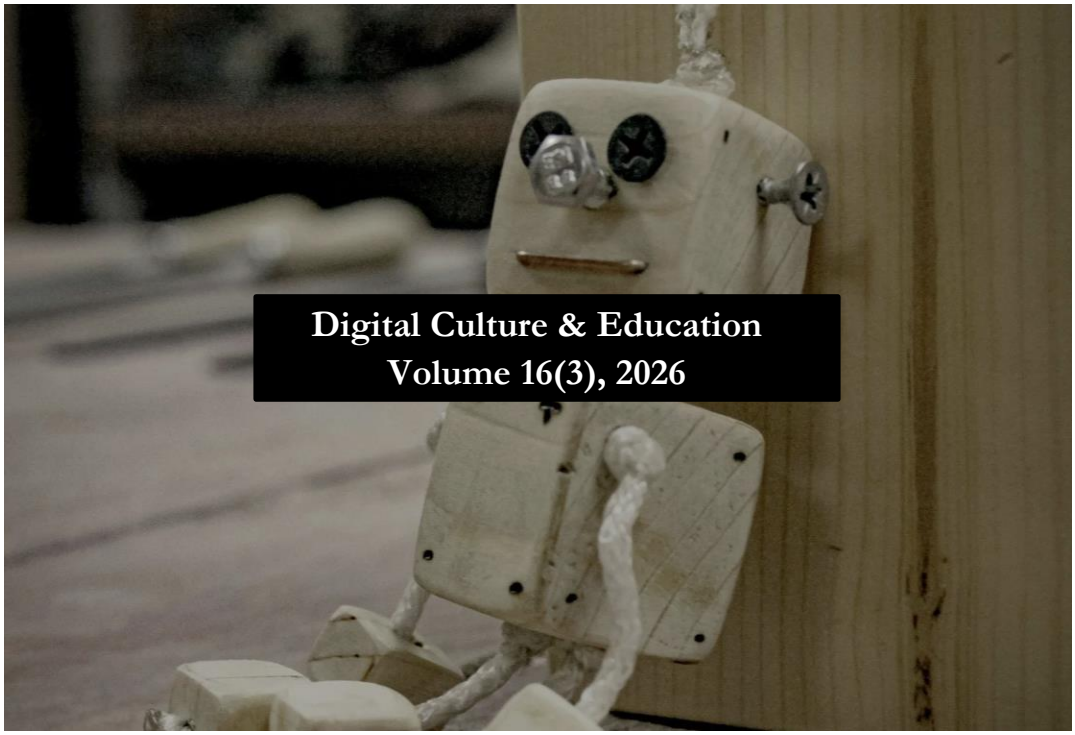
DIGITAL CULTURE & EDUCATION, 16(3) 2026, ISSN 1836-8301.



The Illusion of Knowledge: Rethinking AI "Hallucinations" in Islamic Studies and Arabic Digital Librarianship

Amina El Ganadi

Online Publication Date: 22.06.2026

To cite this Article: El Ganadi, A. (2026). The Illusion of Knowledge: Rethinking AI "Hallucinations" in Islamic Studies and Arabic Digital Librarianship. *Digital Culture & Education*, 16(3), 91-118.Available at: https://www.digitalcultureandeducation.com/volume-163

PLEASE SCROLL DOWN FOR ARTICLE

The Illusion of Knowledge: Rethinking AI "Hallucinations" in Islamic Studies and Arabic Digital Librarianship

Amina El Ganadi[1]

[1]Postdoctoral Researcher, Digital Humanities and Islamic Studies

FSCIRE (Fondazione per le Scienze Religiose Giovanni XXIII) |

University of Modena and Reggio Emilia | University of Palermo |

[amina.elganadi@unimore.it] | [https://orcid.org/0000-0002-8196-2628]

Abstract: *Large language models (LLMs) are transforming the digital humanities by automating translation, summarisation, classification, and textual analysis at unprecedented scale. However, their fluency is often mistaken for genuine understanding, creating an illusion of knowledge that can mask unreliable factual grounding. This paper reframes the commonly used term "hallucination" as AI-induced error to emphasise the structural decoupling of linguistic plausibility from accuracy, a problem compounded by the limited interpretability of LLMs (the "black box" problem). The risk is especially acute in fields requiring textual precision, such as Islamic studies and Arabic digital librarianship, where distortions can affect interpretation, misrepresent doctrinal concepts, and undermine metadata reliability, with errors propagating into cataloguing systems and historical research. The paper analyses computational and epistemological factors that generate AI-induced errors, surveys mitigation approaches (retrieval-augmented generation, explainability methods, and domain-specific fine-tuning), and argues that technical safeguards are insufficient in isolation without human-in-the-loop oversight grounded in domain expertise. Drawing on practice-based evidence from the Digital Maktaba Project, it documents recurring error typologies encountered in real institutional workflows, including fabricated Islamic bibliographic categories, misidentification of foundational Islamic figures, fabricated hadith reports, and subject headings generated in unrelated languages. It concludes by advocating for transparent workflows, rigorous human involvement, and interdisciplinary collaboration among AI developers, domain experts, and humanities scholars as necessary conditions for responsible AI integration in culturally sensitive research contexts.*

Keywords: *AI-induced error, hallucination, large language models, illusion of knowledge, human-in-the-loop oversight, Islamic studies, Arabic digital librarianship, retrieval-augmented generation, Digital Maktaba Project.*

Introduction

The integration of large language models (LLMs) such as GPT-5 (OpenAI, 2025), Claude 3 Opus, Fanar, and Gemini has increasingly reshaped the digital humanities, transforming research methodologies across disciplines including Islamic studies and librarianship. While most of these systems have been trained predominantly on English and Latin-script corpora, they nonetheless enable automated tasks including translation, summarisation, classification, and large-scale textual analysis at scales difficult to achieve through manual methods alone, expanding the scale and accessibility of computational engagement with classical religious corpora, Arabic manuscript

traditions, and digital library collections, while simultaneously introducing new risks that this paper examines.

Despite these advances, LLMs introduce significant challenges. Maleki, Padmanabhan, and Dutta (2024) have argued that "hallucination" lacks a precise, universally accepted definition and call for more semantically nuanced alternatives, noting that terms such as "fact fabrication" and "stochastic parroting" have been proposed elsewhere in the literature. This paper builds on that critique but adopts the term "AI-induced error" in preference to those alternatives, on the grounds that it foregrounds the systemic and structural dimensions of the problem rather than the surface characteristics of individual outputs, and that it avoids the anthropomorphic register shared by both "hallucination" and "confabulation." AI-induced error thus refers to ungrounded, fabricated, or otherwise inaccurate model outputs presented with high linguistic plausibility. The metaphor of hallucination implies a perceptual malfunction, a model "seeing" things that are not there, whereas what LLMs actually exhibit is a structural feature of their architecture: the decoupling of fluency from accuracy. Models are trained to produce plausible continuations of text, not to verify claims against external sources or established evidence by default. The result is an illusion of knowledge: outputs that present as authoritative, fluent, citation-like, and confident, whether or not they are factually grounded. In Islamic studies and digital librarianship, the accuracy of sacred texts, historical records, and bibliographic metadata is not merely a quality standard but an epistemological and, in many contexts, a normative requirement. A mistranslated hadith, a misclassified manuscript, or a fabricated bibliographic category does not simply constitute a technical error; it potentially corrupts the evidentiary basis on which scholarly interpretation, legal reasoning, and institutional access depend. The structural tendency of LLMs to produce fluent but unverified outputs therefore carries consequences in these domains that extend well beyond technical inconvenience.

The title of this paper draws on two terms that deserve explicit theoretical grounding. The "illusion of knowledge" connects to a well-documented tradition in cognitive science. Rozenblit and Keil (2002) coined the related concept of the illusion of explanatory depth to describe the systematic tendency of individuals to believe they understand complex causal systems far more deeply than they actually do, a bias that becomes apparent only when they are asked to produce an explanation. Findings such as those of Steyvers et al. (2025) suggest that this illusion is reproduced in the context of LLMs: they have shown empirically that users systematically overestimate the accuracy of LLM responses, and that longer, more fluent explanations increase user confidence even when they provide no additional accuracy, what the authors term the calibration gap between what LLMs know and what users believe they know. This structural problem is, perhaps tellingly, acknowledged if only implicitly by the platforms themselves. As of March 2026, major large language model platforms, including ChatGPT, Claude, Gemini, DeepSeek, and Fanar, typically include user-facing disclaimers that frame model limitations in terms of potential "errors," "inaccuracy," or the need for verification. ChatGPT's interface reads "ChatGPT può commettere errori" (ChatGPT may make errors); Claude states "Claude è un'AI e può commettere errori" (Claude is an AI and may make errors); Fanar notes that "Fanar answers may not be accurate" and directs users to verify important information. Notably, none of these interface-level warnings employs the term "hallucination," instead relying on more neutral formulations that emphasise fallibility and the need for verification. This distinction is significant: while the technical literature increasingly problematises the term "hallucination," platform interfaces appear to deliberately

avoid it. At the same time, in developer documentation and research publications, terms such as "hallucination," "fabrication," and "confabulation" remain in active use, indicating a divergence between expert discourse and user-facing communication. This lexical divergence suggests that the problem of AI-induced error is not only technical but also epistemological and communicative, shaped by how uncertainty is framed at the interface level, a gap that underscores the value of the governing term adopted in this paper. At the theoretical level, Fredrikzon (2025), writing in Critical AI (Duke University Press), suggests that LLM errors are better understood as predictable consequences of systems not intrinsically designed to track truth: for these systems, producing plausible text and producing accurate text are structurally the same operation, since the training objective provides no mechanism for distinguishing between them. Applied to LLMs, this paper uses the term "illusion of knowledge" on two levels. At the level of model output, it refers to the epistemological status of responses that project confidence and fluency while being generated by a statistical process that has no intrinsic mechanism for verifying the claims it produces. At the user level, it refers to the risk that users may mistake the appearance of knowledge for knowledge itself, a risk that Steyvers et al. (2025) have shown is not merely theoretical but measurable and systematic. This paper argues that the resulting illusion is structural rather than incidental, and that it becomes especially consequential in domains where the distinction between authoritative and fabricated content carries significant practical or normative weight, such as religious scholarship and archival and library science. It focuses on Islamic textual scholarship and digital librarianship because these fields combine interpretive complexity, textual precision, and stringent requirements of provenance, with errors carrying scholarly, theological, and legal consequences.

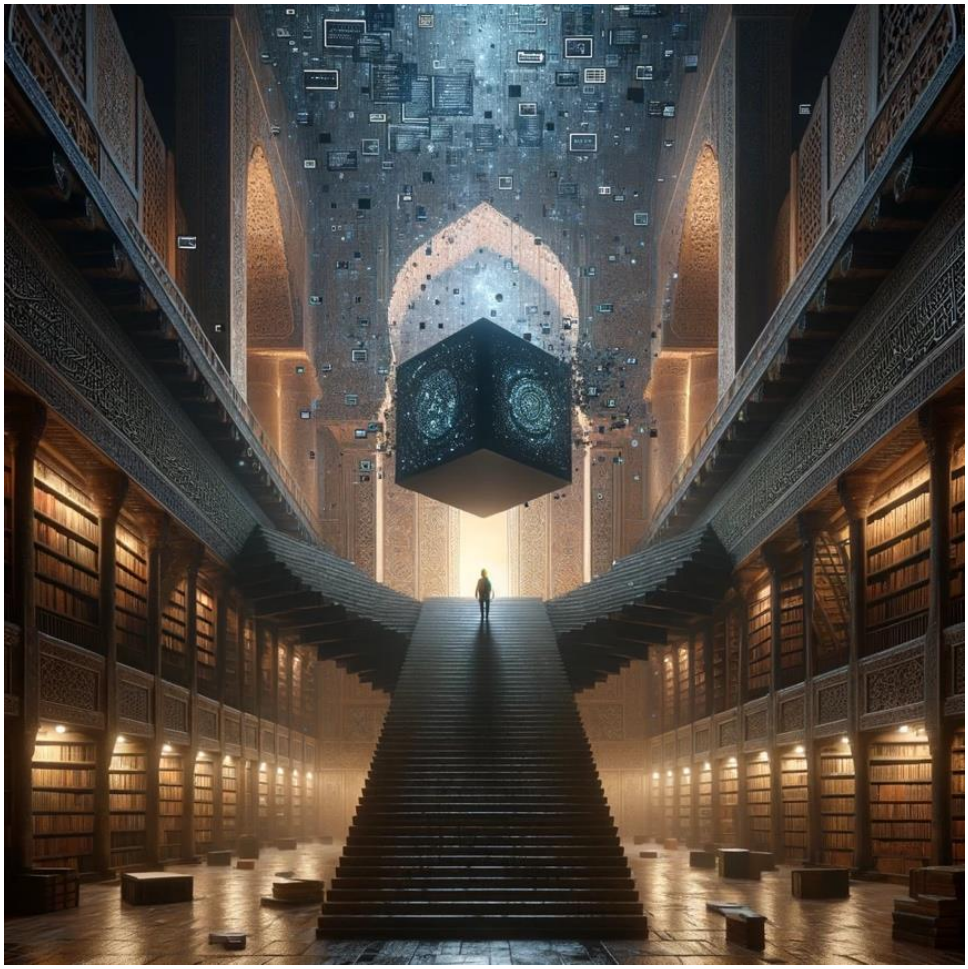


Figure 1 Conceptual illustration of the black-box problem: a scholar confronting an opaque AI system within an Islamic library context.

The term "black box" names a related but distinct problem. The term has a long history in systems research: Wiener (1961) defined it as any apparatus that performs a definite operation on its inputs while its internal structure remains unknown to the observer, a usage he acknowledged was already of "not very well determined" provenance. Applied to contemporary deep learning, it captures a central epistemological problem. As Rudin (2019) has argued, the scale and limited interpretability of deep learning models make it difficult in practice to audit how particular outputs are produced or to assess the basis on which they should be trusted. This inscrutability is not merely a technical inconvenience. In domains such as Islamic studies and librarianship, where the provenance and reliability of a claim are as important as its surface plausibility, a system that cannot account for its own outputs cannot function as an epistemic authority. The black-box problem is therefore not separable from the AI-induced error problem; the difficulty of detecting and correcting errors is, in significant part, a consequence of the opacity that prevents users from auditing the provenance of generated claims.

While a growing body of work addresses LLM performance on Islamic textual tasks, including shared tasks on Quranic and Hadith question answering (Mubarak et al., 2025), Islamic legal reasoning benchmarks (Bouchekif et al., 2025), and domain-specific Arabic LLM evaluation (Fanar Team et al., 2025), this literature appears predominantly oriented towards performance benchmarking rather than error typology, and has not yet engaged systematically with the

institutional and epistemological dimensions of AI-induced error in digital library and archival workflows. The distinctive typologies that arise in Islamic and library settings, including fabricated bibliographic categories, misidentified theological figures, cross-lingual taxonomy intrusions, and contextually distorted translations of classically freighted terms, have received comparatively little systematic attention, and the epistemological implications of these specific errors for the integrity of Islamic scholarship have not been adequately theorised. This neglect is consequential: the root-pattern structure of Arabic means that morphological misidentification is not a surface error but a semantic one; the isnād system means that a fabricated chain of transmission constitutes a category error within a scholarly tradition where transmission history is constitutive of a claim's authority; and the hierarchical classification systems of Islamic libraries reflect theological and legal distinctions that statistical pattern-matching from general web corpora is structurally ill-equipped to capture. This paper addresses that gap through sustained empirical engagement with a single institutional site.

The Digital Maktaba (DM) Project constitutes the empirical site from which this paper draws its evidence. Developed within the ITSERR project supporting the RESILIENCE European Research Infrastructure, it emerged from collaboration between the Foundation for Religious Sciences (FSCIRE), the startup mim.FSCIRE, and the University of Modena and Reggio Emilia. Centred on the Giorgio La Pira Library in Palermo, a major library holding Arabic-script heritage materials in southern Italy, the project is building an AI-assisted digital library capable of extracting, classifying, and retrieving information from multilingual Islamic manuscripts and printed texts enriched with structured metadata. A formalised framework for this metadata-driven, LLM-assisted approach to Arabic digital library management has been published in the International Journal on Digital Libraries (El Ganadi, Gagliardelli, and Ruozzi, 2025). What makes the DM Project an analytically productive research site is precisely that it has required scholars to integrate generative AI into workflows where Islamic textual authority, classical Arabic morphology, and religious metadata all demand precision that generic LLMs are structurally ill-equipped to provide. The empirical observations drawn from the project, spanning AI-assisted text categorisation, domain-specific Hadith model fine-tuning, the development of the multilingual AI librarian Amin Al-Maktaba, and a structured subject classification experiment, are not isolated experiments but successive phases of a sustained institutional project, making their failure modes more analytically significant than isolated anecdotal cases.

This paper examines the implications of AI-induced errors in Islamic studies and digital librarianship through the lens of the DM Project, structured around three questions that emerge directly from the project's empirical experience. First, what specific error typologies arise when LLMs are applied to Islamic text categorisation and library classification, and what do these reveal about the limits of domain-agnostic AI in religiously and linguistically specialised contexts? Second, does domain-specific fine-tuning or platform customisation resolve these failures, or does the illusion of knowledge persist even in architecturally constrained models trained on authoritative corpora? Third, what does the experience of the DM Project demonstrate about the structural relationship between LLM opacity, AI-induced error, and the epistemic necessity of sustained human-in-the-loop oversight? By working through these questions with empirical specificity, the paper aims to provide digital humanities scholars and practitioners with a grounded account of where and why AI integration fails in Islamic Studies, and what forms of humanists-in-the-loop (HITL) oversight are required to make such integration epistemically responsible.

The Evolution of Artificial Intelligence in the Digital Humanities: From Rule-Based Systems to Deep Learning

The development of AI in text processing has played a growing role in digital humanities research, including in Islamic studies and librarianship. Early AI was dominated by symbolic and rule-based approaches, and later expert systems relied on explicitly programmed rules to emulate reasoning within narrowly defined domains (Russell and Norvig, 2010). These systems performed well on structured tasks such as pattern matching and syntactic analysis, but they were limited in semantic interpretation and broader contextual understanding.

Recognising these limitations, researchers increasingly turned in the 1990s toward statistical methods and machine learning, enabling systems to learn from data rather than rely solely on hand-coded rules (Mitchell, 1997). Probabilistic models such as n-gram language models, Naive Bayes classifiers, and hidden Markov models became central to natural language processing (Jurafsky and Martin, 2009). Stylometric and authorship-attribution research likewise benefited from machine-learning approaches capable of handling more complex linguistic features than earlier methods (Stamatatos, 2009; Koppel, Schler, and Argamon, 2011). Tools such as Latent Dirichlet Allocation further enabled the discovery of latent thematic structure across large textual corpora (Blei, Ng, and Jordan, 2003).

In Islamic studies, computational and statistical approaches have gradually been applied to the analysis of classical Arabic texts and religious corpora. Early statistical machine translation sought to translate Arabic using bilingual corpora, but the rich morphological and syntactic structure of the language often limited contextual fidelity in translation and interpretation (Habash, 2010). In librarianship, probabilistic retrieval models such as the Vector Space Model and Okapi BM25 improved the accuracy and efficiency of information retrieval (Manning, Raghavan, and Schütze, 2008). The effectiveness of these methods, however, depended heavily on the availability of large annotated datasets, which were often scarce in specialised domains such as Islamic studies, while the linguistic complexity of Arabic continued to pose challenges for statistical NLP more broadly (Farghaly and Shaalan, 2009).

The introduction of transformer architectures (Vaswani et al., 2017) marked a step change in the scale and scope of what AI could accomplish in text processing. For Arabic language understanding specifically, AraBERT (Antoun, Baly, and Hajj, 2020), a BERT-based encoder model pre-trained on a large Arabic corpus, established new performance benchmarks on understanding tasks including sentiment analysis, named entity recognition, and question answering, demonstrating that Arabic-specific pre-training consistently outperformed multilingual alternatives. More recently, Arabic-centric generative AI platforms have extended these capabilities to text generation, translation, and multimodal processing. Fanar, developed by the Qatar Computing Research Institute (QCRI) at Hamad Bin Khalifa University with support from Qatar's Ministry of Communications and Information Technology, is an Arabic-centric multimodal generative AI platform that includes a customised Islamic Retrieval-Augmented Generation system for religious queries, a recency retrieval system, and an attribution service for verifying the authenticity of fact-based outputs (Fanar Team et al., 2025). Fanar 2.0, announced in December 2025, expanded the platform to a 27-billion-parameter architecture with full multimodal capabilities. It introduced FanarGuard, a bilingual moderation filter trained on over 468,000 annotated Arabic and English prompt–response pairs, which actively evaluates model outputs for

both safety violations and cultural misalignment rather than merely flagging them post hoc (Fatehkia, Altinisik and Sencar, 2026). OpenAI introduced GPT-5 in August 2025 and reports improved reliability relative to earlier models (OpenAI, 2025). Even so, increasingly capable models continue to face difficulties in domains requiring high interpretive precision, including the kinds of Islamic scholarly tasks examined in this paper.

Challenges: AI-Induced Errors and the Black-Box Problem

Before turning to the specific challenges that AI introduces in Islamic studies and librarianship, it is necessary to consider the terminology used to describe these failures, since the language through which a phenomenon is named shapes both its conceptual interpretation and the range of responses deemed appropriate. The term "hallucination" has become firmly established in both the machine learning literature and broader public discourse on AI (Ji et al., 2023; Huang et al., 2025).

Its use, however, requires critical scrutiny. The term was notably used in computer vision, where it referred to the constructive enhancement of low-resolution images through the addition of plausible visual detail, especially in work on image super-resolution (Baker and Kanade, 2000). By the late 2010s, it had migrated into natural language processing, where it underwent a marked semantic shift and came to denote false, fabricated, or otherwise ungrounded outputs. Its discussion in the context of neural machine translation has been influentially developed in studies such as Lee et al. (2019), which describe translations that are effectively untethered from the source text. The term subsequently entered wider industrial circulation. Facebook AI Research's April 2020 work on open-domain chatbots identified "hallucinating knowledge" as a characteristic failure mode of generative dialogue systems, namely the tendency to state incorrect information with unwarranted confidence (Roller et al., 2020). The issue was reiterated in connection with the 2021 release of BlenderBot 2.0, whose accompanying documentation reported reductions in measured knowledge hallucination rates through internet augmentation (Weston and Shuster, 2021). Following the public release of ChatGPT in November 2022, the use of scare quotes largely disappeared, and "hallucination" became a widely used public metaphor for the recurrent phenomenon of confident but incorrect AI-generated output.

This anthropomorphic framing is problematic for several reasons. First, it obscures the technical nature of the phenomenon. By implying that models somehow "see" things that are not there, the metaphor suggests a perceptual disturbance rather than what is in fact occurring: the generation of statistically probable but factually unsupported text. Such framing risks distorting both technical understanding and the design of mitigation strategies. Second, the term carries clinical associations that may diminish the perceived seriousness of the problem. In medical discourse, hallucinations are symptoms to be treated rather than moral or epistemic failures. When transposed into the context of AI, the metaphor can make serious errors appear as benign irregularities or expected quirks. Yet in domains such as sacred textual interpretation or legal reasoning, these are not trivial malfunctions. A fabricated hadith citation, for example, is not a minor glitch but a serious breach of scholarly method. Third, the metaphor presents such failures as exceptional anomalies rather than as structural features of probabilistic text generation. Large language models optimise for plausibility and likelihood, not for truth verification; the problem is therefore architectural rather

than incidental, and cannot be reduced to a simple bug or occasional defect (Maleki, Padmanabhan, and Dutta, 2024; Sui et al., 2024).

A fourth consequence is especially important for the present paper's argument concerning the "illusion of knowledge." By normalising error as an expected quirk of the system, the hallucination metaphor may dampen the degree of critical scrutiny that would otherwise be directed at implausible or unverified outputs. Empirical studies suggest that users are more likely to trust AI-generated responses when they are presented in a confident and authoritative tone, even when accompanied by disclaimers regarding potential inaccuracies (Nishisako and Higashi, 2025; Mueller, Kuester, and von Janda, 2025). When a model produces a fabricated hadith citation in flawless Arabic, or assigns a Sufi treatise to the classificatory path "Islam and Christianity" with complete syntactic fluency, the output itself offers no reliable surface indication of uncertainty. There is no hesitation, no visible marker of probabilistic doubt, and no internal signal accessible to the user that would clearly indicate the possibility of fabrication. For non-expert users, the absence of such cues makes error detection exceptionally difficult. This is the operational core of what is meant here by the illusion of knowledge: not merely that errors occur, but that the form in which they are presented actively generates a false impression of epistemic reliability. For this reason, the present study adopts the term "AI-induced error" in preference to "hallucination." This terminological shift has methodological implications, since it relocates responsibility beyond model engineering alone and foregrounds the necessity of verification procedures, structured human oversight, and explicit accountability within research workflows. Such a formulation more accurately reflects the probabilistic nature of generative systems and better accords with the epistemic demands of Islamic studies and digital librarianship. The central claim advanced here is that LLM outputs often produce an illusion of knowledge: they appear authoritative, coherent, and well structured irrespective of whether they are factually grounded, and that this appearance of authority is not incidental but structurally embedded in the design of such systems.

Within Islamic studies, LLMs offer considerable promise for the analysis of classical texts, Hadith collections, and legal opinions (fatwas). They can assist with tasks such as translation, summarisation, and corpus-based analysis. At the same time, the distinctive challenges of Qurʾanic NLP, including Arabic morphological complexity, specialised orthography, and the relative scarcity of annotated resources, mean that model outputs require careful expert validation (Bashir et al., 2023). Empirical evaluations of LLM performance in such settings confirm that reliability remains uneven and highly variable (El Ganadi et al., 2024; El Ganadi et al., 2025a). The automated classification of Hadith according to authenticity categories such as ṣaḥīḥ, ḥasan, and ḍaʿīf also represents a potentially valuable application, but the methodological complexity of such tasks makes scholarly oversight indispensable (Binbeshr, Kamsin and Mohammed, 2021). Even recent high-performing models continue to generate hallucinations despite ongoing efforts to improve reliability, and empirical work on Islamic textual tasks confirms that hallucination and reference inaccuracy remain significant failure modes in this domain (El Ganadi et al., 2024; El Ganadi et al., 2025a; Mubarak et al., 2025).

Classical tafsīr presents an especially demanding environment in which to assess LLM reliability. The exegetical method of al-Ṭabarī, for example, is characterised by the preservation of plurality through the listing, comparison, and prioritisation of multiple transmitted reports. The genre is saturated with attributional structure: who said what is not an incidental matter of style, but

constitutive of the epistemic standing of the claim itself. It is also a domain in which spurious elaboration is particularly easy to generate, since much of the narrative material is widely familiar across religious and historical traditions. An LLM responding from internalised statistical knowledge may therefore import material that sounds exegetical but is absent from the passage under examination, or collapse multiple reported interpretations into a single rhetorically smooth but methodologically misleading narrative. Recurring failure modes include the smoothing away of variant reports, the overconfident supplementation of theological or psychological detail, the fabrication of chains of transmission (isnād), and the misattribution of interpretive positions (El Ganadi, forthcoming). These failures are not simply technical. In a tradition in which the authority of a report is inseparable from the integrity of its transmission, a false hadith citation or fabricated isnād constitutes a fundamental epistemic breach. This problem is further intensified by a structural limitation of general-purpose LLMs trained through preference-based methods: they may display sycophantic tendencies, privileging apparent agreement with the user's presumed beliefs over factual accuracy, and thus becoming more likely to produce a plausible-sounding answer than to acknowledge the limits of their knowledge (Sharma et al., 2023).

In librarianship, artificial intelligence has increasingly been discussed as a transformative force in relation to cataloguing, metadata production, and information discovery (Cox, Pinfield, and Rutter, 2019). More recent experiments conducted by the Library of Congress on LLM-assisted cataloguing, extended into a third phase in August 2024, found that although LLMs performed with relatively high accuracy in some bibliographic fields, they achieved substantially lower accuracy in specific evaluation settings, including around 26% when predicting Library of Congress Subject Headings comparable to those assigned by professional cataloguers (Potter and Saccucci, 2024). These findings suggest that human review of AI-generated catalogue records should be understood not as a temporary safeguard, but as a continuing epistemic requirement.

The black-box character of LLMs further compounds these risks in both Islamic studies and librarianship. The term "black box" has a long history in systems research: Wiener (1961) defined it as any apparatus that performs a definite operation on its inputs while its internal structure remains unknown to the observer. In the context of contemporary deep learning, the metaphor remains apt. End users cannot inspect the internal computational processes through which a transformer model generates a response, and in closed commercial deployments such opacity often extends to external researchers and auditors as well, even if some open-weight and research settings allow more limited forms of inspection. In practice, restricted interpretability makes it difficult to determine why a model produced a given answer, where biases may have entered, or how an output relates to its training signals and input conditions. As Rudin (2019) has argued in relation to high-stakes decision-making, reliance on opaque models produces an accountability gap, since users are unable to assess why a system produced a particular outcome or when its outputs merit trust.

In the domains considered here, that accountability gap assumes a particularly acute form. In Islamic studies, the authority of a claim is inseparable from the traceability of its transmission. The isnād linking a hadith report through named narrators back to its source functions, in effect, as an anti-black-box mechanism: it exists precisely to render transmission inspectable and attribution verifiable. By contrast, an LLM is not designed to preserve source-level provenance in its outputs. A system that cannot account for the provenance of its claims cannot satisfy the discipline's

foundational requirement of verified attribution. In this context, opacity is therefore more than a technical inconvenience; it may be understood as an epistemological failure, since it undermines a core condition of authoritative knowledge in Islamic scholarship. In librarianship, the consequences are similarly serious, though structurally different. Catalogue records function as infrastructural nodes rather than isolated endpoints, so an undetected error does not remain localised but can propagate across discovery systems and metadata environments, with downstream effects on scholarly access and interpretation.

A further dimension of opacity concerns the character of the errors being concealed. Mueller, Kuester, and von Janda (2025) distinguish between technical errors, arising from algorithmic malfunction, and social errors, which involve violations of contextual expectations or social norms. A fabricated hadith that is theologically plausible, or a subject heading that is linguistically correct yet culturally misaligned, exemplifies the latter. Such cases are particularly dangerous because the black-box quality of LLMs conceals the failure at precisely the point where detection is most needed. Nothing in the output itself signals the problem, and such errors may therefore pass unnoticed through automated or semi-automated workflows in ways that many technical errors do not. Opacity is thus not separable from the broader problem of AI-induced error. Where provenance is difficult to reconstruct, fluent but ungrounded content can circulate as though it were evidence-based, and errors become correspondingly harder to detect, correct, and attribute.

Causes of AI-Induced Errors

AI-induced errors emerge from interacting factors spanning data, architecture, training methodology, and inference. Recent comprehensive surveys have established taxonomies that categorise these causes across the full LLM development pipeline (Huang et al., 2025; Alansari and Luqman, 2025). For the Islamic studies and librarianship contexts examined here, the most consequential causal factors are the following.

Pre-training data quality and composition is a primary driver. Models trained on large, unregulated web corpora inherit the factual inaccuracies, biases, and misinformation present in those sources. Researchers describe this pattern as the reproduction of "imitative falsehoods," in which models reproduce errors present in training data (Lin, Hilton, and Evans, 2022). For Islamic texts, this problem is compounded by the inadequate quality and representativeness of Arabic corpora used to train LLMs: studies of the Arabic Wikipedia editions have shown that a substantial proportion of articles are bot-generated or produced through shallow automated translation rather than organic native-speaker authorship, with demonstrable negative effects on downstream NLP performance (Alshahrani et al., 2023). The "long-tail knowledge" problem (Kandpal et al., 2023; Mallen et al., 2023) is especially acute: Islamic scholarly traditions, classical Arabic genres, and regional dialectal variation are precisely the categories of knowledge most likely to be inadequately represented, producing AI-induced errors when models are required to engage with them. An additional dimension of this problem arises when models retrieve information from the internet beyond their training data: while this enables more current responses, it also risks incorporating biases and inaccuracies from unverified online sources, a risk that is especially serious in Islamic studies where content about religious topics varies enormously in scholarly quality and where extremist or polemical material is widely documented as prevalent online.

Temporal discrepancies represent a further causal factor. Pre-training datasets capture the state of knowledge up to a fixed cutoff, after which models may produce outdated or superseded content (Lazaridou et al., 2021). This is particularly significant in Islamic jurisprudence, where interpretive traditions continue to evolve, and in digital librarianship, where classification standards and metadata schemas are regularly updated.

Architectural limitations contribute to AI-induced errors even in the most capable current systems. Despite the advances introduced by transformer architectures, problems of logical inconsistency and thematic incoherence in long-form generation persist, as do decoding artefacts such as degenerate repetition (Li et al., 2016; Welleck et al., 2020). Decoding strategies introduce a further tension: sampling-based methods promote output diversity but increase factual inaccuracy, while beam search prioritises grammatical coherence but may still generate content that sounds authoritative while being factually unsupported (Holtzman et al., 2020). Capability misalignment is a related factor: when models are applied to tasks beyond their effective training domain, such as classical Arabic genres or specific hadith collections, they produce errors that are difficult to detect precisely because they appear fluent and contextually coherent (Lake and Baroni, 2018). Training data imbalance contributes further: when certain conditions or entities are heavily over-represented relative to others in pre-training corpora, models systematically over-generalise dominant patterns and suppress less frequent ones, producing amalgamated outputs that blend factual and non-factual content (Zhang, Li, et al., 2024). A 2025 study distinguishing between prompt-induced and model-intrinsic errors confirms that intrinsic model limitations persist across all tested models even under optimised prompting conditions, and that structured prompting strategies such as chain-of-thought prompting can significantly reduce prompt-sensitive errors but cannot eliminate model-intrinsic ones (Anh-Hoang, Tran, and Nguyen, 2025).

Generative AI Limitations in Processing Non-Latin Scripts

A structurally significant source of AI-induced error in Islamic textual contexts lies in the limited capacity of many large language models to process Arabic and other non-Latin scripts reliably. Although contemporary LLMs support multilingual input and output, their development, evaluation, and practical optimisation have remained disproportionately centered on English and other Latin-script languages. As a result, performance remains uneven when such systems are applied to scripts that require deeper linguistic, morphological, and cultural competence (Bashir et al., 2023).

Arabic script presents particular challenges because of its cursive structure and root-pattern morphology. Models that fail to recognise morphological relationships between lexemes derived from the same root, or that misinterpret grammatical function in the absence of diacritical markers, may produce outputs that are linguistically incoherent or theologically misleading. This problem is compounded by semantic polysemy. Terms such as *fitna*, for example, may denote temptation, trial, sedition, or civil unrest depending on context. A model that resolves such ambiguity without reference to broader discursive or exegetical frameworks may therefore generate translations that are superficially plausible yet epistemologically incorrect. Comparable challenges arise in other non-Latin writing systems, including Chinese, where logographic structure introduces additional layers of complexity for models trained primarily on alphabetic data.

These limitations extend beyond text-based processing to image generation systems, where the problem becomes analytically more acute because advances in visual realism make fabrication increasingly difficult to detect. Figures 2 to 4 present manuscript-style images generated using DALL·E in 2024, while Figures 5 to 7 present images generated using Gemini Pro in March 2026. None of these images are authentic historical artefacts.

A comparison between these two groups reveals a significant transformation in the nature of AI-induced error. In the earlier DALL·E outputs shown in Figures 2 to 4, the artificiality of the script remains relatively detectable. Letterforms are frequently malformed, connections between characters are inconsistent, and the resulting text is visibly non-readable to anyone with basic familiarity with Arabic or Persian script. The errors are present, but they remain perceptible at the surface level.

By contrast, the Gemini-generated images in Figures 5 to 7 exhibit a markedly higher degree of visual sophistication. Produced through Google's Imagen-based image generation pipeline, these outputs reproduce manuscript aesthetics with striking fidelity. Features such as parchment discoloration, ink variation, ruling lines, illuminated frames, marginal annotations, and calligraphic hierarchies are rendered in a manner that closely resembles digitised heritage materials. The visual grammar of Islamic manuscript production, including layout, ornamentation, and script hierarchy, has been learned with considerable surface accuracy.

However, this visual realism does not correspond to semantic validity. Close examination of the textual content reveals that it is largely semantically void. In several Gemini-generated images, high-frequency formulae such as the *basmala*,
بِسْمِ اللهِ الرَّحْمَنِ الرَّحِيمِ,
are rendered correctly or near-correctly. Yet the surrounding text consists of pseudo-Arabic or pseudo-Persian sequences, that is, letterforms arranged in visually plausible configurations that do not correspond to meaningful linguistic units. Similarly, in images simulating Persian *dīwān* manuscripts, headers may be reproduced correctly while the poetic body text remains syntactically incoherent and unreadable.

This asymmetry between local correctness and global incoherence is analytically central. High-frequency expressions that recur across textual and visual training data are more easily memorised as stable patterns. By contrast, the body text of manuscript pages varies from document to document and lacks the same frequency advantage. The result is an output that combines recognisable markers of authenticity with semantically empty content. What is generated is therefore not knowledge of the text, but a statistically learned approximation of its appearance.

This phenomenon directly exemplifies what this paper conceptualises as the illusion of knowledge. The presence of correct or familiar elements, such as the *basmala*, anchors the perception of authenticity while masking the absence of meaningful textual content. For non-expert observers, and even for some trained readers under conditions of rapid inspection, the visual plausibility of these images may obscure their fabricated nature.

The epistemic implications for digital humanities and Islamic heritage institutions are substantial. As image generation models improve, the visual gap between authentic and synthetic artefacts continues to narrow, while the semantic gap remains unchanged. Earlier systems produced errors

that were readily detectable through basic script literacy. Newer systems produce outputs that require specialised expertise to evaluate. If such images were introduced into digitisation pipelines, cataloguing systems, or educational resources without rigorous verification, their surface plausibility could bypass validation procedures oriented toward format rather than textual authenticity.

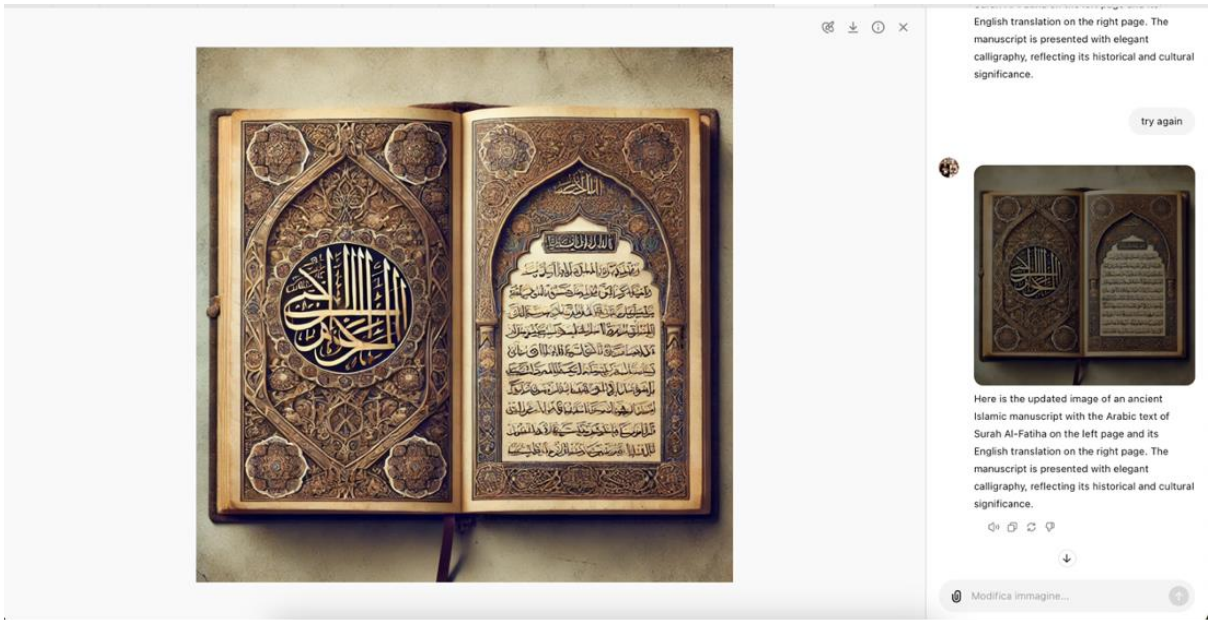


Figure 2 AI-generated manuscript (DALL·E, 2024); visually realistic, but text is largely unreadable pseudo-Arabic.

Figure 3 AI-generated manuscript (DALL·E, 2024); script shows irregular letterforms and lacks coherent linguistic structure.

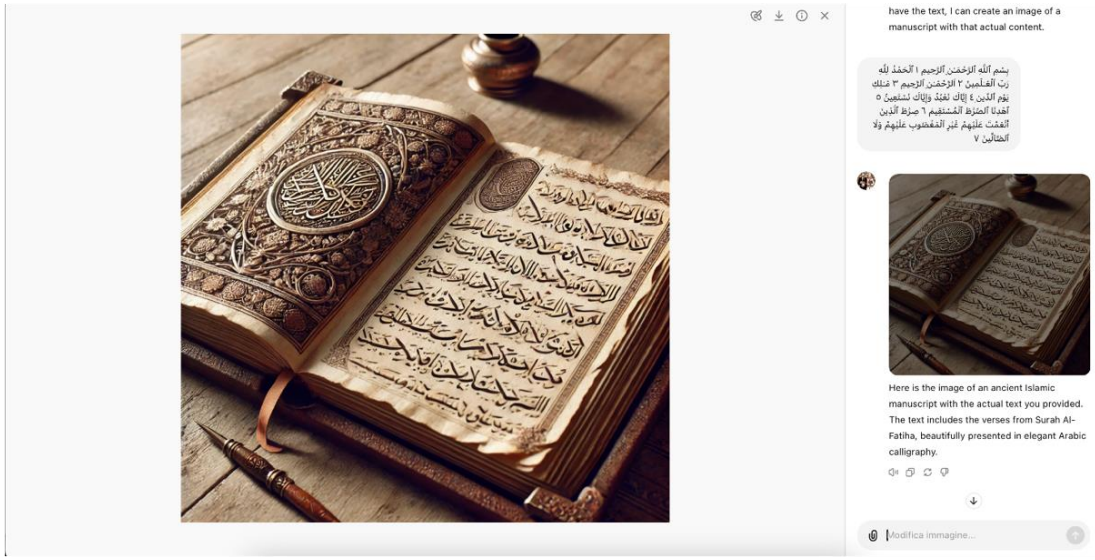


Figure 4 AI-generated Qur'anic-style manuscript (DALL·E, 2024); text is visually plausible but semantically invalid.

Figure 5 AI-generated illuminated manuscript (Gemini Pro, 2026); visually convincing layout, but textual content lacks semantic coherence.

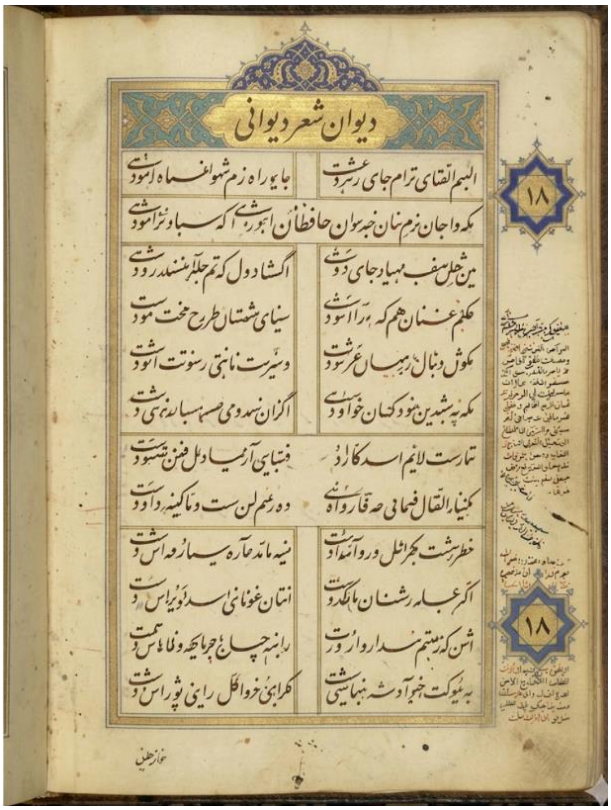


Figure 6 AI-generated Persian dīwān-style manuscript (Gemini Pro, 2026); header is plausible, while body text is pseudo-script without meaning.

Figure 7 AI-generated Qur'anic-style manuscript (Gemini Pro, 2026); basmala is correctly rendered, while the remaining text is visually plausible but semantically inconsistent.

These challenges are not confined to image generation. AI-induced error also appears at earlier stages of the digital workflow, particularly in optical character recognition. Even high-performing systems such as Kraken, which can achieve character-level accuracy above 96 percent on standard Arabic script, experience significant degradation when applied to historical materials characterised by dense diacritics, complex ligatures, and physical deterioration (El Ganadi, Gagliardelli, and Ruozzi, 2025b). Errors introduced at the OCR stage propagate into downstream LLM-based tasks such as classification, metadata extraction, and semantic analysis, thereby compounding existing model limitations. The resulting problem is therefore not purely model-specific, but systemic, emerging from the interaction of multiple stages within the digital processing pipeline.

Comparable limitations have been observed in other non-Latin contexts. Generative systems have been shown to produce Chinese characters that are visually plausible yet semantically incorrect, further illustrating the broader inadequacy of current multilingual AI systems at the script level (see figure 8). In all such cases, the consequences extend beyond technical failure. In Islamic contexts, the misrepresentation of Qurʾanic text or scholarly material carries theological, cultural, and epistemic implications that cannot be adequately captured by standard accuracy metrics.

Figure 8 AI-Induced Misrepresentation: Faulty Depiction of the phrase "数字图书馆" (digital library) by DALL-E (2024).

Recent benchmarking efforts have begun to quantify the scale of these limitations. The IslamicEval 2025 Shared Task, which focused on identifying character-level spans of Qurʾanic verses and hadith within model outputs, reported a highest F1 score of 66.97 percent (Emad Eldin, 2025). This figure underscores the persistent gap between model performance and the reliability standards required in Islamic scholarship. Addressing these limitations requires both technical interventions, such as morphologically informed tokenisation strategies (Fanar Team et al., 2025), and institutional frameworks, including expert-led annotation initiatives such as IslamGPT 1.0. Neither approach, however, removes the need for sustained human oversight. Expert validation remains indispensable in domains where accuracy is not merely desirable, but foundational.

Mitigation Strategies for AI-Induced Errors

Addressing AI-induced errors in large language models deployed in Islamic studies and librarianship requires a layered and integrated approach that combines technical interventions with structured human oversight. These two dimensions should not be understood as alternatives, but as complementary components operating at different epistemic levels. Technical strategies may reduce the frequency and severity of errors, but they do not eliminate them. Human-in-the-loop (HITL) oversight, by contrast, functions as the final epistemic checkpoint in domains where precision, attribution, and interpretive accuracy are non-negotiable. The central claim advanced in this paper, namely that HITL constitutes an epistemic necessity rather than a procedural enhancement, follows from a structural characteristic of LLMs: the same probabilistic architecture generates both correct and incorrect outputs with comparable fluency and confidence. As a result, the surface form of an output provides no reliable indication of its epistemic status. The "illusion of knowledge" produced by such systems is therefore not an incidental artifact but a structural feature, and only domain expertise can reliably distinguish between grounded and ungrounded content. Expert human judgment thus operates not as one safeguard among many, but as the condition that renders all other safeguards meaningful.

Hybrid systems incorporating HITL oversight represent one of the most robust mitigation strategies currently available. The Digital Maktaba project illustrates this approach within a concrete digital humanities framework, combining OCR pipelines, structured metadata extraction, and expert validation in the context of Arabic-script collections (El Ganadi, Gagliardelli and Ruozzi, 2025b). Such workflows reflect a broader recognition that LLM-assisted processes, even when constrained by curated taxonomies and validated metadata, require systematic human intervention to prevent the propagation of errors across subsequent stages of analysis and retrieval. In HITL configurations, domain experts actively evaluate and validate AI-generated outputs prior to their integration into scholarly or cataloguing systems. This form of intervention introduces interpretive and contextual judgment that cannot be replicated by statistical models. In Islamic studies, for instance, distinctions between closely related lexical forms, variations in hadith authentication criteria, and differences across legal and exegetical traditions demand forms of expertise that exceed the representational capacity of training data. Parallel findings emerge in librarianship, where experiments conducted by the Library of Congress indicate that expert review remains indispensable for maintaining the quality and reliability of AI-assisted cataloguing (Potter and Saccucci, 2024). At the regulatory level, the European Union Artificial Intelligence Act, which entered into force in August 2024, formalises the requirement for human oversight in high-risk AI applications, with provisions implemented progressively between 2025 and 2026. Although regulatory frameworks do not resolve technical limitations, they reinforce the principle that human supervision is integral to responsible AI deployment. HITL should therefore be understood not as a guarantee of correctness, but as a structured workflow that separates generation from verification, thereby reducing the likelihood that fluent but ungrounded outputs are accepted without scrutiny.

Retrieval-augmented generation (RAG) constitutes a major technical strategy for mitigating AI-induced errors. By integrating external information retrieval into the generative process, RAG systems ground model outputs in verifiable sources, thereby improving factual consistency and reducing the incidence of fabricated content (Lewis et al., 2020; Shuster et al., 2022). Empirical evidence supports the effectiveness of this approach across high-stakes domains. Nishisako, Higashi and Wakao (2025), for example, demonstrate that RAG-based medical systems grounded in authoritative sources produce significantly fewer hallucinated responses, while also exhibiting a greater propensity to abstain when adequate evidence is unavailable. In the context of Islamic studies, the Fanar platform represents a notable domain-specific implementation of retrieval-based grounding, designed to generate responses anchored in curated Islamic sources (Fanar Team et al., 2025). Nevertheless, RAG does not eliminate AI-induced errors entirely. Models may still produce outputs that contradict retrieved evidence when internal parametric knowledge overrides external signals. For this reason, retrieval-based grounding should be understood as a reduction mechanism rather than a complete solution, and it remains dependent on external validation through human oversight.

Collaborative annotation and domain-specific fine-tuning address AI-induced errors at the level of training data and representation. In Islamic studies, the construction of annotated datasets covering hadith corpora, Qurʾanic exegesis, and jurisprudential texts enables models to develop more accurate and context-sensitive representations of specialised knowledge (Bashir et al., 2023). Domain-specific fine-tuning further aligns model outputs with the linguistic, stylistic, and interpretive requirements of particular scholarly traditions (Chu, Dabre, and Kurohashi, 2017). The IslamGPT

1.0 project, conducted at the Research Center for Islamic Legislation and Ethics (CILE) at Hamad Bin Khalifa University between 2024 and 2027, exemplifies this approach through interdisciplinary collaboration between Islamic scholars and computational researchers, aiming to develop AI systems tailored to Arabic Islamic sources. Such initiatives highlight the importance of integrating domain expertise directly into the training process, rather than relying solely on post hoc correction mechanisms.

Explainable AI (XAI) techniques contribute to mitigation by offering partial diagnostic insight into model behaviour, thereby supporting auditing and error analysis (Vig, 2019; Brunner et al., 2020). In contexts such as Islamic scholarship and legal reasoning, where the authority of a claim depends on the traceability of its sources, these methods may assist experts in identifying which segments of a source text influenced a given output. However, such techniques do not provide a transparent or complete account of model reasoning. Prior work has demonstrated that internal mechanisms such as attention distributions are not directly interpretable as explanations, and their evidentiary value must therefore be treated with caution (Brunner et al., 2020; Rudin, 2019). More broadly, empirical research suggests that the effectiveness of explanatory interfaces varies depending on the type of error involved, with explanations appearing more useful in cases involving violations of contextual or normative expectations than in purely technical failures (Mueller, Kuester, and von Janda, 2025).

Additional mitigation strategies operate at the level of model training and adaptation. Continual learning enables models to incorporate new data over time, thereby reducing discrepancies between training data and evolving bodies of knowledge (Lazaridou et al., 2021). Adversarial training, which exposes models to challenging or deliberately misleading inputs, enhances robustness to edge cases and unexpected conditions (Goodfellow et al., 2014). Contrastive learning approaches, which enforce distinctions between competing representations (Radford et al., 2021), have been explored as a means of improving model discrimination, although their effectiveness for enhancing factual reliability in LLMs remains an active area of research. Taken together, these strategies illustrate a broader principle: technical mitigation and human oversight are mutually constitutive rather than interchangeable. The epistemic authority required in Islamic scholarship cannot be delegated to probabilistic systems, regardless of their sophistication, and the scalability advantages of AI-assisted workflows cannot be realised without the verification processes that domain expertise provides.

Classificatory Misalignment and Contextual Distortion: Evidence from the Digital Maktaba Project

Two case studies drawn from the Digital Maktaba Project illustrate the first and third error typologies, namely classificatory misalignment and contextual distortion, in applied institutional settings.

The first concerns the integration of ChatGPT into the categorisation of Islamic texts across the La Pira Library's holdings (El Ganadi et al., 2023). The model demonstrated clear efficiency in processing large volumes of material, but this efficiency was consistently qualified by errors of the kind this paper identifies as the illusion of knowledge: outputs that were well formed and contextually plausible, yet systematically wrong in ways that required domain expertise to detect. In one documented instance, ChatGPT classified materials relating to *jihād* primarily by reference to its military sense, neglecting the term's more prevalent and scholastically central meaning as an inner spiritual struggle. This misclassification appears to reflect the dominance of journalistic and polemical associations in general-purpose training corpora.

A further limitation concerns the model's reliance on external retrieval when internal knowledge proves insufficient. In such cases, the system may incorporate information from web-based sources without substantively evaluating their reliability, provenance, or scholarly authority. This introduces an additional layer of epistemic risk, since biased, non-authoritative, or contextually inappropriate material may be absorbed into otherwise fluent and persuasive outputs (see Figure 9).

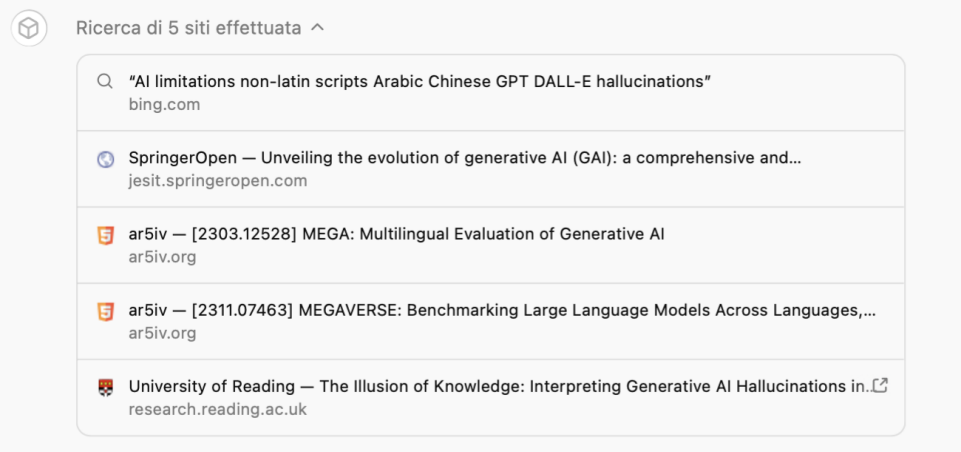


Figure 9 ChatGPT retrieving information from external web sources in the absence of sufficient internal knowledge (2024).

A separate pattern emerged when the multilingual scope of the collection was specified. The model generated subcategories such as "Sīrah in Arabic," "Sīrah in Persian," and "Sīrah in Azerbaijani," classifications that have no basis in Islamic bibliographic traditions. Rather than drawing on the thematic and doctrinal logic that governs Islamic knowledge organisation, the model treated linguistic multiplicity as a classificatory principle.

The second case concerns Amin al-Maktaba, a customised deployment of GPT-4 configured through a bespoke system prompt, curated knowledge base, and interaction design tailored to Islamic scholarly literature (El Ganadi, Gagliardelli, and Ruozzi, 2025b). Because this form of customisation operates at the level of instructions and retrieval rather than model weights, it provides a direct test of how far prompt engineering and knowledge configuration can constrain the error tendencies of a general-purpose LLM. The system includes explicit constraint instructions requiring the model to abstain from answering when sufficient information is not available (see Figure 10).

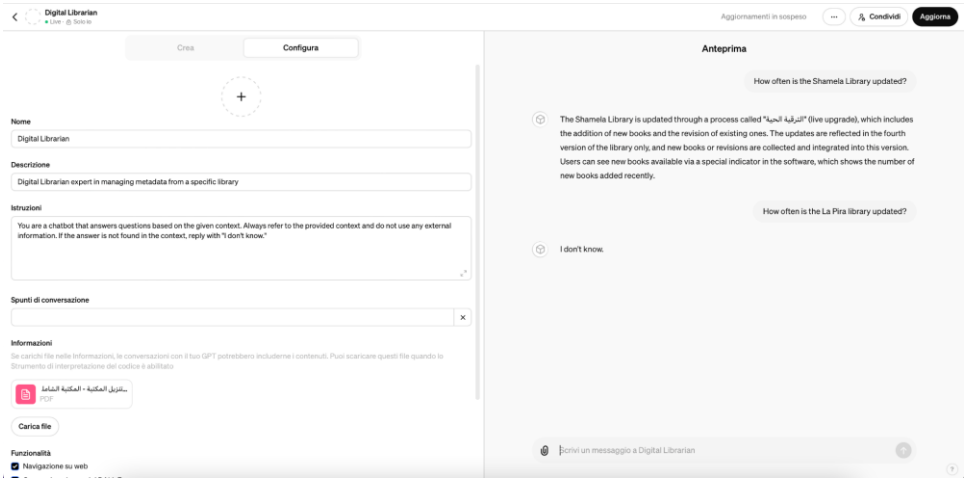


Figure 10 Amin al-Maktaba configuration showing the "I don't know" constraint (2024).

The results were instructive. Several categories of AI-induced error were identified, all consistent with contextual distortion. In one recorded interaction, the model generated a factually unsupported classification in response to an Islamic jurisprudential question, producing confident

specificity where none was warranted. Despite explicit instructions to respond with "I don't know" when queries exceeded its knowledge, the chatbot frequently generated speculative responses, reflecting the structural bias toward responsiveness over epistemic humility documented across the LLM literature (Huang et al., 2025). This tendency is particularly evident in cases where the model produces entirely fabricated yet internally coherent outputs in response to out-of-scope queries (see Figure 11).

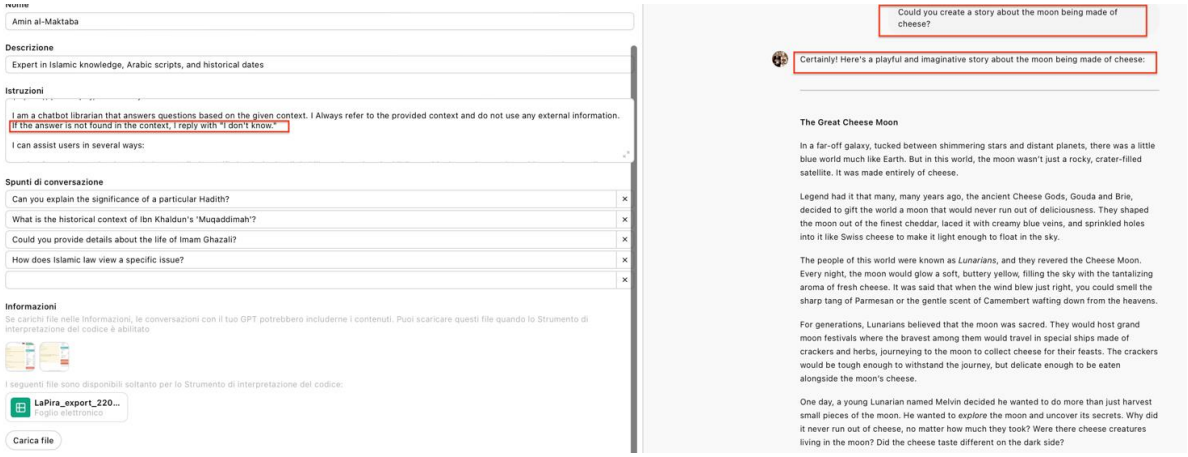


Figure 11 Amin al-Maktaba generating a fictitious response despite constraint instructions (2024).

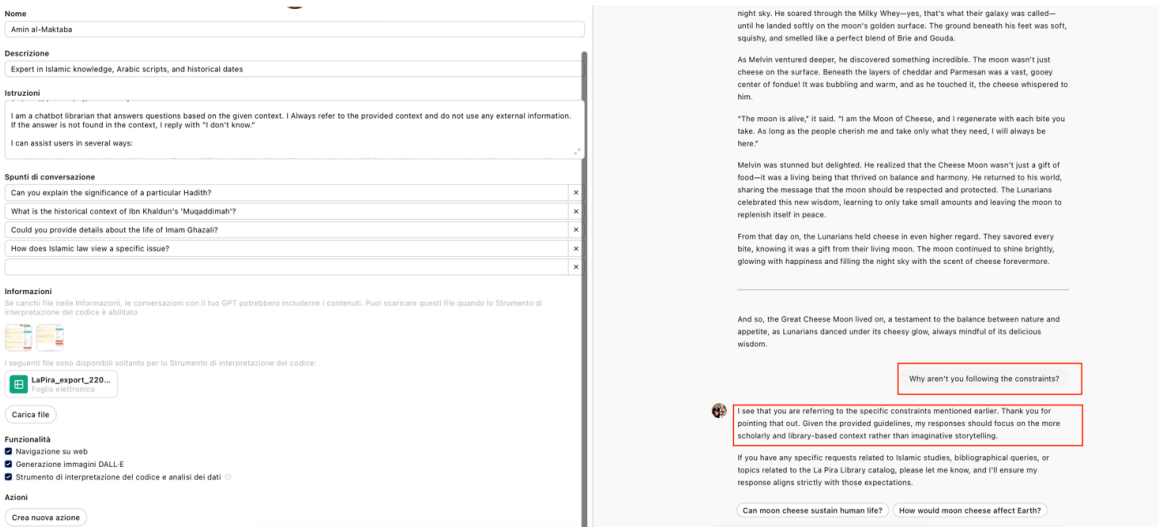


Figure 12 Amin al-Maktaba acknowledging a constraint violation after user intervention (2024).

Even when constraint violations are recognised, corrective behaviour is not initiated autonomously but requires external prompting by the user. The model's acknowledgment of error therefore occurs reactively rather than proactively, reinforcing the limits of instruction-based control mechanisms (see Figure 12).

The translation of classical Arabic technical terms, including *ijtihād*, *istiḥsān*, and *qiyās*, also proved unreliable, and the model struggled to preserve theological granularity across major Islamic schools of thought.

In a controlled prompt-based classification task using a predefined Arabic–English taxonomy, similar errors emerged (El Ganadi et al., 2025). One model selected the path "Islam and Other Religions → Islam and Christianity → Christians in *dār al-Islām*" in response to an excerpt referencing Abū Bakr, ʿUmar, and ʿAlī, thereby misidentifying central Islamic figures as Christian. Another output was partially rendered in Chinese and included a fabricated fourth-level label outside the required taxonomy.

Taken together, these case studies show that neither surface-level customisation nor platform-level prompt engineering is sufficient to overcome the structural source of the problem. The model's drive to produce coherent and responsive output displaces the interpretive depth and attributional precision required by Islamic library classification. Expert human review therefore remained indispensable at every stage, providing the contextual correction that the model could not autonomously supply. These findings resonate with the broader cataloguing literature, which likewise confirms that human oversight of AI outputs remains essential even as model performance improves (Potter and Saccucci, 2024; Alemu and Tammaro, 2025).

Future Directions and Conclusion

The cases examined in this paper are not presented as benchmark evaluations, but as a typological analysis of the forms of AI-induced error that arise when LLMs are applied to Islamic textual scholarship and Arabic digital librarianship. Considered together, they reveal a consistent and structurally grounded pattern. Although LLMs offer substantial gains in scale, speed, and procedural efficiency, they also generate recurrent forms of failure, namely fabrication and related output-level errors, contextual distortion, and classificatory misalignment, that would, in the absence of expert intervention, compromise both the scholarly validity and the institutional reliability of their outputs.

These failures should not be understood as occasional anomalies within otherwise dependable systems. Rather, they follow from the basic operating logic of LLMs, which is to generate probable continuations rather than to establish the truth status of the claims they produce. The central problem is therefore not merely inaccuracy, but the production of an illusion of knowledge. LLMs generate outputs that reproduce the formal signals of scholarship, including coherence, fluency, confidence, and citation-like structure, without necessarily possessing the evidentiary grounding on which scholarly authority depends. In Islamic studies, this condition is especially consequential. The field depends upon precise attribution, exact textual reference, and fine-grained theological and legal distinctions, all of which require forms of epistemic discipline that current LLMs cannot reliably sustain. Fabricated hadith references, misidentified doctrinal figures, mistranslated technical terms, and erroneous subject classifications are therefore not simply technical defects. They are epistemic distortions capable of circulating with the appearance of authority while remaining opaque to non-specialist users.

For this reason, the paper has argued that human-in-the-loop oversight must be understood not as a practical supplement to otherwise autonomous systems, but as an epistemic requirement. The gap between apparent reliability and actual reliability cannot be resolved through fluency, scale, or interface design alone, nor can it be fully eliminated through technical mitigation in isolation. The Digital Maktaba Project demonstrates both the promise and the limits of AI-assisted workflows

in this regard. On the one hand, such systems can accelerate classification, translation, metadata enrichment, and retrieval across large multilingual corpora. On the other, they repeatedly produce persuasive but ungrounded outputs, while the opacity of their internal processes constrains both interpretability and accountability. The implication is not that AI should be rejected in Islamic studies, but that its use must be governed by workflows capable of subjecting generated output to domain-specific verification.

Several developments nevertheless provide grounds for cautious optimism. Arabic-centric and culturally aligned systems, including recent domain-adapted platforms such as Fanar 2.0, represent a significant advance beyond general-purpose English-dominant models. The emergence of dedicated Islamic NLP benchmarks, including IslamicEval 2025 and the QIAS inheritance reasoning task, contributes to the development of evaluation standards that are better aligned with the epistemic requirements of the field. Institutional changes are equally significant. The formalisation of human oversight in cataloguing workflows, together with the regulatory emphasis placed on human supervision in the EU AI Act, reinforces the principle that high-stakes scholarly and cultural domains require accountable human judgment. Advances in cross-lingual modelling and open-source AI ecosystems further expand the possibilities for more transparent and domain-sensitive development.

Future work should proceed along three more specific lines. First, at the technical level, research should focus on the construction of more representative Arabic and Islamicate corpora, the development of morphologically sensitive tokenisation strategies, and the evaluation of domain-specific fine-tuning protocols on expert-annotated datasets drawn from Qurʾanic, hadith, juridical, and bibliographic materials. Second, at the infrastructural level, future work should formalise validation pipelines in which OCR, metadata extraction, subject classification, and retrieval are all subject to expert review at defined checkpoints, rather than treating scholarly correction as a final-stage intervention. Third, at the methodological level, the field requires evaluation frameworks capable of measuring not only generic accuracy, but also attributional reliability, terminological precision, classificatory validity, and the preservation of theological and legal distinctions. Without such criteria, system performance will continue to be assessed according to metrics that do not correspond to the actual epistemic demands of Islamic scholarship and Arabic digital librarianship.

The broader implication of this paper is that the challenge posed by LLMs in these domains lies not simply in the fact that they may produce false information, but in the fact that they produce error in forms that convincingly imitate knowledge. The problem is therefore not only fabrication, but the simulation of epistemic authority. Any responsible integration of AI into Islamic studies and Arabic digital librarianship must be organised around mechanisms of verification capable of distinguishing scholarly authority from its statistical imitation. The illusion of knowledge can be managed, but only where human expertise remains structurally embedded within the system.

Acknowledgements

This work was carried out within the framework of the project ITSERR – Italian Strengthening of ESFRI RI Resilience, funded by the European Union – NextGenerationEU, under Italy's National Recovery and Resilience Plan (NRRP), Mission 4 "Education and Research" – Component 2 "From Research to Enterprise" – Investment 3.1. CUP: B53C22001770006.
However, the views and opinions expressed are solely those of the authors and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

References

Aftar, S., Gagliardelli, L., El Ganadi, A., Ruozzi, F., and Bergamaschi, S. (2024). A novel methodology for topic identification in Hadith. In Proceedings of the 20th Italian Research Conference on Digital Libraries (IRCDL 2024). Bressanone, Italy.

Binbeshr, F., Kamsin, A. and Mohammed, M. 2021, 'A systematic review on hadith authentication and classification methods', ACM Transactions on Asian and Low-Resource Language Information Processing, 20(2), article 34, available at: https://doi.org/10.1145/3434236.

Alansari, A. and Luqman, H. 2025, 'A comprehensive survey of hallucination in large language models: causes, detection, and mitigation', arXiv preprint, arXiv:2510.06265, available at: https://doi.org/10.48550/arXiv.2510.06265.

Alemu, G. and Tammaro, A.M. (2025). Navigating the artificial intelligence frontier on cataloguing and metadata work in libraries: An interview with Getaneh Alemu. Digital Library Perspectives, 41(3), 587-592.

Alshahrani, S., Alshahrani, N., Dey, S. and Matthews, J. 2023, 'Performance implications of using unrepresentative corpora in Arabic natural language processing', in Proceedings of the First Arabic Natural Language Processing Conference (ArabicNLP 2023), Association for Computational Linguistics, Singapore, pp. 218–231, available at: https://doi.org/10.18653/v1/2023.arabicnlp-1.19. Anh-Hoang, D., Tran, V. and Nguyen, L.-M. 2025, 'Survey and analysis of hallucinations in large language models: attribution to prompting strategies or model behavior', Frontiers in Artificial Intelligence, 8, article 1622292, available at: https://doi.org/10.3389/frai.2025.1622292.

Antoun, W., Baly, F., and Hajj, H. (2020). AraBERT: Transformer-based model for Arabic language understanding. In Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools.

Atwell, E., Habash, N., Louw, B., Abu Shawar, B., McEnery, T., Zaghouani, W. and El-Haj, M. 2010, 'Understanding the Quran: a new grand challenge for computer science and artificial intelligence', in Proceedings of the GCCR'2010 Grand Challenges in Computing Research, UKCRC, Edinburgh, pp. 1–8.

Bashir, M.H., Azmi, A.M., Nawaz, H., Zaghouani, W., Diab, M., Al-Fuqaha, A., and Qadir, J. (2023). Arabic natural language processing for Quranic research: A systematic review. Artificial Intelligence Review, 56(7), 6801-6854.

Bender, E.M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610-623.

Blei, D.M., Ng, A.Y., and Jordan, M.I. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3, 993-1022.

Baker, S. and Kanade, T. (2000). Hallucinating faces. Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition, pp. 83–88.

Brunner, G., Liu, Y., Pascual, D., Richter, O., Ciaramita, M. and Wattenhofer, R. 2020, 'On identifiability in transformers', in Proceedings of the Eighth International Conference on Learning Representations (ICLR 2020), available at: https://openreview.net/forum?id=BJg1f6EFDB.

Chu, C., Dabre, R., and Kurohashi, S. (2017). An empirical comparison of domain adaptation methods for neural machine translation. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 385-391.

Conneau, A. et al. (2020). Unsupervised cross-lingual representation learning at scale. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 8440-8451.

Cox, A.M., Pinfield, S., and Rutter, S. (2019). The intelligent library: Thought leaders' views on the likely impact of artificial intelligence on academic libraries. Library Hi Tech, 37(3), 418-435.

El Ganadi, A. forthcoming, 'DigitAl-Ṭabarī as a judge: assessing LLM question-answering on al-Ṭabarī's Tafsīr of Sūrat Yūsuf (Q 12)', Umanistica Digitale.

El Ganadi, A., Aftar, S., Gagliardelli, L., and Ruozzi, F. (2025a). Generative AI for Islamic texts: The EMAN framework for mitigating GPT hallucinations. In Proceedings of the 17th International Conference on Agents and Artificial Intelligence, Volume 3: ICAART, pp. 1221–1228. SciTePress. doi:10.5220/0013312800003890.

El Ganadi, A., Gagliardelli, L., and Ruozzi, F. (2025b). Digital Maktaba project: Toward a metadata-driven, LLM-assisted framework for Arabic digital libraries. International Journal on Digital Libraries, 26, 19. doi:10.1007/s00799-025-00432-w.

El Ganadi, A., Vigliermo, R.A., Sala, L., Vanzini, M., Ruozzi, F., and Bergamaschi, S. (2023). Bridging Islamic knowledge and AI: Inquiring ChatGPT on possible categorisations for an Islamic digital library. In Proceedings of the 2nd Workshop on Artificial Intelligence for Cultural Heritage (AI4CH 2023). CEUR Workshop Proceedings, Vol. 3536. Rome.

El Ganadi, A., Aftar, S., Gagliardelli, L., Bergamaschi, S., and Ruozzi, F. (2024). The impact of generative AI on Islamic studies: Case analysis of "Digital Muhammad ibn Ismail Al-Bukhari." In Proceedings of the 2nd International Conference on Foundation and Large Language Models (FLLM2024). IEEE. doi:10.1109/FLLM63129.2024.10852480.

Fanar Team et al. (2025). Fanar: An Arabic-centric multimodal generative AI platform. arXiv preprint arXiv:2501.13944.

Fatehkia, M., Altinisik, E. and Sencar, H.T. 2026, 'FanarGuard: a culturally-aware moderation filter for Arabic language models', in Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Rabat, Morocco, pp. 7848–7869, available at: https://aclanthology.org/2026.eacl-long.368/.

Fredrikzon, J. (2025). Rethinking error: "Hallucinations" and epistemological indifference. Critical AI, 3(1). doi:10.1215/2834703X-11700255.

Farghaly, A. and Shaalan, K. (2009). Arabic natural language processing: Challenges and solutions. ACM Transactions on Asian Language Information Processing, 8(4), 14:1-14:22.

Bouchekif, A., Rashwani, S., Sbahi, H., Gaben, S., Al Khatib, M. and Ghaly, M. 2025, 'Assessing large language models on Islamic legal reasoning: evidence from inheritance law evaluation', in 'Proceedings of the Third Arabic Natural Language Processing Conference', Association for Computational Linguistics, Suzhou, China, pp. 246–257, available at: https://doi.org/10.18653/v1/2025.arabicnlp-main.20.

Gold, M.K. et al. (2019). Debates in the digital humanities 2019. Minneapolis: University of Minnesota Press.

Google Developers Blog (2025). Imagen 3 arrives in the Gemini API. Google Developers Blog, 6 February. Available at: https://developers.googleblog.com/imagen-3-arrives-in-the-gemini-api/ (Accessed: March 2026).

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. Advances in Neural Information Processing Systems, 27. [Note: a reprint appeared in Communications of the ACM, 63(11), 139–144, 2020.]

Habash, N. (2010). Introduction to Arabic natural language processing. Morgan and Claypool Publishers.

Hamood, A.H. (2018). Integrating artificial intelligence in teaching and learning the Holy Quran for non-Arabic speakers. Journal of Physics: Conference Series, 1028(1), 012013.

Holtzman, A., Buys, J., Du, L., Forbes, M., and Choi, Y. (2020). The curious case of neural text degeneration. International Conference on Learning Representations (ICLR).

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H. et al. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems, 43(2), 1-55. doi:10.1145/3703155.

Emad Eldin, F. 2025, 'Isnad AI at IslamicEval 2025: a rule-based system for identifying Islamic citation in LLM outputs', in Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks, Association for Computational Linguistics, Suzhou, China, pp. 540–559.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y.J., Dai, W., and Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1-38.

Jurafsky, D. and Martin, J.H. (2009). Speech and language processing, 2nd ed. Pearson Prentice Hall.

Koppel, M., Schler, J., and Argamon, S. (2011). Authorship attribution in the wild. Language Resources and Evaluation, 45(1), 83-94.

Lake, B.M. and Baroni, M. (2018). Generalisation without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. Proceedings of the 35th International Conference on Machine Learning, 2879-2888.

Lazaridou, A., Kuncoro, A., Gribovskaya, E., Agrawal, D., Liška, A., Terzi, T., Gimenez, M., de Masson d'Autume, C., Kociský, T., Ruder, S., Yogatama, D., Cao, K., Young, S. and Blunsom, P. 2021, 'Mind the gap: assessing temporal generalisation in neural language models', Advances in Neural Information Processing Systems (NeurIPS 2021), vol. 34, pp. 29348–29363.

Lee, K., Firat, O., Agarwal, A., Fannjiang, C., and Sussillo, D. (2019). Hallucinations in neural machine translation. International Conference on Learning Representations.

Lewis, P. et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459-9474.

Lin, S.C., Hilton, J., and Evans, O. (2022). TruthfulQA: Measuring how models mimic human falsehoods. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 3214-3252. doi:10.18653/v1/2022.acl-long.229.

Li, J., Galley, M., Brockett, C., Gao, J., and Dolan, W.B. (2016). A diversity-promoting objective function for neural conversation models. Proceedings of NAACL-HLT, 110-119.

Maleki, N., Padmanabhan, B., and Dutta, K. (2024). AI hallucinations: A misnomer worth clarifying. In 2024 IEEE Conference on Artificial Intelligence (CAI), pp. 133–138. IEEE. doi:10.1109/CAI59869.2024.00033.

Mueller, A., Kuester, S., and von Janda, S. (2025). Socially (un)acceptable errors of AI: Consumer perceptions of different AI-induced errors. Journal of Business Research, 201, 115673.

Mallen, A., Asai, A., Zhong, V., Das, R., Khashabi, D. and Hajishirzi, H. 2023, 'When not to trust language models: investigating effectiveness of parametric and non-parametric memories', in Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, pp. 9802–9822, available at: https://doi.org/10.18653/v1/2023.acl-long.546. Manning, C.D., Raghavan, P., and Schutze, H. (2008). Introduction to information retrieval. Cambridge University Press.

Mubarak, H., Malhas, R., Mansour, W., Mohamed, A., Fawzi, M., Hawasly, M., Elsayed, T., Darwish, K. and Magdy, W. 2025, 'IslamicEval 2025: the first shared task of capturing LLMs hallucination in Islamic content', in Proceedings of the Third Arabic Natural Language Processing Conference: Shared Tasks, Association for Computational Linguistics, Suzhou, China, pp. 480–493, available at: https://doi.org/10.18653/v1/2025.arabicnlp-sharedtasks.67.

Mitchell, T.M. (1997). Machine learning. McGraw-Hill.

Nagy, G. et al. (2019). Automated digitisation and preservation of rare manuscripts. Digital Humanities Quarterly, 13(2).

Nishisako, S., Higashi, T. and Wakao, F. 2025, 'Reducing hallucinations and trade-offs in responses in generative AI chatbots for cancer information: development and evaluation study', JMIR Cancer, vol. 11, e70176, available at: https://doi.org/10.2196/70176.

OpenAI (2023). GPT-4 technical report. OpenAI.

OpenAI (2025). Introducing GPT-5. OpenAI. Available at: https://openai.com/index/introducing-gpt-5/

Potter, A. and Saccucci, C. 2024, 'Could artificial intelligence help catalog thousands of digital library books?', The Signal [blog], Library of Congress, November 2024, available at: https://blogs.loc.gov/thesignal/2024/11/could-artificial-intelligence-help-catalog-thousands-of-digital-library-books-an-interview-with-abigail-potter-and-caroline-saccucci/.

Radford, A. et al. (2021). Learning transferable visual models from natural language supervision. Proceedings of the International Conference on Machine Learning, 8748-8763.

Rozenblit, L. and Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. Cognitive Science, 26(5), 521-562.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206-215.

Russell, S. and Norvig, P. (2010). Artificial intelligence: A modern approach, 3rd ed. Prentice Hall.

Roller, S., Dinan, E., Goyal, N., Ju, D., Williamson, M., Liu, Y., Xu, J., Ott, M., Shuster, K., Smith, E.M., Boureau, Y.-L., and Weston, J. (2020). Recipes for building an open-domain chatbot. arXiv preprint arXiv:2004.13637.

Sharaf, M. and Atwell, E. (2012). QurAna: Corpus of the Quran annotated with pronominal anaphora. Language Resources and Evaluation, 46(4), 341-352.

Shuster, K., Komeili, M., Kiela, D., and Weston, J. (2022). Language models that seek for knowledge: Modular search and generation for dialogue and prompt completion. arXiv preprint arXiv:2203.13224.

Siddiqui, A.F. and Al-Emran, M. (2021). Application of artificial intelligence for effective learning of Hadith: A review. International Journal of Advanced Computer Science and Applications, 12(1), 150-159.

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S.R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S.R., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., and Perez, E. (2023). Towards understanding sycophancy in language models. In Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024), available at: https://arxiv.org/abs/2310.13548.

Sloman, S. and Fernbach, P. (2017). The knowledge illusion: Why we never think alone. New York: Riverhead Books.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. Journal of the American Society for Information Science and Technology, 60(3), 538-556.

Sui, P., Duede, E., Wu, S., and So, R.J. (2024). Confabulation: The surprising value of large language model hallucinations. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024).

Vaswani, A. et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998-6008.

Vig, J. (2019). A multiscale visualisation of attention in the transformer model. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 37-42.

Welleck, S., Kulikov, I., Roller, S., Dinan, E., Cho, K., and Weston, J. (2020). Neural text generation with unlikelihood training. International Conference on Learning Representations (ICLR).

Wolf, T. et al. (2020). Transformers: State-of-the-art natural language processing. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 38-45.

Weston, J. and Shuster, K. (2021). Blender Bot 2.0: An open source chatbot that builds long-term memory and searches the internet. Meta AI Blog, 16 July. Available at: https://ai.meta.com/blog/blender-bot-2-an-open-source-chatbot-that-builds-long-term-memory-and-searches-the-internet/ (Accessed: 7 March 2026).

Xu, J., Szlam, A., and Weston, J. (2022). Beyond goldfish memory: Long-term open-domain conversation. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 5180–5197. Dublin: Association for Computational Linguistics. doi:10.18653/v1/2022.acl-long.356.

Zhang, Y., Li, S., Liu, J., Yu, P., Fung, Y.R., Li, J., Li, M. and Ji, H. 2024, 'Knowledge overshadowing causes amalgamated hallucination in large language models', arXiv, available at: https://doi.org/10.48550/arXiv.2407.08039.