and systematic approach that balances scalability, contextual accuracy, and source fidelity. The EMAN framework also integrates a detailed modeling pipeline tailored for Hadith analysis. This pipeline provides a structured and reproducible method for mitigating hallucinations, emphasizing precision, reliability, and strict adherence to source material. By combining these strategies,

EMAN offers a robust solution for deploying generative AI in sensitive scholarly domains, particularly the study of Islamic texts. Figure 2 details the methodologies and outcomes of different approaches.

#### 5.5 Evaluation Metrics

To evaluate the effectiveness of the EMAN framework, we tested various models including GPT-4, Claude, Gemini, and the custom-developed *Bukhari GPT*. We assessed their performance using precision, recall, F1-score, and accuracy metrics on two sets of prompts: one with 50 authentic Ahadith queries and another featuring fabricated Ahadith prompts. This evaluation, detailed in Table 2, illustrates the EMAN framework's superior performance in accurately referencing and interpreting religious texts.

### 6 RESULTS AND DISCUSSION

The evaluation of Large Language Models aimed to compare the performance of the proposed EMAN framework with widely used general-purpose models in handling religious texts, specifically Ahadith. This assessment focused on the models' ability to generate precise, contextually accurate responses while minimizing errors and hallucinations, a critical requirement given the sensitivity of the domain. Fifty queries were designed to evaluate the models across key metrics, including precision, recall, F1-score, accuracy, and hallucination rate. These metrics provided a comprehensive assessment of each model's capacity to retrieve authentic Hadith and address nuanced religious queries.

The EMAN framework emerged as the best-performing system, achieving a precision of 0.65, recall of 0.70, and an F1-score of 0.67. These results underscored the advantages of domain-specific optimization. In contrast, the customized GPT model, developed using tools like GPT Builder, demonstrated moderate improvements with a precision of 0.42, recall of 0.52, and an F1-score of 0.47, revealing persistent challenges in achieving higher precision and contextual reliability. General-purpose models like GPT-4 showed moderate performance (precision 0.38, recall 0.49, F1-score 0.43), struggling to adapt effectively to specialized contexts. Claude and Gemini performed the weakest, with Gemini particularly underperforming due to inadequate domain-specific training. These findings highlighted the superior reliability of the EMAN framework while revealing the limitations of both general-purpose and customized GPT models in handling high-stakes applications.

Table 1: Evaluation of Model Accuracy and Hallucination Frequency.

Model	Constraints followed	Misquotation	Accuracy%	Remarks
<b>Embedding-Based Integration</b>	Fully	No	100%	Example-based guidance improves style adherence but struggles with hallucinations.
Few-Shot Learning	Partial	Yes	72%	Understands response style but still faces issues with hallucinations and misquotation.
Zero-Shot Learning	Partial	Yes	72%	Lacks examples, leading to guideline inconsistencies and high hallucination rates.

Table 2: Model Evaluation Metrics.

Model	Precision	Recall	F1 Score
<b>EMAN</b>	<b>0.65</b>	<b>0.70</b>	<b>0.67</b>
<i>Bukhari GPT</i>	0.42	0.52	0.47
Chat GPT-4	0.38	0.49	0.43
Claude	0.35	0.33	0.34
Gemini	0.20	0.23	0.26

Further experiments evaluated EMAN’s ability to reliably detect fabricated Ahadith, comparing three approaches: Embedding-Based Integration, Few-Shot Learning, and Zero-Shot Learning (see Table 1). The Embedding-Based Integration approach emerged as the most effective, achieving 100% accuracy with no misquotations and full adherence to constraints. This performance was attributed to its grounding in a verified database and strict enforcement of response constraints, making it highly reliable for identifying fabricated Hadith. In comparison, Few-Shot Learning and Zero-Shot Learning both achieved 72% accuracy but exhibited partial adherence to constraints and frequent misquotations. Few-Shot Learning benefitted from example-based guidance, improving adherence to response styles, but it struggled to eliminate hallucinations entirely. Zero-Shot Learning, lacking contextual examples, faced greater inconsistencies and higher hallucination rates.

Overall, these findings emphasize the effectiveness of the Embedding-Based Integration approach within the EMAN framework in achieving precise, reliable outputs, particularly in sensitive domains like religious studies. While Few-Shot and Zero-Shot Learning showed moderate success, their limitations further underscore the need for structured, data-grounded methodologies to ensure reliability and ethical compliance in high-stakes applications.

## 7 CONCLUSIONS

This study highlights the critical role of grounding AI systems in structured, authenticated datasets when

processing sensitive religious content. By consistently relying on verified sources, the EMAN framework substantially decreases errors and boosts contextual accuracy, enabling more trustworthy access to complex Islamic texts such as Ahadith. Building on these findings, future research will extend the EMAN framework to the Arabic text of *Sahih al-Bukhari*, incorporating *Sanad* (chains of narrators) to assess whether enriched metadata can further boost model performance. This exploration directly follows from our observation that domain-specific data, rigorously verified, can significantly enhance contextual accuracy. We also plan to evaluate Arabic-developed models like Fanar, focusing on their ability to process Islamic texts with high fidelity. By comparing Fanar to more generalized models, we aim to determine whether its cultural and linguistic grounding provides a distinct advantage in handling religious content.

A key objective moving forward is to enhance the scalability of our embedding-based methods for larger, multilingual datasets, an improvement critical to broadening the EMAN framework’s applicability. By examining strategies such as more robust indexing or distributed architectures, we aim to ensure that grounded AI systems can serve diverse contexts and languages with the reliability and precision demanded by sensitive content.

## ACKNOWLEDGEMENTS

This work was conducted within the PNRR project ITSEERR - Italian Strengthening of the ESFRI RI RESILIENCE” (Avviso MUR 3264/2022) funded by EU – NextGenerationEU - Grant No IR0000014.

## REFERENCES

- Aftar, S., Gagliardelli, L., El Ganadi, A., Ruoizzi, F., and Bergamaschi, S. (2024a). Robert2vectm: A novel approach for topic extraction in islamic studies. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9148–9158.

- Aftar, S., Gagliardelli, L., El Ganadi, A., Ruoizzi, F., and Bergamaschi, S. (2024b). A novel methodology for topic identification in hadith. In *Proceedings of the 20th Conference on Information and Research Science Connecting to Digital and Library Science (formerly the Italian Research Conference on Digital Libraries)*.
- Almazrouei, E., Alobeidli, H., Alshamsi, A., Cappelli, A., Cojocar, R., Debbah, M., Goffinet, É., et al. (2023). The falcon series of open language models. *arXiv preprint arXiv:2311.16867*.
- Alnefaie, S., Atwell, E., and Alsalka, M. A. (2023). Is gpt-4 a good islamic expert for answering quran questions? In *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*.
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. *arXiv preprint, arXiv:2005.14165*.
- Chelli, M., Descamps, J., Lavoué, V., Trojani, C., Azar, M., Deckert, M., Raynier, J.-L., Clowez, G., Boileau, P., and Ruetsch-Chelli, C. (2024). Hallucination rates and reference accuracy of chatgpt and bard for systematic reviews: Comparative analysis. In *Journal of Medical Internet Research*, volume 26, page e53164.
- El Ganadi, A., Vigliermo, R. A., Sala, L., Vanzini, M., Ruoizzi, F., and Bergamaschi, S. (2023). Bridging islamic knowledge and ai: Inquiring chatgpt on possible categorizations for an islamic digital library (full paper). In *CEUR Workshop Proceedings*, volume 3536, pages 21–33.
- Gallifant, J., Fiske, A., Strekalova, Y. A. L., Osorio-Valencia, J. S., Parke, R., Mwavu, R., Martinez, N., Gichoya, J. W., Ghassemi, M., Demner-Fushman, D., et al. (2024). Peer review of gpt-4 technical report and systems card. *PLOS Digital Health*, 3(1):e0000417.
- Gao, L., Dai, Z., Pasupat, P., Chen, A., Chaganty, A. T., Fan, Y., Zhao, V. Y., Lao, N., Lee, H., Juan, D.-C., et al. (2022). Rarr: Researching and revising what language models say, using language models. *arXiv preprint arXiv:2210.08726*.
- Ilyas, M. (2018). A multilingual datasets repository of the hadith content. *International Journal of Advanced Computer Science and Applications*, 9(2):165–172.
- Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. (2023a). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.
- Ji, Z., Liu, Z., Lee, N., Yu, T., Wilie, B., Zeng, M., and Fung, P. (2023b). RHO: Reducing hallucination in open-domain dialogues with knowledge grounding. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4504–4522, Toronto, Canada. Association for Computational Linguistics.
- Kamali, M. H. (2014). *A Textbook of Hadith Studies: Authenticity, Compilation, Classification and Criticism of Hadith*. Kube Publishing Ltd.
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.
- Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.
- Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegrefe, S., Alon, U., Dziri, N., Prabhunoye, S., Yang, Y., Gupta, S., Majumder, B. P., Hermann, K., Welleck, S., Yazdanbakhsh, A., and Clark, P. (2023). Self-refine: Iterative refinement with self-feedback.
- Mghari, M., Bouras, O., and Hibaoui, A. E. (2022). Sanad-set 650k: Data on hadith narrators. *Data in Brief*, 44:108540.
- Mündler, N., He, J., Jenko, S., and Vechev, M. (2023). Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. In *arXiv preprint*, volume arXiv:2305.15852.
- Qiu, Y., Embar, V., Cohen, S. B., and Han, B. (2023). Think while you write: Hypothesis verification promotes faithful knowledge-to-text generation. In *arXiv preprint*, volume arXiv:2311.09467.
- Reddy, G. P., Kumar, Y. V. P., and Prakash, K. P. (2024). Hallucinations in large language models (llms). In *2024 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream)*. IEEE.
- Rizqullah, M. R., Purwarianti, A., and Aji, A. F. (2023). Qasina: Religious domain question answering using sirah nabawiyah. In *2023 10th International Conference on Advanced Informatics: Concept, Theory and Application (ICAICTA)*. IEEE.
- Siino, M. and Tinnirello, I. (2024). Gpt hallucination detection through prompt engineering. In *Working Notes of CLEF*.
- Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., et al. (2024). Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*.
- Varshney, N., Yao, W., Zhang, H., Chen, J., and Yu, D. (2023). A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation. In *arXiv preprint*.
- Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N. A., Khashabi, D., and Hajishirzi, H. (2023). Self-instruct: Aligning language models with self-generated instructions.